Social and Linguistic Speech Prosody

Proceedings of the 7th international conference on Speech Prosody

SpeechProsody 7

(Trinity College Dublin) May 20-23, 2014

ISSN: 2333-2042

Social and Linguistic Speech Prosody

1 Frontmatter/Preface

1.1 Statistics by Country (showing number of authors)

Authors from 45 countries sent in submissions to Speech Prosody 2014
5 countries didn't make it - we hope they'll try again for SP8!

1.2 Accepted authors by country:

Algeria	1
Australia	6
Austria	1
Bangladesh	1
Belgium	6
Brazil	17
Canada	12
China	16
Costa Rica	1
Czech Republic	7
Denmark	1
Estonia	9
European Union	281
Finland	8
France	59
Germany	70
Hong Kong	6
Hungary	26
India	4
Iran	1
Ireland	12
Israel	3
Italy	15
Japan	29
Mexico	1
Netherlands	16
Norway	1
Poland	8
Portugal	6
Qatar	1
Russian Federation	2
Saudi Arabia	1
Slovakia	3
South Africa	2
Spain	18
Swaziland	1
Sweden	7
Switzerland	12
Taiwan	7
UK	25
USA	76

1.3 Acceptance rates:

We have seen a 35% increase in acceptances since Speech Prosody in Nara, 10 years ago, and a 68% increase in submissions.

2004 (in Nara): 164/180; 29 oral and 135 poster
2014 (in Dublin): 222/303; 42 oral and 180 poster

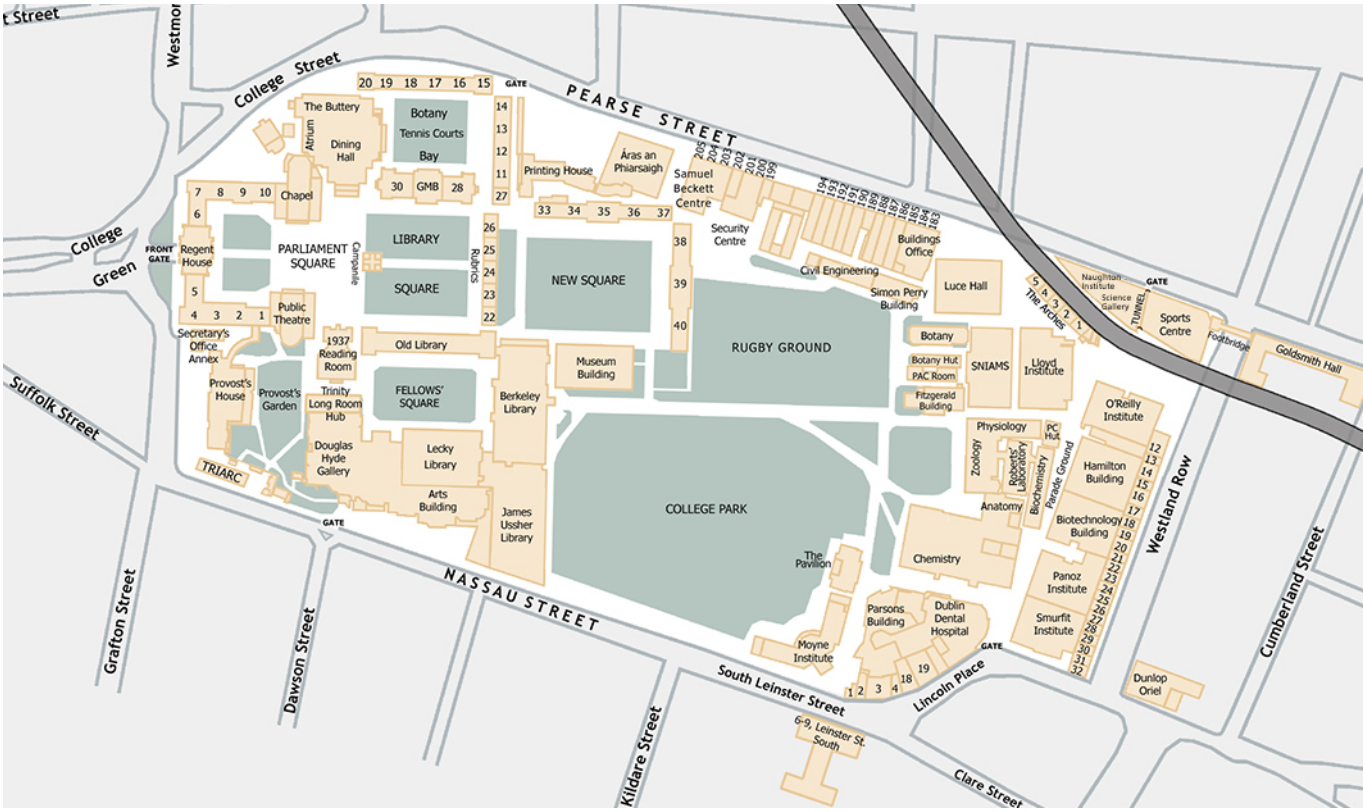
2004: 91% acceptance rate
2014: 73% acceptance rate

2004: 21% oral/poster ratio
2014: 23% oral/poster ratio

altogether 480 authors and 4 keynote speakers are represented at SP7

Speech Prosody - brought to you in Dublin by the Pros Bros!

(photo courtesy of Jolanta Bachan, Brighton 2012)



Finding us!

Speech Prosody 2014 will be hosted in the 'Ed Burke' Theatre, deep in the Arts Building of Trinity College Dublin (*the* University of Dublin ;-)

Trinity College is an academic island of quiet history right in the centre of Dublin City (the Aircoach stops just outside) and the Arts Building has an entrance in Nassau Street, facing Dawson Street but it is much nicer to enter from Fellows' Square (through Front Gate) and enjoy the old College

(please don't forget to visit the Long Room (Old Library) while you are here ;-)

and actually, while you're here, the Chester Beatty Library in Dublin Castle is also well worth a visit

and while we're on the subject of libraries, Marsh's Library is offering a great exhibition on Japan!

. . . . . but this is all for Saturday - we hope you'll be with us full-time until then . . .

An essay on 'ProsBros' (1st version, DG)

'ProsBros', alternatively 'Pros-Bros', 'Pros Bros' … what's that? By induction from context and some help from Monty Python, you will have concluded that 'ProsBros' is the name of a set:

ProsBros = {Nick, Daniel, Dafydd}

That's it, a nickname for the three of us. (Why NICKname, by the way? Not fair.) So why 'ProsBros'? Analogical thinking will no doubt yield the tentative hypothesis that 'Pros' is an abbreviation for 'prosody'. Well done, correct! But 'Bros'? Analogical thinking may lead you to think that it is the plural of 'Bro', which is an abbreviation of 'brother', almost correct, though as you will see it is actually a direct abbreviation for 'brothers'. Now, having clarified the extensional semantics and the morphology, the intensional semantics, pragmatics and phonetics remain to be clarified.

First, semantics. Yes, we wrote our dissertations on prosody (Daniel and Dafydd on intonation, Nick on timing), and have since continued to work in the field, with occasional deviations (the most deviant being Dafydd). Yes, we have all worked with computational phonetic tools, yes, we have all worked extensively with speech corpora (the most well-known being Dan's), yes, we have all worked with speech synthesis (most of all, Nick).

Second, pragmatics. Yes, we are roughly the same generation. Yes we have been friends for decades. We are three expat Brits (hence the undeniable influence of Monty Python on this note) who have each worked mainly on other languages than English: Nick on Japanese, Dan on French, and Dafydd on German (and a collection of African and Asian languages). We have been heavily involved in creating and supporting international infrastructures in these fields: Speech Prosody, COCOSDA, international project consortia. Ironically, we have been often been vicarious representatives of these speech communities in committees and conferences: "What is the Japanese perspective on this, er, Nick?" ... "What is the French perspective on this, er, Dan?" ... "What is the German perspective on this, er, Dafydd?"

Third, phonetics. Yes this one of the main areas in which we work. Now note that 'Pros' and 'Bros' could both rhyme with 'Oz', 'boss', 'rose' or 'gross', yielding 16 possible combinations. Following Occam's Razor, we reject 'rose' and 'gross' as too complex (alternatively: too emotional), leaving 4 combinations, and to restrict the search space we propose a new markedness constraint 'Disyllabic Nickname Rhyme Harmony' (*DNRH) in Optimal Nickname Theory:

**XAYB*, where onset(*X*), onset(*Y*), rhyme(*A*), rhyme(*B*), for *A*≠*B*

The *DNRHC permits only [prɔsbrɔs] and [prɔzbrɔz] (ignoring other segmental details). Controversially, in acknowledgment of a plethora of languages with final devoicing, [prɔsbrɔs] is marginally preferred to [prɔzbrɔz], and together with the English Compound Stress Rule, ['prɔsbrɔs] emerges as the favoured pronunciation, though ['prɔzbrɔz] is a close second.

However, the selection is finally clinched by further analogical thinking, which will initially only be accessible to Brits. So a little cultural history: in 1898 the legendary sartorial hire business 'Moss Bros', well known throughout the UK and beyond, was established in London by the brothers Alfred and George Moss, who deserve the Noble Prize [sic] for achieving the remarkable goal of ensuring that businessmen worldwide (and some businesswomen) wear the same style of suit, shirt and tie. No, we do not normally wear these suits, shirts and ties, but we were strongly influenced by the name of the business (and, as noted above, by Monty Python).

Now the gentle reader may wish to face the challenge of a final exercise in induction and analogical thinking: How is 'Moss Bros' pronounced?

Programme

Nick Campbell	Trinity College Dublin, the University of Dublin, Ireland
Dafydd Gibbon	Universität Bielefeld, Germany
Daniel Hirst	CNRS & Université de Provence, France

International Advisory Committee

Paavo Alku	Aalto University
Véronique Aubergé	Grenoble LIG
Christophe D'Alessandro	LIMSI-CNRS
Plinio Barbosa	University of Campinas
Fred Cummins	University College Dublin
Grazyna Demenko	Adam Mickiewicz University
Hongwei Ding	Tongji University
Jens Edlund	Royal Technical Institute (KTH)
Mária Gósy	Hungarian Academy of Sciences
Mark Hasegawa-Johnson	University of Illinois at Urbana-Champaign
Keikichi Hirose	University of Tokyo
Oliver Jokisch	Leipzig University of Telecommunication
Haruo Kubozono	National Institute for Japanese Language and Linguistics
Philippe Martin	Université Paris Diderot
Hansjörg Mixdorff	Beuth University of Applied Sciences
Bernd Möbius	Dept. of Comp. Ling. and Phonetics, Saarland University
Toshiyuki Sadanobu	Kobe University
Yoshinori Sagisaka	Waseda University
Jan van Santen	Center for Spoken Langauge Processing
M.G.J. Swerts	Tilburg University, School of Humanities
Jürgen Trouvain	Saarland University
Khiet Truong	University of Twente
Chiu-Yu Tseng	Institute of Linguistics, Academia Sinica
Petra Wagner	Universität Bielefeld
Nigel Ward	University of Texas at El Paso
Yi Xu	University College London

Special Session Organisers

Hiroya Fujisaki	University of Tokyo
Toshiyuki Sadanobu	Kobe University
Véronique Aubergé	Laboratory of Informatics of Grenoble (LIG)
Marzena Żygis	Zentrum für Allgemeine Sprachwissenschaft, Berlin
Zofia Malisz	Universität Bielefeld

Programme Committee

IAC (see above)	**all IAC members were active reviewers**
Noam Amir	Tel Aviv university
Bistra Andreeva	Department of Computational Linguistics and Phonetics
Amalia Arvaniti	University of Kent
Véronique Aubergé	LIG Grenoble
Cinzia Avesani	ISTC-CNR
Anton Batliner	Lehrstuhl fuer Mustererkennung
Stefan Baumann	IfL Phonetik, Cologne University
Pier Marco Bertinetto	Scuola Normale Superiore
Roxane Bertrand	Laboratoire Parole et Langage, UMR 6057 CNRS

Maria Paola Bissiri	Technische Universität Dresden
Antonio Bonafonte	UPC
Francesca Bonin	Trinity College Dublin
Genevieve Caelen-Haumont	MICA laboratory
Aoju Chen	Utrecht University
Sin-Horng Chen	National Chiao Tung University
Robert Clark	The University of Edinburgh
Jennifer Cole	University of Illinois
Ricardo Cordoba	Grupo de Tecnologia del Habla, Madrid
Snezhina Dimitrova	University of Sofia
Elisabeth Delais-Roussarie	CNRS-Université Paris 7 Paris Diderot,
Gorka Elordieta	University of the Basque Country
John Esling	University of Victoria
Sascha Fagel	zoobe message entertainment GmbH
Zsuzsanna Fagyal-Le Mentec	University of Illinois at Urbana-Champaign
Isabel Falé	Universidade Aberta/CLUL
Janet Fletcher	School fo Languages & Linguistics University of Melbourne
Sónia Frota	Universidade de Lisboa
Hiroya Fujisaki	University of Tokyo
Dafydd Gibbon	Universität Bielefeld
Matt Gordon	UC Santa Barbara
Björn Granström	KTH, Sweden
Carlos Gussenhoven	Radboud University, Nijmegen
Mária Gósy	Research Institute for Linguistics, HAS
Sophie Herment	Université de Provence
Daniel Hirst	CNRS & Université de Provence
Merle Horne	Lund University
David House	KTH, Sweden
Jill House	University College London
Sarmad Hussain	CLE-KICS, UET
Ignasi Iriondo	Enginyeria i Arquitectura La Salle. Universitat Ramon Llull
Stefanie Jannedy	Center for Linguistics (ZAS)
Sun-Ah Jun	UCLA
Maciej Karpiński	Adam Mickiewicz University
Hideki Kawahara	Wakayama University
Tatsuya Kawahara	School of Informatics, Kyoto University, Kyoto, Japan
Roland Kehrein	Uni Marburg, Germany
Esther Klabbers	OHSU
Jody Kreiman	University of California, Los Angeles
Frank Kügler	Potsdam University
Haizhou Li	Institute for Infocomm Research
Yuan-Fu Liao	National Taipei University of Techonology
Joaquim Llisterri	Universitat Autònoma de Barcelona
Madureira	PUCSP
Zofia Malisz	Universität Bielefeld
Piet Mertens	K.U.Leuven
Nobuaki Minematsu	University of Tokyo
Helena Moniz	FLUL/INESC-ID
Hiroki Mori	Utsunomiya University
Shrikanth Narayanan	University of Southern California
Eva Navas	University of the Basque Country
Oliver Niebuhr	Dept. of General Linguistics, University of Kiel
Elmar Nöth	University of Erlangen-Nuremberg
Michael O'Dell	University of Tampere
John Ohala	University of California, Berkeley
Zdena Palkova	Institute of Phonetics, Charles University Prague
Prem C. Pandey	Indian Institute of Technology Bombay
Gabor Pinter	Kobe University
Bernd Pompino-Marschall	HU Berlin
Heather Pon-Barry	Arizona State University
Cristel Portes	Universite de Provence

Brechtje Post	University of Cambridge
Hugo Quené	Utrecht University
César Reis	Universidade Federal de Minas Gerais
Eduardo Rodriguez Ba]nga	University of Vigo
Daisuke Saito	UNversity of Tokyo
Elina Savino	University of Bari
Amy Schafer	University of Hawaii
Antje Schweitzer	Stuttgart University
Jane Setter	University of Reading
Chilin Shih	University of Illinois at Urbana-Champaign
Elizabeth Shriberg	SRI International
Miquel Simonet	University of Arizona
Shari Speer	Ohio State University
Marc Swerts	Tilburg University
Jianhua Tao	Chinese Academy of Sciences
Shu-Chuan Tseng	Institute of Linguistics, Academia Sinica
Alice Turk	University of Edinburgh
Vincent Van-Heuven	University of Leiden
Nanette Veilleux	Simmons College
Céu Viana	inesc-id, Portugal
Yue Wang	Simon Fraser University
Duane Watson	University of Illinois Urbana-Champaign
Stefan Werner	University of Eastern Finland
Laurence White	Plymouth University
Marcin Wlodarczak	Universität Bielefeld
Chai Wutiwiwatchai	Human Language Technology Laboratory, NECTEC
Jiahong Yuan	University of Pennsylvania

Assistant Reviewers

Amengual, Mark
Arnold, Denis
Batista, Fernando
Casillas, Joseph
Correia, Susana
Jügler, Jeanin
Rohena-Madrazo, Marcos
Šimko, Juraj
Steiner, Ingmar
Wang, Yang

Local Organising Committee

Mai & Sarah & Lucy @ Odyssey

Odyssey International Incentives & Meetings
6-8 Garville Lane, Rathgar, Dublin 6, Ireland

Tel : + 353 1 497 4866
Fax : + 353 1 496 1396

with sincere thanks also to Fáilte Ireland and the SFI for their kind help!

Speech Prosody - a brief history

(according to the pros bros)

The first international meeting on Speech Prosody that we know of was a three-day seminar on Intonation and Discourse organised by the British Association for Applied Linguistics in Birmingham in April 1982. Strangely enough, the second meeting was held just two weeks later in Paris: a workshop on Prosody organised by the European Association for Psycholinguistics. This was soon followed by a Working Group on Intonation that was a satellite event preceding the 13th International Congress of Linguists in Tokyo.

And then, after that, nothing for more than ten years...

Before the next prosody meeting, the 12th ICPhS meeting was held in Aix-en-Provence in August 1991 where we were struck by the large number of papers on the topic of Speech Prosody - more than 20% of the papers which were directly related to this topic and the small room assigned was overflowing into the corridors! Dan asked Mario Rossi, the conference chairman, to announce an ad-hoc meeting before one of the plenary sessions and over 100 people turned up, no doubt wondering what to expect. As it happened we had no more idea than they did but among suggestions made then were: setting up an international organisation - organising conferences on prosody - and setting up a prosody mailing list.

George Allen volunteered to set up an email list for prosody and this list was a valuable resource for many years, although the fact that it included both literary prosody (versification) and speech prosody was slightly confusing for some people. There would be a flurry of engineers signing off the list after a posting on mediaeval versification, followed by an equally urgent flurry of linguists quitting the prosody ship after an engineer had posted something on hidden Markov models for speech recognition. We eventually decided to replace the list by one specifically devoted to Speech Prosody.

In the years following this meeting, ICSLP (International Conference on Spoken Language Processing) and ESCA (European Speech Communication Organisation) organised some more workshops on prosody (Lund 1993, Yokohama 1994, Athens 1995, Kraków 1999) and workshops on prosody also became a regular feature of the ICPhS meetings (Stockholm 1995, San Francisco 1999). Despite this welcome increase in the number of meetings, there was still something missing. From one meeting to the next there was no way of telling when or where the next meeting would be held. Each meeting was organised as a separate event with no co-ordination. This made it hard for potential participants to plan to present a paper on prosody. It was particularly true for doctoral students, who had no way of knowing if there would be an appropriate meeting somewhere where they could present their research before the end of their thesis preparation.

In September 1999, ESCA and ICSLP agreed to combine and were renamed ISCA (International Speech Communication Association) and the new organisation soon set up Special Interest Groups to promote research on specific topics of speech communication or for specific languages. With the wider availability of internet, it was now quite easy to contact a large number of specialists on Speech Prosody, asking if they would agree to support the creation of a Special Interest Group on Speech Prosody (SPròSIG). The response was enthusiastic - 72 established researchers in the field from 23 different countries agreed to support the SIG and in January 2000 the group was recognised by ISCA.

The 'Prosody 2000' Kraków meeting, chaired by Wiktor Jassem, was a kind of Proto-Speech Prosody, actually a combination of two workshops - one on Speech Recognition and Synthesis, organised by our prosody sister Grazyna Demenko with the help of Dafydd Gibbon, and one on Prosodic Transcription and Modelling organised by Esther Grabe, Kai Alter and Hansjörg Mixdorff.

The ISCA/SPròSIG event, the First International Conference on Speech Prosody, Aix-en-Provence, April 3 2002 chaired by Daniel Hirst, was a success with 152 submitted papers plus 6 invited keynote speakers, 12 invited co-speakers. It was attended by 317 people. Of course the question everybody asked was: would it last? The list of Speech Prosody meetings speaks for itself: 2004 Nara Japan, 2006 Dresden Germany 2008, Campinas Brazil, 2010 Chicago USA, 2012 Shanghai China, 2014 Dublin, Ireland; and the next will be in 2016 - but where ???

A modification of the constitution of SPròSIG in 2010 added a Permanent Advisory Committee, consisting of the founder officers of SPròSIG (Daniel Hirst, Nick Campbell, Bernard Bel), the current elected officers (currently Keikichi Hirose, Yi Xu, Mark Hasegawa-Johnson, Hansjorg Mixdorff) as well as the chair and co-chair of the last 5 conferences.

At the last meeting in Shanghai (2012), the PAC decided to nominate Hiroya Fujisaki as Honorary Life Member of the PAC. The 2011 board meeting of the International Phonetic Association resolved to co-sponsor the Speech Prosody Special Interest Group - this resolution was adopted formally in October 2012. The Speech Prosody SIG is, consequently, now officially affiliated with both ISCA and IPA.

Speech Prosody is too serious a matter to be left in the hands of just engineers... ... or just linguists. We really need each other, and Speech Prosody is a way to bring us all together!

SP7 - another SPròSIG event - with special thanks to :

The Team:

EasyChair:

Odyssey:

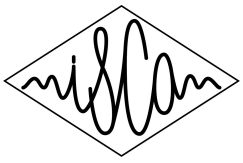
Science Foundation Ireland:

The University of Dublin:

Fáilte Ireland:

SPròSIG:

ISCA



IPA

and YOU!

Daily Event Programme

Speech Prosody will feature 222 papers in plenary oral and poster format: there will be 4 invited keynotes each followed by 3 oral presentations on a related theme, and 3 oral thematic sessions of 6 papers each. Each day will have a poster session and posters can remain in place throughout the day. There is no distinction in rank between an oral and a poster presentation. There will be 2 special sessions, a Round Table, and a Panel Discussion, with a special commemoration of Prof Miyoko Sugito, the honorary co-chair of Speech Prosody 2004, to remember her contributions to our field. Lunch will be provided each day.

Day-1.1 Tuesday May 20th, 1:30pm - 2pm : Opening (3 bros:-welcome!)
Day-1.2 Tuesday May 20th, 2pm - 3:30pm : plenary (1+3 oral) Invited Keynote
Day-1.3 Tuesday May 20th, 4pm - 5:30pm : special (Booster & Round Table)
Day-1.4 Tuesday May 20th, 5:30pm - 6pm : SSSpASSS (20 Special Session posters)
Day-1.5 Tuesday May 20th, 6pm - 7:30pm : (28 Posters with **WELCOME RECEPTION**)

Day-2.1 Wednesday May 21st, 9am - 10:30am : plenary (1+3 oral) Invited Keynote
Day-2.2 Wednesday May 21st, 11am - 1pm : (6 oral) - Speech Rhythm and Timing
Day-2.3 Wednesday May 21st, 2pm - 4pm : special session - Slavic Prosody
Day-2.4 Wednesday May 21st, 4:30pm - 6pm : (48 posters) Perception and Intonation
Day-2.5 Wednesday - evening - **Reviewers' Reception** in Trinity College Long Room

Day-3.1 Thursday May 22nd, 9am - 10:30am : 3-1-plenary (1+3 oral) Invited Keynote
Day-3.2 Thursday May 22nd, 11am - 1pm : (48 posters) Theoretical and Linguistic Prosody
Day-3.3 Thursday May 22nd, 2pm - 3:30pm : Panel Session & memorial
Day-3.4 Thursday May 22nd, 4pm - 6pm : (6 oral) Perception and Production
Day-3.5 Thursday May 22nd, 7pm - 9:30pm : **BANQUET - Old Dining Hall** - *ALL WELCOME!*

Day-4.1 Friday May 23rd, 9am - 10:30am : (48 posters) Intonation and Speaking Style
Day-4.2 Friday May 23rd, 11am - 1pm : (6 oral) Intonation - general
Day-4.3 Friday May 23rd, 2pm - 3:30pm : plenary (1+3 oral) Invited Keynote
Day-4.4 Friday May 23rd, 3:30pm - 4pm : closing

Day-5 Saturday May 24th, **FREE DAY**

Day One

Tuesday May 20th, 1:30pm - 2pm : Opening (welcome to Speech Prosody!)

Tuesday May 20th, 2pm - 3:30pm : plenary (1+3 oral presentations) (p.1)

Invited Keynote: Fred Cummins - followed by 3 oral presentations
Chair: Ailbhe Ní Chasaide

Coffee break - 3:30 - 4pm

Tuesday May 20th, 4pm - 5:30pm : (Booster & Round Table) (p.2)

SSSpASSS: Special Session: **Social prosody: Affective Social Speech Signals**
Chair: Véronique Aubergé
Véronique and team to introduce SSSpASSS posters (booster: 2-min per paper) with Round Table discussion to follow, then posters in the exhibition space with welcome reception

Tuesday May 20th, 5:30pm - 6pm : SSSpASSS Posters) (p.2)

Chair: Gu Wentao

Tuesday May 20th, 6pm - 7:30pm : - Prominence & Phrasing - (28 posters) (p.7)

Chair: Juraj Šimko

Day Two

Wednesday May 21st, 9am - 10:30am : plenary (1+3 presentations) (p.15)

Invited Keynote: Stefanie Shattuck-Hufnagel - followed by 3 oral presentations
Chair: Julia Hirschberg

Coffee break - 10:30 - 11am

Wednesday May 21st, 11am - 1pm : oral (6 presentations) (p.16)

- Speech Rhythm and Timing - Chair: Amalia Arvaniti

Lunch break - 1 - 2pm (lunch provided)

Wednesday May 21st, 2pm - 4pm : special session - Slavic Prosody (p.18)

special session - Slavic Prosody (another SP!) Chairs: Zofia Malisz & Marzena Żygis

Coffee break - 4:00 - 4:30pm

Wednesday May 21st, 4:30pm - 6pm : poster (48 presentations) (p.19)

- Perception and Intonation - Chair: Oliver Jokisch

evening - Reviewers' Reception in Trinity College Long Room (by invitation)

Day Three

Thursday May 22nd, 9am - 10:30am : plenary (1+3 presentations) (p.31)

Invited Keynote: Jürgen Trouvain followed by 3 oral presentations
Chair: Nigel Ward

Coffee break - 10:30 - 11am

Thursday May 22nd, 11am - 1pm : poster (48 presentations) (p.32)

- Theoretical and Linguistic Prosody - Chair: Alice Turk

Lunch break - 1 - 2pm (lunch provided)

Thursday May 22nd, 2pm - 2:30pm : Remembering Sugito Miyoko (p.44)

Chair: Sadanobu Toshiyuki

Thursday May 22nd, 2:30pm - 3:30pm : Terminology in Prosody Research (p.44)

Chair: Fujisaki Hiroya

Coffee break - 3:30 - 4pm

Thursday May 22nd, 4pm - 6pm : oral (6 presentations)

- Perception and Production - Chair: Agnieszka Wagner

Thursday May 22nd, 7pm - 9:30pm : BANQUET - Old Dining Hall - ALL WELCOME!

Day Four

Friday May 23rd, 9am - 10:30am : poster (48 presentations) (p.46)

- intonation and speaking style - Chair: Laura Dilley

Coffee break - 10:30 - 11am

Friday May 23rd, 11am - 1pm : oral (6 presentations) (p.59)

- Intonation - Chair: Hansjörg Mixdorff

Lunch break - 1 - 2pm (lunch provided)

Friday May 23rd, 2pm - 3:30pm : plenary (1+3 presentations) (p.61)

Invited Keynote: Anne Cutler - followed by 3 oral presentations
Chair: Chiu Yu Tseng

Friday May 23rd, 3:30pm - 4pm : closing

Saturday May 24th, Free Day - go explore Dublin!!!

The Authors index

(sorted by given name - for an index sorted by family name see the Author index at the end of the proceedings (p.1180))

Page numbers refer to the abstract entry where a link can be found to the full paper.

Abstracts

NOTE: page numbers refer to the digital form of the proceedings where full papers are included - these can be downloaded from http://www.speechprosody2014.org/proceedings.pdf

1 Day One - May 20th

Tuesday - Opening Session

1:30pm - 2pm : 1-0-opening (3 bros:welcome!etc)

1.1 Tuesday Session One

2pm - 3:30pm : 1-1-plenary (1+3 presentations)

1.1.1 KeyNote 1

Fred Cummins - 30-min
From Prayer to Protest: An Initial Look at Joint Speech
Joint speech is an umbrella term covering choral speech, synchronous speech, chant, and all forms of speech where many people say the same thing at the same time. Prosodists, more than most, should be aware of the incompleteness of a structuralist description of language. Much of our use of language is ignored or missed when linguistic behaviour is viewed through the narrow lens of phonological/syntactic structure. I will discuss Joint Speech, as found in prayer, protest, classrooms, and sports stadia around the world. Despite its deep embedding in practices we value very much, joint speech has not hitherto attracted the attention of scientists as a distinct form of language behavior, because it is uninteresting from a structuralist point of view. If we merely take the time to look, however, there is much to be found in joint speech that is crying out for elaboration and investigation. I will attempt to sketch the terra incognita that opens up and present a few initial findings (phonetic, anthropological, neuroscientific) that suggest that Joint Speech is far from being a peripheral and exotic special case. It is, rather, a central example of language use that must inform our theories of what language and languaging are.

1.1.2 p.65

Zhihua Xia, Rivka Levitan, Julia Hirschberg,
Prosodic Entrainment in Mandarin Chinese and English: A Cross-Linguistic Comparison
Entrainment is the propensity of speakers to begin behaving like one another in conversation. We identify evidence of entrainment in a number of acoustic and prosodic dimensions in conversational speech of American English speakers and Mandarin Chinese speakers. We compare entrainment in the Columbia Games corpus and the Tongji Games Corpus and find some remarkable similarities between the two.

1.1.3 p.70

Rob Voigt, Robert J. Podesva, Dan Jurafsky,
Speaker Movement Correlates with Prosodic Indicators of Engagement
Recent research on multimodal prosody has begun to identify associations between discrete body movements and categorical acoustic prosodic events such as pitch accents and boundaries. We propose to generalize this work to understand more about continuous prosodic phenomena distributed over a phrase - like those indicative of speaker engagement - and how they covary with bodily movements. We introduce movement amplitude, a new vision-based metric for estimating continuous body movements over time from video by quantifying frame-to-frame visual changes. Application of this automatic metric to a collection of video monologues demonstrates that speakers move more during phrases in which their pitch and intensity are higher and more variable. These findings offer further evidence for the relationship between acoustic and visual prosody, and suggest a previously unreported quantitative connection between raw bodily movement and speaker engagement.

1.1.4 p.75

Štefan Beňuš, Marián Trnka,
Prosody, voice assimilation, and conversational fillers
Conversational fillers (CFs), commonly transcribed as uh, um, or er, typically start with a schwa-like vowel, and signal multiple social, interactive, meta-cognitive, and pragmatic functions. They also co-occur with prosodic boundaries, increase saliency of inter-word disjunctures, and participate thus in coding the prosodic structure. Contrary to these functions, CFs are assumed not to participate in the phonological system of a language. This paper uses two types of Slovak conversational speech corpora for investigating the the prosodic and phonological behavior of CFs. In Slovak, the vowel inventory does not include a schwa, and word-final obstruents undergo voice assimilation that is triggered by word-initial vowels but interacts with the strength of the prosodic boundary between the two words. Our data show the propensity of CFs to neutralize word-final voicing, and function thus as prosodic breaks, but also non-negligible number of cases of CFs triggering voicing of word-final obstruents, supporting their relevance for cognitive phonology.

1.2 Tuesday Special Session - SSSpASSS

4pm - 5:30pm : Booster & Round Table

Special Session: Social Prosody: **Affective Social Speech Signals**

Véronique and team to introduce SSSpASSS posters (booster: 2-min per paper) with Round Table then poster session to follow after break (posters with Welcome Reception)

5:30pm - 6pm : (20 poster presentations)

Tuesday Session Two - SSSpASSS Posters

1.2.1 p.81

Marta Maffia, Elisa Pellegrino, Massimo Pettorino,
Labeling expressive speech in L2 Italian: the role of prosody in auto-and external annotation
The present study is intended to compare two approaches of labeling expressive corpora: auto-annotation and annotation by external lay listeners. These two methods have been applied to the semi-spontaneous emotional speech produced by Chinese learners of L2 Italian, involved in the CardTask, a mood-induction procedure that permits to control the context of interaction, preserving the spontaneity of reactions. The emotional responses to the stimuli presented in the task were object of an auto-annotation session. The same samples were then administered only in the auditory mode to 20 Italian and 20 Chinese lay listeners. The results of perceptual tests have underlined some similarities and differences both between auto- and external annotation, and between the rates given by Italian and Chinese external listeners. The labels chosen by native Italians were similar to those selected in the auto-annotation session, particularly in the case of anxiety, fear and disgust. The correspondence between the results of the two annotation methods may be ascribed to the different prosodic patterns characterizing the emotional states. The results of the annotation made by Chinese listeners show that they found it hard to give a specific emotional label to utterances produced in a second language relying only on prosodic patterns.

1.2.2 p.86

Yuko Sasa, Véronique Aubergé,
Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue"
The aim of this preliminary study of feasibility is to give a glance at interactions in a Smart Home prototype between the elderly and a companion robot that is having some socio-affective language primitives as the only vector of communication. Through a Wizard of Oz platform (EmOz), a robot is introduced as an intermediary between the technological environment and some elderly who have to give vocal commands to the robot to control the Smart Home. The robot vocal productions increases progressively by adding prosodic levels: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. The elderly subjects' speech behaviors confirm the hypothesis that the socio-affective "glue"

effect increase towards the prosodic levels, especially for socio-isolated people.

1.2.3 p.91

Noor Alhusna Madzlan, Jingguang Han, Francesca Bonin, Nick Campbell,
Towards Automatic Recognition of Attitudes: Prosodic Analysis of Video Blogs
Understanding of speakers' attitude is essential for establishing successful human interaction. In this paper we analyse attitude manifestations in video blogs. We describe the main features of this novel communication medium and focus our attention on its possible exploitation as a rich source of information for human-human and human-machine communication. We describe the manual annotation of attitudes and the prosodic analyses. Finally we present a preliminary attitude automatic annotation system that attains 65% accuracy.

1.2.4 p.95

Pan Liu, Marc Pell,
Processing emotional prosody in Mandarin Chinese: A cross-language comparison
To understand how emotional prosody is processed in Mandarin Chinese and whether it differs from that of other languages, we conducted a perceptual-acoustic study on a set of Chinese vocal emotional stimuli and examined how they were perceived and acoustically characterized, in comparison with four other languages, English, Arabic, German, and Hindi, reported by Pell et al. [1]. Chinese pseudo-utterances spoken in seven emotions (anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality) were first identified by a group of native Mandarin speakers in a seven forced choice task, and then subjected to acoustic analyses. Results revealed that among the seven emotions, neutrality, anger, sadness, and fear tended to be recognized most accurately. Acoustic analysis demonstrated the importance of three acoustic parameters (f0 mean, f0 range, and speech rate) in characterizing vocal emotions in Mandarin. Both the perceptual and acoustic characteristics are highly similar, although not identical, to that observed by Pell et al. [1] in English, Arabic, German, and Hindi, indicating a set of universal principles in vocal emotion communication across languages.

1.2.5 p.100

Carlos Ishi, Hiroaki Hatano, Miyako Kiso,
Acoustic-prosodic and paralinguistic analyses of "uun" and "unun"
The speaking style of an interjection contains discriminative features on its expressed intention, attitude or emotion. In the present work, we analyzed acoustic-prosodic features and the paralinguistic functions of two variations of the interjection "un", a lengthened pattern "uun" and a repeated pattern "unun", which are often found in Japanese conversational speech. Analysis results indicate that there are differences in the paralinguistic function expressed by "uun" and "unun", as well as different trends on F0 contour types according to the conveyed paralinguistic information.

1.2.6 p.105

Jean Philippe Goldman, Tea Prŝir, George Christodoulides, Antoine Auchlin,
Speaking style prosodic variation: an 8-hour 9-style corpus study
This paper presents the results of a prosodic and phonostylistic analysis based on C-PhonoGenre, a 9-hour-long spoken French corpus, including 10 speakers on average recorded in 10 speaking situations. The corpus was automatically segmented at phonetic, syllabic, word levels (EasyAlign), and larger pause-separated units. Part-of-speech annotation (DisMo) and prominent syllable detection (ProsoProm) was added automatically. The corpus was also manually annotated at the syllabic level for stylistic variants, such as post-tonic schwas, liaisons, elisions, disfluencies, audible breaths and noises. Acoustic analyses (ProsoReport, DurationAnalyser) provide more than 100 micro- and macro-prosodic measures, which we correlate with the phonostylistic and linguistic annotation. This analysis finally yields a contrastive, fine-grained prosometric description of phonostylistic and situational variation, over 4 situational gradual dimensions: audience, media, preparation, and interactivity. Further statistical analysis was carried out to explore the discriminative and explanatory power of combinations of prosodic measures.

1.2.7 p.110

Leandra Antunes, Véronique Aubergé, Yuko Sasa,
Certainty and uncertainty in Brazilian Portuguese: methodology of spontaneous corpus collection and data analysis
This work presents a methodology used to collect some spontaneous social affect corpus and preliminary prosodic analysis of certainty and uncertainty in Brazilian Portuguese. The corpus was collected by a Wizard of Oz (Emoz)method, the scenario to induce certainty and uncertainty is based on the situation of a job interview, for which a companion robot (Emox) is supposed to be a trainer. The subjects were convinced to benefit of a free training of this "revolutionary" method to train to job interview. In this scenario the linguistic expressions are partially controlled, in order to focus the certainty/uncertainty expression mainly on paraphrasing and prosody. Data were preliminary analyzed for audiovisual prosody: videos analysis were made regarding eyes, mouth and face/head movements, while audio analysis were made about acoustic prosody parameters of fundamental frequency and duration. The first results show that using Emoz within such a scenario is an efficient way to induct spontaneous but comparable speech production. Prosodic results show that fundamental frequency and duration measurements, as well as eyes, mouth and face/head movements, are differently used in certainty and in uncertainty production in Brazilian Portuguese.

1.2.8 p.115

Jan Volín, Lenka Weingartová, Oliver Niebuhr,
Between Recognition and Resignation The Prosodic Forms and Communicative Functions of the Czech Confirmation Tag "jasně"
Like question tags, confirmation tags such as the Czech affirmative particle jasně can be used with various prosodic characteristics that augment, reverse or otherwise modify their relatively unspecific lexical meaning. We extracted 172 instances of jasně from several dialogues and assessed their discourse function. 36 prosodic correlates in temporal, amplitude and fundamental frequency domains were measured and used in three computational classifiers: linear discriminant analysis, classification trees and artificial neural networks. All three methods significantly reflected the functional assessments and additionally indicated the relative importance of individual predictors in a mutually consistent manner.

1.2.9 p.120

Ting Wang, Hongwei Ding, Qiuwu Ma, Daniel Hirst,
Automatic Analysis of Emotional Prosody in Mandarin Chinese:Applying the Momel Algorithm
Based on the Momel algorithm, a set of acoustic parameters was analyzed automatically on Chinese emotional speech. Global prosodic features were calculated on the sentence level, which showed a concordance with the usual pattern reported in the literature. Local constraints were also considered on the syllable layer. An ANOVA showed that there were interactive effects among emotions, syllable positions and syllable tones on certain parameters. Further more, by examining the pitch movements, no significant difference was found between neutral speech and active emotional speech, which was different from the performance in non-tonal languages. However when reducing the tonal influence by using only tone 1 syllables in the utterance, this inverse effect disappeared. Hence we posited an interpretation that due to the existence of lexical tone in Mandarin Chinese, the paralinguistic use of pitch movements has been reduced.

1.2.10 p.125

Yan Lu, Véronique Aubergé, Albert Rilliard,
Prosodic Profiles of Social Affects in Mandarin Chinese
An acted corpus of 19 prosodic social affects is devoted to this work, which investigates the production side of prosodic attitudes in Mandarin Chinese, with the aim of extracting the more prominent patterns of acoustical variations. Results are then compared to previous perception data obtained on the same expressions. The F0, intensity and duration characteristics of 76 utterances conveying 19 prosodic attitudes are statistically examined in this study. All attitudes are regrouped into 5 clusters according to their prosodic features. The result of the statistical analysis shows that the prominent differentiation between clusters is mostly related to F0 and duration parameters; some similarities are noted between the clustering of attitudes from acoustic features and from perceptual confusions obtained in previous

experiments; inside each cluster, some attitudes show typical characteristics in F0 and duration.

1.2.11 p.130

Houwei Cao, Štefan Beňuš, Ruben C. Gur, Ragini Verma, Ani Nenkova,
Prosodic cues for emotion: analysis with discrete characterization of intonation
In this paper we study the relationship between acted perceptually unambiguous emotion and prosody. Unlike most contemporary approaches which base the analysis of emotion in voice solely on continuous features extracted automatically from the acoustic signal, we analyze the predictive power of discrete characterizations of intonations in the ToBI framework. The goal of our work is to test if particular discrete prosodic events provide significant discriminative power for emotion recognition. Our experiments provide strong evidence that patterns in breaks, boundary tones and type of pitch accent are highly informative of the emotional content of speech. We also present results from automatic prediction of emotion based on ToBI-derived features and compare their prediction power with state-of-the-art bag-of-frame acoustic features. Our results indicate their similar performance in the sentence-dependent emotion prediction tasks, while acoustic features are more robust for the sentence-independent tasks. Finally, we combine ToBI features and acoustic features together and further achieve modest improvements in sentence-independent emotion prediction, particularly in differentiating fear and neutral from other emotion.

1.2.12 p.135

Massimo Pettorino, Elisa Pellegrino, Marta Maffia,
"Young" and "Old" Voice: the prosodic auto-transplantation technique for speaker's age recognition
The present study is intended to figure out the extent to which prosody and intonation entail listeners' ability to estimate the speaker's age. The performance of a 40-year old anchorman and that produced by the same speaker at the age of 80 were spectro-acoustically analyzed in order to identify the prosodic features of the "young" and the "old" voice. The results of the analyses have shown relevant differences between the two voices on suprasegmental level. To test the effects of these differences on perceptual level, through the prosodic transplantation technique, the F0 values and the durations of segments and silences were transferred from the "young" to the "old" voice and viceversa. Two age recognition tests, based on original and transplanted voices, were administered to Italian listeners. The results of perceptual tests have confirmed the strict relationship between some rhythmic and prosodic features and the speaker's age and have demonstrated the effectiveness of the transplantation technique. With advancing age, articulation rate and speech rate slow down, voice register raises and tonal range widens. Moreover, the "old" voice is also characterized by a higher percentage of vocalic portion that determines a shift of Italian rhythm towards the isomoraic pattern

1.2.13 p.140

Julien Magnier, Maya Gratier, Anne Lacheret,
Expressive prosody vs neutral prosody : From descriptive binary to continuous features
In this paper, we propose to compare expressive and neutral oral renditions of a children's tale in french by examining the segmentations performed by twelve high level french readers. We used a software dedicated to this kind of analysis (Analor) which takes into account different parameters (pause, pitch gesture, pitch jump) and their relative strength to determine pertinent prosodic units (phrases). The extraction of these phrases and their features enables us to observe the influence of both the type of oralisation (expressive or neutral) and punctuation signs on the organization of speech flow. Results show that prosodic phrases in the expressive readings are more numerous (specially at comma locations), that their boundaries are more clearly demarcated, and that they have more varied contours than those in neutral readings.

1.2.14 p.144

Heather Pon-Barry, Arun Reddy Nelakurthi,
Challenges for Robust Prosody-based Affect Recognition
Prosody-based affect recognition has great potential impact for building adaptive speech interfaces. For example, in intelligent systems for personalized learning, sensing a student's level of certainty, which is

often signaled prosodically, is one of the most interesting states to interpret and respond to. However, robust uncertainty recognition faces several challenges, including the lack of gold-standard labels, and differences in expressivity among speakers. In this paper we explore the intersection of these two issues. We have collected a corpus of spontaneous speech in a question-answering task. Three kinds of certainty labels are associated with each utterance. First, speakers rated their own level of certainty. Second, a panel of listeners rated how certain the speaker sounded. Third, an externally crowdsourced difficulty score is generated for each stimulus (the question). We present an analysis of the prosodic characteristics of individual speaking styles, as they relate to these three different measurements of certainty.

1.2.15 p.149

Dominique Fourer, Takaaki Shochi, Jean Luc Rouas, Jean Julien Aucouturier, Marine Guerry,
Prosodic analysis of spoken Japanese attitudes
The aim of this paper is to provide cues for prosodic characterization of attitudes in Japanese. This is comparable with similar researches made on others languages (e.g. American english, Brazilian portuguese, etc.). The presented work focuses on an objective analysis results of the Japanese based on the audio signal structure. In the proposed experiments, the speech signal of several Japanese native speakers is analyzed, The used signals were recorded in a particular context where the corresponding attitude is clearly identified and segmented. The presented results are based on a previous study where 16 attitudes were defined to describe the emotional content of human spoken language. Thus we compare the signal properties which can characterize each attitude.

1.2.16 p.154

Noam Amir, Eitan Globerson,
On the Role of Pitch in Perception of Emotional Speech
Two experiments investigated the role of intonation in perception of basic emotions. In the first experiment, pitch contours of stimuli from a corpus containing portrayals of anger, joy, fear and sadness were manipulated with respect to range, mean and smoothness. In the second experiment, pitch contours of identical words portraying different emotions were exchanged. In each experiment, the emotional category and intensity of the original and manipulated stimuli were evaluated by two separate groups of 20 participants. Results of the first experiment show mainly that pitch mean and range should vary congruently to portray activation correctly, and demonstrate the interaction in varying these two parameters. Results of the second experiment show that a pitch contour conveying high activation is not sufficient in conveying the appropriate emotion, if the other paralinguistic cues are not also in accordance. A pitch contour indicating low activation, on the other hand, is apparently a more powerful cue and thus less reliant on other cues.

1.2.17 p.159

Sven Grawunder, Marianne Oertel, Cordula Schwarze,
Politeness, culture, and speaking task - paralinguistic prosodic behavior of speakers from Austria and Germany
This paper tests previous findings for polite speech of low pitch, low intensity, higher number of hesitation markers and filled pauses against those parameters in a different socio-cultural background. Two similar groups of (19+13) participants, from Austria and from Germany, were recorded. The adopted experimental approach used 16 tasks aiming at different speech acts in situations that evoke either polite or informal speech. The analyzed acoustic and electroglottographic signals reveal main effects for lower pitch, lower intensity and HNR only for the German group. Open quotient values differ only for female speakers. In both groups significantly lower word rate and lower speaking rate as well as higher rates of filled pauses and hesitation markers are found in formal (polite) conditions. However individual speakers can show indifferent or opposing behavior for a given parameter with compensatory utilisation of other parameters in order to express politeness (formality).

1.2.18 p.164

Miguel Oliveira Jr, Ayane Nazarela Santos De Almeida, René Alain Santana De Almeida, Ebson Wilkerson Silva,

Speech rate in the expression of anger: a study with spontaneous speech material
The study of the acoustic expression of emotion is, in general, the analysis of whether prosodic variables such as intonation (F0), speech rate, pauses, rhythm, intensity and duration, are reliable clues for the characterization of the emotional states of the speaker. The present paper aims to verify whether an association exists in Brazilian Portuguese between the basic emotion of "anger" and the prosodic variable "speech rate", as the literature often suggests there is for other languages. The corpus consisted of fragments of spontaneous speech recorded from a radio program. The fragments were selected on the basis of a perceptual test. For the production analysis, only excerpts that were identified by more than 75% of the participants of the perceptual test as associated to the categories "anger" and "neutral" were selected. The results demonstrated that, for the data that were used for the analysis, there is a general reduction in speech rate when utterances are associated with the emotion of "anger", if compared to utterances spoken in a "neutral" mode by the same speaker, contrary to what literature often indicates for other languages.

1.2.19 p.169

Yan Lu, Véronique Aubergé, Nicolas Audibert, Albert Rilliard,
Audiovisual perception of expressions of Mandarin Chinese social affects by French L2 learners
This study focuses on confusions made by French L2 learners vs. native subjects in the perception of 11 audiovisual Mandarin Chinese attitudes, selected from a broader set of 19 attitudes previously evaluated in audio condition by both native Chinese and naïve French listeners. Two groups of French L2 learners of Mandarin Chinese were selected according to their level assessed by the Common European Framework of Reference for Languages: 9 beginners (A1) vs. 10 intermediate learners (A2). Subjects evaluated the 11 attitudes in audio, visual and audiovisual condition. Comparison of confusions between learners of level A1 vs. A2 indicates few significant differences, mostly in audiovisual condition and without a clear gain for one group over the other: confusions patterns are closer to the native reference for group A1 in expression of doubt, and for group A2 in expression of contempt. The comparison of French L2 learners pooled together vs. native speakers reference sheds light on major confusions to be targeted by specific methods and exercises. In audio-only condition, neutral surprise and politeness are less recognized by learners, who confuse contempt with question and question with obviousness. In visual-only condition, obviousness is more confused with declaration, contempt with irritation, and disappointment with doubt. In audio-visual condition, recognition of neutral surprise is lower, while infant-directed speech is better recognized; neutral surprise is more confused with irritation and contempt with disappointment. Cross-modality comparisons suggest a limited contribution of informations conveyed by acoustic prosody in the identification of audiovisual social affects by L2 learners.

1.2.20 p.174

Nuzha Moritz, Christophe Damour,
Coordination between gesture and prosody in two versions of"The Great Gatsby": 1974, 2013"
The cross-disciplinary study (phonetics and film study) aims at highlighting the coordination between posture and prosody in two versions of "The Great Gatsby". The central aim of the study is to understand how prosodic variations are related to gesture in different acting schools. Formal and functional analysis of gesture and their relation to prosody, shows striking contrast between the acting styles

1.3 Tuesday Session Three - Poster

- Prominence & Phrasing -

1.3.1 p.183

Fred Cummins, Judit Varga,
Explorations in the prosodic characteristics of synchronous speech, with specific reference to the roles of words and stresses
We examine the prosodic characteristics of read speech produced alone or in synchrony with a co-speaker in English. Previous work has demonstrated a marked difference between these two speaking conditions in Mandarin, but not English. We employ word lists that are either simple sequences of trochees, or

complex lists with regular stress alterantion but irregular word boundaries. Inter-onset intervals are examined and no major differences between solo and synchronous interval sequences are found. Viewed from the perspective of two generative models, however, there is weak evidence for some small difference in the dependence of interval duration on serial position.

1.3.2 p.187

Zhen Qin, Annie Tremblay,
Effects of native dialect on Mandarin listeners' use of prosodic cues to English stress
This study investigates the effect of native dialect on the use of prosodic cues to English stress by Standard Mandarin (SM) listeners, Taiwanese Mandarin (TM) listeners, and English listeners. Both SM and TM use fundamental frequency (F0) to realize lexical tones, but only SM uses duration together with F0 to realize lexically contrastive full-full vs. full-reduced stress patterns. Native English listeners and second language learners of English who spoke SM or TM as native language and were at similar proficiencies in English completed a sequence-recall task. English disyllabic non-words that differed in stress placement were resynthesized to contain only F0 cues, only duration cues, or converging F0 and duration cues. The results showed that SM-speaking learners used duration more than TM-speaking learners to recall English non-words. Native dialect is suggested to be considered in second language speech processing models.

1.3.3 p.192

Caterina Petrone, Mariapaola D'Imperio, Susanne Fuchs, Leonardo Lancia,
The interplay between prosodic phrasing and accentual prominence on articulatory lengthening in Italian
The distribution of preboundary lengthening within the phrase-final word is controversial. In CV syllables immediately preceding a prosodic boundary, the acoustic duration of the syllable onset C is less involved than that of the following rime V in the lengthening phenomenon. Moreover, preboundary lengthening might be extended to the stressed/accented rime within the phrase final word. On the other hand, articulatory constriction gesture for the onset consonant can be lengthened despite not being immediately adjacent to a boundary. In this study, we explore the effects of prosodic boundary and prominence in Italian, at both acoustic and articulatory level. Bilabial consonants in CV onset position were examined. The consonants were inserted in unstressed (word final) and stressed (penultimate vs. antepenultimate) syllables occurring in the vicinity of prosodic boundaries of different levels. In final syllables, the acoustic duration of the onset consonant was not affected by the prosodic boundary manipulation whereas the closing gesture duration showed a pattern of lengthening which was stronger for higher level prosodic boundaries. In non-final syllables, no acoustic/articulatory effect was found for onset consonants but only on the stressed vowels in penultimate position. Structural, phonological and phonetic constraints might be at work in determining preboundary lengthening.

1.3.4 p.197

Francisco Torreira, Miquel Simonet, José Ignacio Hualde,
Quasi-neutralization of stress contrasts in Spanish
We investigate the realization and discrimination of lexical stress contrasts in pitch-unaccented words in phrase-medial position in Spanish, a context in which intonational pitch accents are frequently absent. Results from production and perception experiments show that in this context durational and intensity cues to stress are produced by speakers and used by listeners above chance level. However, due to substantial amounts of phonetic overlap between stress categories in production, and of numerous errors in the identification of stress categories in perception, we suggest that, in the absence of intonational cues, Spanish speakers engaged in online language use must rely on contextual information in order to distinguish stress contrasts.

1.3.5 p.202

Miquel Simonet, Joseph Casillas, Yamile Díaz,
The effects of stress/accent on VOT depend on language (English, Spanish), consonant (/d/, /t/) and linguistic experience (monolinguals, bilinguals)
This study examines Voice Onset Times of coronal stops in utterance-initial position in two languages.

Crucially, the effects of lexical stress (stressed, unstressed syllable) on VOT are analyzed. The study investigates aspirated stops (English /t/), short-lag voiceless stops (English /d/, Spanish /t/) and prevoiced stops (Spanish /d/). Three groups of speakers provide data: English monolinguals, Spanish monolinguals, and proficient Spanish-English bilinguals. The study finds that lexical stress lengthens aspiration (English /t/) and prevoicing (Spanish /d/) but it does not alter significantly short-lag stops (Spanish /t/, English /d/). Monolinguals and bilinguals differ slightly in their phonetic behavior. Implications for gestural coordination as well as for feature theory are discussed.

1.3.6 p.??

Bogdan Ludusan, Guillaume Gravier, Emmanuel Dupoux,
Incorporating Prosodic Boundaries in Unsupervised Term Discovery
We present a preliminary investigation on the usefulness of prosodic boundaries for unsupervised term discovery (UTD). Studies in language acquisition show that infants use prosodic boundaries to segment continuous speech into word-like units. We evaluate whether such a strategy could also help UTD algorithms. Running a previously published UTD algorithm (MODIS) on a corpus of prosodically annotated English broadcast news revealed that many discovered terms straddle prosodic boundaries. We then implemented two variants of this algorithm: one that discards straddling items and one that truncates them to the nearest boundary (either prosodic or pause marker). Both algorithms showed a better term matching F-score compared to the baseline and higher level prosodic boundaries were found to be better than lower level boundaries or pause markers. In addition, we observed that the truncation algorithm, but not the discard algorithm, increased word boundary F-score over the baseline.

1.3.7 p.212

Chiu Yu Tseng, Chao Yu Su,
Binary Contrast and Categorical DifferentiationProsodic Characteristics of English Word Stress in Broad and Narrow Focus Positions
Assuming that categorical differentiation is major acoustic characteristics of English lexical stress through binary instead of more complex 3-way distinction, we investigated lexical stress in broad and narrow focus positions and found how binary distinction is achieved by the concomitancy of secondary stress defined by its position and distance in relation to primary stress. Similar results are found in broad (sentence initial) and narrow focus as well. These results suggest that binary categorical contrast is the optimal choice while differentiation is dependent on robust contrast patterns in the speech signal.

1.3.8 p.217

Adrian Leemann, Marie José Kolly, Volker Dellwo,
Crowdsourcing regional variation in speaking rate through the iOS app 'Dialäkt Äpp'
It is a common stereotype in Switzerland that speakers from Bern speak slowly and speakers from Zurich speak quickly. Are these differences in perception at all mirrored in production? We present a new method of crowdsourcing speaking rate through a free of charge iOS application. Astonishingly, results indicate that the temporal structure of a few words alone as spoken by a few hundred speakers are sufficient to tell apart the two dialects in speaking rate. In line with previous literature, females articulate more slowly than males. Further potential fields of application of the introduced method are discussed.

1.3.9 p.222

Núria Esteve-Gibert, Ferran Pons, Laura Bosch, Pilar Prieto,
Are gesture and prosodic prominences always coordinated? Evidence from perception and production
This study explores the temporal coordination between gesture and speech by addressing two main questions: (1) Are speakers sensitive to the misalignment between gesture prominence and prosodic prominence? (2) Is this sensitivity modulated by the semantic information conveyed by gesture and speech modalities in production? Experiment 1 tested question (1) and Experiment 2 tested question (2). Results from Experiment 1 revealed that the combinations in which prominences were misaligned were less acceptable than combinations with aligned prominences, and that the metrical pattern of the target word had an effect on the speakers' sensitivity: unsynchronized trochees (with the gesture prominence at the

post-tonic syllable) were frequently accepted, while unsynchronized iambs (with the gesture prominence at the pre-tonic syllable) were rejected. Results from Experiment 2 revealed that when the pointing gesture adds information to speech, i.e. it is supplementary to speech, the prominences are frequently misaligned (with gesture occurring after the speech), as if two different speech acts were produced. These findings suggest that the semantic content of gesture-speech combinations might influence the speakers' sensitivity of the misalignment between prosodic and gesture prominences.

1.3.10 p.227

Stefan Baumann, Anna Roth,

Prominence and Coreference On the Perceptual Relevance of F0 Movement, Duration and Intensity

We conducted a web-based experiment on German testing the perception of an element's prosodic prominence in relation to its status as a potential coreferent of an antecedent. Data were elicited by asking subjects to judge the probability of a coreference relation between a context noun (antecedent) and a target word (anaphor), whose lexically stressed syllable was manipulated as to the parameters F0 movement, duration and intensity. Results suggest a direct but inverse relationship between prominence and coreference judgements indicating that the likelihood of a coreference interpretation decreases with increasing prosodic prominence. F0 movement turned out to be the dominant cue for prominence as the main trigger for the perception of pitch accents with rises being perceived as more prominent than falls. In turn, lack of tonal movement probably led to perceived deaccentuation and thus favoured the evaluation of a target word as being coreferential with an antecedent. Duration was found to be a significant factor as well, while intensity did not prove to be relevant for the task given. Thus, the present study with its revised methodology adds new aspects to the debate of which parameters are crucial for prominence perception, directly linking it to the investigation of information structure.

1.3.11 p.232

Pärtel Lippus, Eva Liina Asu, Mari-Liis Kalvik Mari,

An acoustic study of Estonian word stress

This study investigates the acoustic correlates of word stress in Estonian. It forms part of a broader international collaboration the aim of which is to develop a universal language independent model for evaluating lexical stress regardless of the phonological structure of a given language. To this aim the characteristics of word stress in a range of languages is studied using unified methodology. For the present study, four acoustic measures were analysed as a function of speaking style and stress: vowel duration, F0 mean, F0 standard deviation, and spectral emphasis. The results show that the strongest correlate of style and stress in Estonian is vowel duration, but stress has a strong interaction with the Estonian three-way quantity system.

1.3.12 p.236

Lenka Weingartová, Kristýna Poesová, Jan Volín,

Prominence Contrasts in Czech English as a Predictor of Learner's Proficiency

The study investigates prominence patterns in Czech-accented English comparing the production of non-native speakers of English at two distinct stages of phonological acquisition (beginners and intermediates) with a native performance. Word stress in Czech is entirely different from English, it has a fixed position, a delimitative function and rather impalpable acoustic manifestations. Alternations in the realization of word stress were analyzed by measuring the ratios or differences of acoustic correlates of prominence: duration, fundamental frequency, sound pressure level and spectral slope. Since word stress is a relational phenomenon, these characteristics were measured in two adjacent syllables one of which was a canonical stress bearer. The results reveal a clear difference between native and non-native treatment of word stress in all parameters examined. In the non-native sample distinct interferences of L1 across the two groups were detected: the subjects displayed different exploitation of duration, spectral slope and SPL with relation to their proficiency in L2 English. Out of these, duration ratio proves to be the most significant correlate. Furthermore, our findings indicate a strong effect of prosodic context coinciding with the prominence features, particularly in intonation declination and phrase-final lengthening.

1.3.13 p.241

Alice Turk, Stefanie Shattuck Hufnagel,
A sketch of an extrinsic timing model of speech production
In this paper, we motivate and present a sketch of an extrinsic timing model of speech production. It is a three-stage model, involving 1) a phonological planning stage, where symbolic segmental representations are sequenced and slotted into an appropriate prosodic structure, and where appropriate acoustic cues are selected for each segment in its context, and 2) a phonetic planning stage, where cues are mapped onto sets of articulators and appropriate values for spatial and temporal parameters of movement are computed, and 3) a motor-sensory implementation stage, where articulator movements are generated and tracked. We cite model components from the literature that accomplish many of the functions this type of model requires.

1.3.14 p.246

Nicolas Obin, Julie Beliao, Christophe Veaux, Anne Lacheret,
SLAM: Automatic Stylization and Labelling of Speech Melody
This paper presents SLAM : a simple method for the automatic Stylization and LAbelling of speech Melody. This main contributions over existing methods are : the alphabet of melodic contours is fully data-driven, an explicit time-frequency representation is used to derive complex melodic contours, and melodic contours can be determined over arbitrary prosodic/syntactic units. Additionally, the system can handle some specificities of spontaneous speech (e.g., multi speakers, speech turns and speech overlaps). A preliminary experiment conducted on 3 hours of spoken French indicates that a small number of contours is sufficient to explain most of the observed contours. The method can be easily adapted to other stressed languages. The implementation is open-source and freely available.

1.3.15 p.251

Antonio Simoes,
Lexical Stress in Brazilian Portuguese in Contrast with Spanish
This study discusses stress assignment in prosodic, non-verbal words in Brazilian Portuguese, in comparison with descriptions of stress assignment for Spanish [9, 13, 15, 16, 17, 18]. Given the conflicting claims regarding stress assignment in Brazilian Portuguese (see [11, 1, 2, 10, 3]) there is still a need to revisit discussions on stress assignment in Portuguese. In general, stress assignment in Spanish has been satisfactorily explained through the interplay between the morphological and phonological domains. Similar descriptions for Portuguese still requires far more abstraction and use of artifacts than in Spanish, which makes Mattoso Cmara Jr.'s [4, 5] claim that lexical stress is unpredictable in Brazilian Portuguese surprisingly unchallenged.

1.3.16 p.256

Rena Nemoto,
Prosodic Characteristics of Vocalic Hesitations in Comparison with Overlong Vowels in Estonian
The goal of this paper is to investigate vocalic hesitations in Estonian and compare them to the related vowels of overlong (Q3) quantity degree. We wonder if there are some languagespecific characteristics of hesitations. If yes, which kind of characteristics can be observed in Estonian language? We analyze duration, fundamental frequency (f0), intensity, and first two formants using 39.5 hours of manually transcribed monoor dialogue speech from a spontaneous speech corpus. Investigated vocalic hesitations and Q3 vowels are: /ee, ää, aa, , öö/. The characteristics of hesitations as compared to those of Q3 vowels show that hesitations have longer duration range. Hesitations generally include lower f0 and intensity values. However, the values vary in terms of vowels. First two formants of hesitations tend to be located at more centralized positions in a vocalic triangle than related Q3 vowels.

1.3.17 p.261

Alexsandro Meireles, Plínio Barbosa,
Articulatory Reorganizations of Speech Rhythm due to Speech Rate Increase in Brazilian Portuguese
This paper examines how speech rate increase acts to change speech rhythm at the articulatory level.

Main results show that speech rate increase worked to change articulatory parameters in the following way: a) decrease of acceleration duration; b) decrease of y-extremum; c) decrease of constriction displacement; d) decrease in modulus of peak and/or valley velocity; e) decrease of gestural duration; and f) constant proportional time-to-peak (or valley) velocity. Besides, results have shown that speech rate tends to affect all gestures in an utterance independently of their phrasal position. Nevertheless, there was evidence that some articulatory parameters could, if properly manipulated, provide cues for rhythmic restructurings in speech. Finally, results show that the dynamical speech rhythm model (Barbosa, 2007) is more appropriate to deal with Brazilian Portuguese acoustical data than the pi-gesture model (Byrd &Saltzman, 2003), and that both models could explain articulatory reorganizations due to speech rate increase.

1.3.18 p.265

Sabine Zerbian, Jane Kühn, Christoph Schroeder, Svenja Schuermann,
Prosody in Turkish learners of German as a Foreign Language
Results of a pilot study are reported which investigates the prosodic realization of information structure by six learners of German as a Foreign Language (GFL) with Turkish as first language. Question-answer pairs were read out loud which systematically varied the position of narrow focus in the response by means of a preceding wh-question. A qualitative analysis of the results shows deaccentuation of postfocal constituents in the case of subject focus for 4/6 GFL speakers but no consistent pitch increase on focused constituents. Two speakers did not change prosody due to information structure. The results are discussed in connection with the acquisition of prosody as a marker of information structure. Deaccentuation has been reported to cause problems in L2 prosody. In Turkish, deaccentuation occurs postfocally. The claim will be motivated that the occurrence of deaccentuation in the L1 is a necessary but not sufficient condition for early acquisition of deaccentuation in a foreign language.

1.3.19 p.270

Sandrine Brognaux, Thomas Drugman, Marco Saerens,
Synthesizing sports commentaries: One or several emphatic stresses?
Emphatic stresses are known to fulfill essential functions in expressive speech. Their integration in speech synthesis usually relies on a prosodic annotation of the training corpus. Emphasized syllables are then assigned a single label or can receive several labels according to their acoustic realization. While it is more complex to predict those various labels for a new text to synthesize, it might allow for a better rendering of the stress in the synthesized speech. This paper examines whether the use of more than one emphatic label improves the perceived expressivity of the synthesized speech. It relies on a manually-annotated expressive corpus of sports commentaries. Statistical acoustic analyses show that four distinct realizations of emphatic stresses can be distinguished. However, perceptual tests indicate that the integration of this distinction in HMM-based speech synthesis does not lead to a significant improvement in expressivity. This seems to imply that the different acoustic realizations of the stress are not required to be explicitly annotated in the training corpus.

1.3.20 p.275

Jürgen Trouvain, Bernd Möbius,
Sources of variation of articulation rate in native and non-native speech: comparisons of French and German
Speech tempo including articulation rate is often considered as a good predictor in the diagnosis of foreign language proficiency and its comprehension. In this study we investigate various sources of variation of articulation rate such as the L2 proficiency level, individual tempo habits in L1 and L2, and more extensive exposure to native speech. In addition, we also discuss the difficulty of the most informative unit for rate metrics which allows comparisons between French and German. The materials used are French and German read sentences, produced as L1 and L2 speech. In contrast to other studies individual habits of articulation rate in the L1 was only partially observed in the corresponding L2 data (a slow L1 speaker does not necessarily articulate slowly in the L2). The convergence of most French learners to the German model speakers shows the advantage of having additional input for phonetic exercises. The fastest German learners also converge to the rather slow French model speaker.

1.3.21 p.280

Helena Moniz, Ana Isabel Mata, Julia Hirschberg, Fernando Batista, Andrew Rosenberg, Isabel Trancoso,

Extending AuToBI to prominence detection in European Portuguese

This paper describes our exploratory work in applying the Automatic ToBI annotation system (AuToBI), originally developed for Standard American English, to European Portuguese. This work is motivated by the current availability of large amounts of (highly spontaneous) transcribed data and the need to further enrich those transcripts with prosodic information. Manual prosodic annotation, however, is almost impractical for extensive data sets. For that reason, automatic systems such as AuToBi stand as an alternate solution. We have started by applying the AuToBI prosodic event detection system using the existing English models to the prediction of prominent prosodic events (accents) in European Portuguese. This approach achieved an overall accuracy of 74% for prominence detection, similar to state-of-the-art results for other languages. Later, we have trained new models using prepared and spontaneous Portuguese data, achieving a considerable improvement of about 6% accuracy (absolute) over the existing English models. The achieved results are quite encouraging and provide a starting point for automatically predicting prominent events in European Portuguese.

1.3.22 p.285

Fabio Tamburini, Chiara Bertini, Pier Marco Bertinetto,

Prosodic prominence detection in Italian continuous speech using probabilistic graphical models

Prosodic prominence, a speech phenomenon by which some linguistic units are perceived as standing out from their environment, plays a very important role in human communication. In this paper we present a study on automatic prominence identification using Probabilistic Graphical Models, a family of Machine Learning Systems able to properly handle sequences of events. We tested the most promising members of such models on utterances selected from a manually annotated Italian speech corpus, obtaining very good recognition results crucially converging with the prominence detection responses provided by a pool of native speakers.

1.3.23 p.290

Robert Fuchs,

Integrating variability in loudness and duration in a multidimensional model of speech rhythm: Evidence from Indian English and British English

Most research on speech rhythm has focussed on duration. For example, [1] suggested the normalised Pairwise Variability Index for vocalic intervals (nPVI-V) in order to measure the variability of vocalic durations. This paper argues that speech rhythm research should also take into account other correlates of prominence as well as their interaction. The duration-based nPVI, or nPVI-V(dur), is supplemented by an nPVI-V(avgLoud) that measures variability in average loudness. These two metrics account for variability in duration and loudness, but cannot measure if loudness and duration reinforce each other by varying simultaneously in the same direction. This simultaneous variability is accounted for by the combined nPVI-V(dur+avgLoud), which is higher than the average of the other two measures, if vocalic intervals that are longer than average are also louder than average. The three metrics are subsequently applied to recordings of a reading task performed by 20 speakers of Indian English (IndE) and 10 speakers of British English (BrE). Results indicate that IndE has less variability in duration and less variability in loudness than BrE. In addition, IndE has less simultaneous variability in duration and loudness than BrE. This indicates that duration and loudness are less often used together as cues to prominence in IndE than in BrE.

1.3.24 p.295

Hongwei Ding, Rüdiger Hoffmann,

A Durational Study of German Speech Rhythm by Chinese Learners

This study focuses on the temporal and metrical features of the German speech produced by Chinese speakers. German is described to be a stress-timed language, while standard Chinese is regarded as a syllable-timed language. It has been suggested that the rhythm of the target language can be influenced by the learners' native language. In this study we conducted an investigation of ten sentences with 18

Chinese students in the low intermediate proficiency level in comparison with six native German speakers. We compared the duration values in terms of pairwise variability indices, and found that most of these Chinese speakers have a lower nPVI-V and a higher rPVI-C than the German speakers. We illustrate that the conventional duration measures of nPVI-V can be influenced by the syllable structures of the utterance and the classification approach of vocalic intervals, and a comparable nPVI-V can hardly be expected from different investigations. Furthermore, we argue that duration values alone cannot fully capture the rhythmic patterns of speech because other prosodic parameters such as pitch and energy also join to contribute to rhythmic characteristics of the speech.

1.3.25 p.300

Donna Erickson, Shigeto Kawahara, J.C. Williams, Jeff Moore, Atsuo Suemitsu, Yoshiho Shibuya,
Metrical Structure and Jaw Displacement: An Exploration
The current experiment using EMA shows that the amount of jaw displacementor mandible movement-may reflect the metrical organization of English sentences. The experiment also supports F1 as a reliable acoustic correlate of jaw displacement, hence metrical organization, but also demonstrates that F0 does not have a similar relationship to mandible movement.

1.3.26 p.305

Erwan Pépiot,
Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers
Many studies have been conducted on acoustic differences between female and male speech. However, they have generally been led on speakers of only one language, and have focused on a single acoustic parameter. The present study is an acoustic analysis of dissyllabic words or pseudo-words produced by 10 Northeastern American English speakers (5 females, 5 males) and 10 Parisian French speakers (5 females, 5 males). Several prosodic parameters were measured: mean f0, f0 range, phonation type (through H1-H2 intensity differences) and words' duration. Significant cross-gender differences were obtained for each tested parameter. Moreover, cross-language variations were observed for f0 range, and H1-H2 differences. These results suggest that cross-gender acoustic differences are partly language-dependent and could be socially constructed.

1.3.27 p.310

Hansjörg Mixdorff, Angelika Hönemann, Oliver Niebuhr, Christoph Draxler,
Perceived Prominence Reflected by Imitations of Words with and without F0 Continuity
This paper continues our work on the perception of prominence as a function of F0 continuity. In an earlier study the first author had shown that F0 intervals occurring at lexically stressed syllables and measured using the amplitude of Fujisaki model accent commands strongly contribute to the perceived prominence of that syllable. More recent work explored how F0 continuity influenced prominence ratings of single word utterances. The outcome indicated that listeners made use of the physically available F0 information and therefore words containing gaps in the contour were perceived as less prominent. It was also shown that subjects were able to interpolate missing parts as long as the F0 peak was still present. The current study explores whether subjects compensate the lack of prominence in words containing F0 gaps by asking them to produce a word with the same accent strength as that of a spoken word stimulus, the spoken word being either the same or different from the one they are asked to utter. We evaluated word durations, F0 intervals and intensities of the responses as correlates of prominence and found that listeners indeed seem to adjust depending on the kind of stimulus they have heard.

1.3.28 p.315

Toshiyuki Sadanobu,
The Structure of Japanese Phrase in Accordance with Speaking Modes
While English is often spoken in an increment of clause (i.e. subject and predicate), Japanese of a smaller phrase called "bunsetsu" (e.g. noun phrase and case particle). Previous studies on Japanese language, however, have traditionally been focusing on clause structure, and little attention has been paid on the

structure of “bunsetsu” (non-predicate one, especially). This paper describes the basic structure of non-predicate “bunsetsu” from grammatical point of view, and elucidates that the structure of non-predicate “bunsetsu” varies in accordance with four speaking modes ((i) Sentence mode A; (ii) Sentence mode B; (iii) “Bunsetsu” mode; and (iv) Character mode), which are identified on the criteria of compatibility among seven phenomena attested in Japanese speech. To be more concrete, this paper shows that it is only the mode (iii) that enables copula, “bunsetsu”-final particle (“Kantoujoshi” in Japanese), final leaping, and combination of breaking and prolongation in non-predicate “bunsetsu”).

2 Day Two - May 21st

2.1 Wednesday Session One

May 21st, 9am - 10:30am : 2-1-plenary (1+3 presentations)

2.1.1 KeyNote 2 (p.64)

Stefanie Shattuck-Hufnagel 30-min
Cue-based analysis of speech: implications for prosodic labelling systems

Over the past few decades it has become clear that an adequate account of systematic context-driven variation in word forms requires representations below the level of the abstract symbolic phoneme or even the allophone. One proposal for this sub-allophonic level of description is in terms of feature cues, such as the cues to articulator-free features and articulator-bound features proposed by Halle (1992) and by Stevens (2002), also assumed in the concept of enhancing cues in Stevens and Keyser (2010), Keyser and Stevens (2006) and Stevens, Keyser and Kawasaki (1986). This proposal of a level of representation of discrete feature cues, along with continuous-valued cue parameters, has the potential to bridge the gap between abstract symbolic categories of the phonology and the concrete spatial and temporal specifications that drive the articulatory-acoustic implementation of word forms in continuous communicative speech. Such an approach suggests that phonetic transcription might benefit from a focus on capturing the individual cues to feature contrasts that are realized in the speech signal. Does this approach to understanding phonetic variation in word forms have implications for prosodic labelling? We will explore this possibility, taking as our point of departure Arbisi-Kelm’s (2006) proposal for labelling the separate correlates of prosodic disfluency in stuttered speech, adapted by Brugos and Shattuck-Hufnagel (2012) for prosodic disfluencies in utterances produced by typical speakers. Our hypothesis is that variation in cue selection and cue parameter values is systematically governed by context, and that cue-level transcription may be needed to capture systematicity in the phonetic implementation of prosodic phonology as well as of lexical phonology.

2.1.2 p.321

Yuki Asano,
Stability in perceiving non-native segmental length contrasts
Previous studies have demonstrated that listeners show high sensitivity in discriminating non-native prosodic contrasts thanks to auditory memory (Hayes and Masuda 2008; Hirano 2011). We tested the limits of discriminating Japanese consonantal length contrasts with three groups of listeners (German learners of Japanese, German non-learners and Japanese natives) under increasing task demands. We increased auditory memory load through a longer inter-stimulus interval (=ISI) (2500ms vs. 300ms) and added psycho-acoustic complexity (trials with task-irrelevant pitch falls that occurred simultaneously with the consonant vs. with monotonous pitch). Results showed very good discrimination in all groups when task demands were lowest. With increasing task demands, only non-natives’ discrimination abilities decreased: non-learners were strongly affected by both ISI and pitch, while learners only by pitch. The psycho-acoustic complexity of the stimuli had a stronger impact on performance than increased memory load. Our findings suggest that L2 learners can establish novel phonological representations, but the ability to use them can be applied still only under favorable listening conditions with no distracting acoustic information. The non-native listeners’ reduced sensitivity under increasing task demands appears to be the reason why even advanced learners still face difficulties in natural learning situations.

2.1.3 p.326

Jessica Siddins, Jonathan Harrington, Ulrich Reubold, Felicitas Kleber,
Investigating the relationship between accentuation, vowel tensity and compensatory short-

ening
The aim of this study was to investigate the relationship between compensatory shortening and coarticulation in German tense and lax vowels and to determine whether this relationship was influenced by prosodic accentuation. While previous studies focussed on temporal vowel reduction due to compensatory shortening, and often found conflicting results, our study extends previous results by including a formant analysis of spatial reduction in two types of compensatory shortening. Polysyllabic shortening was tested in monosyllabic versus disyllabic words, while incremental coda shortening was tested in words with final singleton versus final cluster. Speakers produced minimal pairs differing in vowel tensity in accented and deaccented contexts for both shortening conditions. Vowel duration was influenced primarily by vowel tensity as well as by accentual lengthening for tense but not lax vowels. While vowel duration was not affected by compensatory shortening, formant analyses revealed an effect of coda cluster for tense vowels as well as clear effects of accentuation and vowel tensity. There was no effect of polysyllabic shortening on formants. Further to previous studies on compensatory shortening, these results reveal that compensatory shortening is not limited to temporal reduction, but can have an impact on vowel quality as well.

2.1.4 p.331

Amanda Ritchart, Amalia Arvaniti,
The form and use of uptalk in Southern Californian English
This study examines the phonetics, phonology and pragmatic function of uptalk, utterance-final rising pitch movements, as used in Southern Californian English. Twelve female and eleven male speakers were recorded in a variety of tasks. Instances of uptalk were coded for discourse function (statement, question, confirmation request, floor holding) based on context. The excursion of the pitch rise and the distance of the rise start from the onset of the utterance's last stressed vowel were also measured. Confirmation requests and floor holding showed variable realization. Questions, on the other hand, showed a rise that typically started within the stressed vowel and had a large pitch excursion, while uptalk used with statements exhibited both a smaller pitch excursion and a later rise that often started after vowel offset. This pattern suggests that statements have a L* L-H% melody while questions have L* H-H%. Gender differences were also found: female speakers used uptalk more often than males, and showed greater pitch excursion and later alignment, all else being equal. Other social parameters, however, such as social class and linguistic background did not affect the use of uptalk.

2.2 Wednesday Session Two

11am - 1pm : 2-2-oral (6 presentations)
- speech rhythm and timing -

2.2.1 p.337

Agnieszka Wagner,
Rhythmic structure of utterances in native and non-native Polish
This paper presents results of an ongoing study concerning speech rhythm in native and non-native Polish. The goal of the analyses described in the paper was to characterize rhythmically Polish utterances realized by native and non-native speakers with German and Korean accent. The analyses are limited to the domain of duration, but in the future other prosodic parameters will also be investigated. In the current study, different rhythm metrics (%V, V, C, PVIs and Varcos) were applied to provide quantitative description of temporal patterning in native and non-native Polish. Following the assumption that perceived speech rhythm is the effect of meter and grouping which are closely related to prominence and phrasing, durational marking of various levels of prominence and prosodic edges was also analyzed between the three accents (native Polish and German- and Korean-accented Polish). The analyses aimed also at rhythmic classification of Polish - for that purpose the results of quantitative description with rhythm metrics and phonotactic properties of the speech material used in the current study were compared with the data for other languages presented in the literature.

2.2.2 p.342

Hugo Quené, Rosemary Orr,
Long-term convergence of speech rhythm in L1 and L2 English

When talkers from various language backgrounds use L2 English as a lingua franca, their accents of English are expected to converge, and talkers' rhythmical patterns are predicted to converge too. Prosodic convergence was studied among talkers who lived in a community where L2 English is used predominantly. Speech rhythm was operationalized here as the peak frequency in the spectrum of the intensity envelope, normalized to the speaking rate (in syll/s). Results indicate that talkers produced intensity contours with maximum periodicity at frequencies of about 0.32 times their syllable rates, i.e., peaks in intensity tend to occur every 1/0.32 syllables. These results were collected repeatedly, from 5 recordings conducted over 3 years with the same talkers. We found that variance between talkers in their rhythm decreases over time, thus confirming the predicted convergence in speech rhythm in L2 English. These findings show that speech rhythm in L2 English tends to converge, and that this prosodic convergence continues to proceed over several years, as well as over communicative settings.

2.2.3 p.346

Andreas Windmann, Juraj Šimko, Petra Wagner,
Probing Theories of Speech Timing using Optimization Modeling
We implement two theories about the temporal organization of speech in an optimization-based model of speech timing and conduct simulation experiments in order to test whether both theories can account for the phenomenon of foot-level shortening (FLS) observed in English speech corpora. Results suggest that a model that induces compensatory timing relations between syllables and feet predicts empirical results very accurately. However, we also observe that the FLS effect can equally well be explained under the assumption that suprasegmental timing is confined to localized lengthening effects at the heads and edges of prosodic domains. Implications for theories of speech timing are discussed.

2.2.4 p.351

Sandra Peters, Felicitas Kleber,
The influence of accentuation and onset complexity on gestural timing within syllables
This paper presents results from a production experiment using electromagnetic articulography. The main aim of the study was to investigate how phrasal accent and the number of onset consonants influence the gestural timing of syllable constituents in German. Five speakers of German with sensors attached to the tongue tip, tongue body and lower lip were recorded reading sentences with either accented or unaccented target words that contained simplex (one consonant) and complex (two consonants) onsets. The nucleus was always /a/ and the coda consonant was always /p/. We analyzed acoustic segment duration and gestural overlap (in terms of lag measurements). Onset complexity influenced both CV and VC overlap and accentuation affected gestural overlap to a greater extent than acoustic vowel duration. However, the extent of overlap differed between segment sequences and accentuation patterns: while for CC and VC sequences trends for greater overlap in deaccented than in accented condition were found, CV overlap decreased with deaccentuation. Shorter plateau durations in this context explain the diminished CV overlap in a prosodically weak context. The findings are discussed with respect to the predictions made by articulatory phonology regarding gestural timing and with respect to timing stability in weak versus strong prosodic contexts.

2.2.5 p.356

Núria Esteve-Gibert, Joan Borrs-Comes, Marc Swerts, Pilar Prieto,
Head gesture timing is constrained by prosodic structure
There is an increasing consensus to regard gesture and speech as parts of an integrated communication system, in part because of the findings related to their temporal coordination at different levels. In general, results for different types of gestures show that the most prominent part of the gesture (the apex) is typically aligned with accented syllables. The aim of the present study is to test for this coordination by focusing on head movements taken from a semi-spontaneous setting in order to look at the effects of upcoming phrase boundaries on their timing. Our results show that while apexes of head gestures are synchronized with accented syllables, upcoming phrase boundaries have an effect on the timing of three gestural points, namely the start, apex, and end time of head gestures. Crucially, these points are aligned differently with respect to the stressed syllable for trochees as compared with iambs/monosyllables, showing that head nods are retracted before upcoming phrase boundaries. This result corroborates previous results by Esteve-Gibert &Prieto for pointing gestures in laboratory settings.

2.2.6 p.361

Helen Türk, Pärtel Lippus, Karl Pajusalu, Pire Teras,
The ternary contrast of consonant duration in Inari Saami
The three-way distinction of quantity occurs in several Finnic and Saami languages. The paper focuses on the length contrast of consonants in Inari Saami. Similarly to Estonian and other Finno-Ugric languages where three quantities are described, in Inari Saami the distinction between single consonants, short geminates or consonant clusters, and long geminates or consonant clusters appears only on the boundary of a stressed and unstressed syllable of a disyllabic foot. Our results show that in Inari Saami the duration of consonants is inversely related to the duration of both preceding and following vowels, and there is a tendency towards foot isochrony. The results are in line with previous studies on quantity opposition in Inari Saami and in other Finnic languages, showing the ternary distinction of consonant quantities as a foot-level feature of the language.

2.3 Wednesday Session Three

2pm - 4pm : Special Session - Slavic Prosody

2.3.1 p.366

Štefan Beňuš - 30-min (invited)
Slovak prosody in the phonetics-phonology debate: Yers and emergent prosodic breaks.
Prosody is central for understanding the cognitive system underlying human speech and relates to both more granular aspects of our phonological competence as well as more continuous aspects of observable articulatory movements and resulting acoustic characteristics. The understanding, and formal treatment, of the relationship between these two inter-related components of human speech is at the core of the cognitive approach to speech. In this presentation I contribute to this discussion by drawing links between two seemingly unrelated lines of my research on Slovak, and argue that understanding the continuous prosodic nature of speech is critical for improving our understanding of cognitive competence underlying it. The first aspect concerns yer vowels as the prototypical problem of Slavic phonology, the second involves the nature of prosodic boundaries.

2.3.2 no paper

Tamara Rathcke - 30-min (invited)
Time and timing in intonational phonology: analysing pitch categories in Russian (and other Slavonic languages)
Many Slavonic languages are still lacking a comprehensive description of their intonational phonologies. Given that decisions regarding the number of relevant categories and their types are the key issues of any phonological analysis, this presentation will concentrate on how time and timing can inform intonational phonology. Evidence from Russian (and also Bulgarian, Czech and Polish) will demonstrate that time pressures arising from intermittent voicing and an upcoming phrase boundary have different effects on Slavonic vs. Germanic languages. Potential implications of these findings for prosodic typology will be discussed.

2.3.3 p.368

Jaye Padgett, - 30-min (invited)
On the origins of the prosodic word in Russian
The Prosodic Word is a foundational notion in phonological theories, being relevant for the statement of many phonological generalizations. In spite of their importance, there are basic open questions about prosodic words. Where do they come from? Can their structure in one language vs. another be predicted? In this paper I suggest a research program that attempts to address such questions by viewing prosodic words as emergent over time from the interaction of phonetics, phonologization, and syntactic structure.

2.3.4 p.372

Bistra Andreeva, Jacques Koreman, William Barry,
Local and Global Acoustic Correlates of Information Structure in Bulgarian
In this study the prosodic exponents of information structure are examined in the production of six Bulgarian sentences under different focus conditions (broad focus and non-contrastive and contrastive

narrow focus). Results show that speakers consistently discriminate broad and narrow focus by both local and global acoustic cues. Local cues are the phonetic properties of the accented syllables, while global cues reflect broader phonetic patterns in the intervals before and after the accented syllable, which vary independently of the tonal accent. Contrastive and non-contrastive accents are differentiated exclusively by local cues, but only when the focus is early in the sentence.

2.3.5 p.377

Agnieszka Wagner,
Description of Polish speech rhythm using rhythm metrics and time-delay approach: A comparative study
The goal of this study is to provide a multidimensional description of rhythmic structure of Polish utterances. For this purpose a time-delay approach proposed in [1] is applied and results of qualitative and quantitative analyses based on time-delay plots are compared with results obtained with selected rhythm metrics. The study shows that description that relies on a combination of rhythmic scores is inconclusive and difficult to interpret, because it does not account for rhythmic structuring nor grouping. The time-delay approach, on the contrary, appears to be very efficient in exploring short-time and long-term timing variability that determines Polish speech rhythm.

3:40pm - 4:00: Panel discussion

2.4 Wednesday Session Four

4:30pm - 6pm : 2-4-poster (48 presentations)

- perception and intonation -

2.4.1 p.383

Marion Aguilera, Radouane El Yagoubi, Robert Espesser, Corine Astésano,
Event-Related Investigation of Initial Accent Processing in French
This study investigates stress processing through the Event-Related brain Potential (ERP) technique. It aims at evaluating whether French listeners can perceive and discriminate the Initial Accent (IA) and whether IA is encoded in the phonological representation. Participants listened to trisyllabic words in two stress-pattern conditions, with (+IA) or without (-IA) initial accenting, in an oddball paradigm. The EEG was recorded in both a passive and an active listening task, and in two different oddball versions: one where standard stimuli were +IA words and deviants -IA words, and the reverse for the other version (-IA standard, +IA deviant). Behavioral results show faster processing and less errors for +IA stimuli. ERP results show larger MisMatch Negativity component for -IA words, pointing out 1) that French listeners are sensitive to f0 manipulation, and 2) that +IA is the preferred stress template in French. Altogether, our results indicate that French listeners not only discriminate stress patterns but that IA is encoded in long-term memory, hence phonologically relevant.

2.4.2 p.388

Alejna Brugos, Jonathan Barnes,
Effects of dynamic pitch and relative scaling on the perception of duration and prosodic grouping in American English
Results of two perception experiments suggest that using timing measures alone to compute prosodic structure misses valuable information from pitch. Previous research showed that pitch can distort perceived duration: tokens with dynamic or higher f0 are perceived as longer than comparable level-f0 or lower-f0 tokens, and silent intervals bounded by tokens of widely differing pitch are heard as longer than those bounded by tokens closer in pitch (the kappa effect). Phrase edges (signalled by increased duration, pause, phrase tones, and f0 reset) set the scene for pitch to modulate perceived duration. Two new experiments used the same duration and f0 manipulations (level vs. varying-slope rises, at varying pitch ranges) of segmentally-identical base files, in two separate tasks: 1) a linguistic grouping task using an ambiguously-structured phrase and 2) a psychoacoustic study on perceived duration. Results show that effects on perceived duration due to dynamic pitch can be either strengthened or nullified depending on relative scaling of compared tokens. These same manipulations push grouping judgments beyond what

would be expected from distortions of perceived duration. This suggests that listeners integrate pitch and timing cues when judging linguistic structure, supporting measures of relative boundary size that combine duration and pitch measures.

2.4.3 p.393

Canan Ipek, Sun-Ah Jun,
Distinguishing Phrase-Final and Phrase-Medial High Tone on Finally Stressed Words in Turkish
The goal of this paper is to investigate the nature of the high tones realized on finally stressed words in Turkish. Following Ipek &Jun's [1] AM model of intonational phonology of Turkish, it was hypothesized that the high tone realized on the last syllable of a phrase (i.e., Intermediate Phrase (ip)) is realized differently from that of a phrase-medial prosodic word (PW), reflecting the prosodic hierarchy. Acoustic data show that an ip-final High tone shows larger f0 rise than a PW-final High tone, and the ip-final syllable is longer than the PW-final syllable. Furthermore, the degree of coarticulation is weaker across an ip boundary than a PW boundary. These findings support the prosodic structure and tonal categories proposed in Ipek &Jun's [1] model of Turkish intonation.

2.4.4 p.398

Willemijn Heeren, Vincent van Heuven,
The interaction of accent and boundary tone in perception of whispered speech
We investigated how the perception of Dutch whispered boundary tones depends on the presence of an accent in the utterance-final word, i.e. the boundary tone landing site. Listeners performed near ceiling in normal speech, whereas the same listeners' performance dropped about 30% in whisper, while processing speed decreased in whisper compared to normal speech. Accent position furthermore influenced boundary tone perception. Initial-stress words showed a question bias that affected recognition of that speech act when accent and boundary tone did not coincide. On final-stress words, in which boundary tone and accent coincided, statements and questions were identified equally well.

2.4.5 p.403

Katarina Bartkova, Denis Jouvet,
Links between Manual Punctuation Marks and Automatically Detected Prosodic Structures
This paper presents a study of the links between punctuation and automatically detected prosodic structures, as observed on large speech corpora that were manually annotated during speech transcription evaluation campaigns in French. These corpora contain more than 3 million words and almost 350 thousands punctuation marks. The detection of the prosodic boundaries and of the prosodic structures is based on an automatic approach that integrates little linguistic knowledge and mainly uses the amplitude and the inversion of the F0 slopes as described in [1], as well as phone durations. The paper first analyzes the occurrences of the punctuation marks with respect to various sub-corpora, which also highlights the variability among annotators. Then, the paper focuses on analyzing prosodic parameters with respect to the punctuation marks, followed or not by a pause, and on analyzing the links between the automatically detected prosodic structures and the manually annotated punctuation marks.

2.4.6 p.408

Timo Roettger, Rachid Ridouane, Martine Grice,
Perception of Peak Placement in Tashlhiyt Berber
Previous production studies on Tashlhiyt Berber have demonstrated that questions and statements have similar intonation contours, i.e., a final rise to a F0 peak and subsequent fall. The contours tended to differ in overall pitch register and peak location: questions (a) revealed a stronger tendency to be realized with the F0 peak on the final syllable than statements and (b) even within the same syllable, peaks were often aligned later in questions than in statements. The peak location, however, was reported to vary strongly both within and across speakers, interpreted as free alternation of tonal association. Given this high degree of variation, the question arises as to how relevant this variation is for communication. The present perception study shows that both pitch register (low vs. high) and tonal placement (peak on

penultimate vs. final syllable) affect listeners' judgments on sentence modality as well as reaction times. Whereas peak alignment within the syllable (early vs. late) did not affect judgments, it did have an effect on reaction times. By demonstrating their perceptual impact, this study confirms that the patterns found in production are communicatively relevant.

2.4.7 p.413

Cristel Portes, Uwe Reyle,
The meaning of French"implication"contour in conversation
French intonational contours inventory has a rising-falling tune which presents very interesting semantic properties. It has been called "intonation d'implication" by Delattre (1966) suggesting that the contour triggers an implicit meaning, i.e. an implicature in Gricean terms. Besides, the "implication" contour have been claimed to convey various attitudinal meanings from obviousness to exasperation, and also to mark contrastive focus. The aim of the present paper is to give a unified account of these seemingly differing semantic descriptions of the "implication" contour in French, using a dynamic semantic framework, namely Discourse Representation Theory (DRT). We claim that the main semantic component of the "implication" contour is to convey a contradiction (or a contrast). We first present our DRT-theoretical approach, and then apply it to occurrences of the "implication" contour in a corpus of conversational dialogue.

2.4.8 p.418

César González Ferreras, Carlos Vivaracho-Pascual, David Escudero-Mancebo, Valentín Cardeñoso-Payo,

Combination of variations of pairwise classifiers applied to multiclass ToBI pitch accent recognition
In this paper we present some experiments on multiclass ToBI pitch accent classification. The system is based on the fusion of pairwise classifiers, which are specialized in the distinction of pairs of prosodic labels. Several machine learning techniques, including neural networks, decision trees and support vector machines, are combined in different ways in order to find the best overall combination. Variations of pairwise classifiers are introduced in order to take into account the influence of the samples of the remaining classes during the training of the binary classifiers. The use of these techniques allowed us to improve the results, both the overall classification accuracy and the balance across the different ToBI pitch accent classes.

2.4.9 p.423

Aoju Chen,
Production-comprehension (A)Symmetry: individual differences in the acquisition of prosodic focus-marking
Previous work based on different groups of children has shown that four- to five-year-old children are similar to adults in both producing and comprehending the focus-to-accentuation mapping in Dutch, contra the alleged production-precedes-comprehension asymmetry in earlier studies. In the current study, we addressed the question of whether there are individual differences in the production-comprehension (a)symmetricity. To this end, we examined the use of prosody in focus marking in production and the processing of focus-related prosody in online language comprehension in the same group of 4- to 5-year-olds. We have found that the relationship between comprehension and production can be rather diverse at an individual level. This result suggests some degree of independence in learning to use prosody to mark focus in production and learning to process focus-related prosodic information in online language comprehension, and implies influences of other linguistic and non-linguistic factors on the production-comprehension (a)symmetricity.

2.4.10 p.428

Sandrine Brognaux, Thomas Drugman,
Phonetic variations : Impact of the communicative situation
While speech synthesis research is now focussing on the generation of various speaking styles or emotions,

very few studies have considered the possibility of including phonetic variations according to the communicative situation of the targeted speech (sports commentaries, TV news, etc.). This paper proposes a phonetic analysis of large French corpora to assess the influence exerted by three situational 'traits': read/spontaneous, media/non-media and expressive/non-expressive. It shows that some variations, like elision, tend to be more frequent in spontaneous and non-media speech, conversely to liaisons which appear more often in read and media speech. Interestingly, no phonetic variation draws a clearcut distinction between expressive and non-expressive speech. Finally, a prosodic analysis indicates that the phonetic variations are not directly correlated with the rhythmic features of their corresponding situational 'trait'.

2.4.11 p.433

Netta Weinstein, Konstantina Zougkou, Silke Paulmann,

Differences between the acoustic typology of autonomy-supportive and controlling sentences

The current study was first to describe distinct patterns of prosody that discriminate motivationally laden speech. To do this we applied self-determination theory, a widely used motivational framework. Participants in the US and UK were asked to read out loud either autonomy-supportive sentences (that support choice and volition) or controlling (pressuring and coercive) sentences. Data analyses were conducted using a conservative hierarchical linear modeling approach to account for nesting of sentences within individuals. Across both countries and controlling for gender, autonomy-supportive sentences were read using lower pitch, less intensity, and a slower speech rate than were controlling sentences. Multiple regression analyses showed links between these patterns of prosody for each participant and his or her current level of motivation, providing additional validity to results. Findings inform both the motivation and prosody literatures and offer a first description of how different kinds of motivational speech may sound.

2.4.12 p.438

Jasmin Pfeifer, Silke Hamann, Mats Exter,

Congenital Amusia in linguistic and non-linguistic pitch perception: What behavior and reaction times reveal

Congenital Amusia is a developmental disorder that has a negative influence on pitch perception. While it used to be described as a disorder of musical pitch perception, recent studies indicate that congenital amusics also show deficits in linguistic pitch perception. This study investigates the perception of linguistic and non-linguistic pitch by ten German amusics and their matched controls. To test the influence of amusia on linguistic pitch perception, the present study parametrically varied pitch differences in steps of one semitone in resynthesized statement-question pairs. In addition, we looked at the influence of stimulus duration, continuity of pitch and direction of pitch change (statement or question). Performance accuracy and reaction times were recorded. Behavioral results show that amusics performed worse than controls over all conditions. The reaction time analysis supports these findings, as amusics were significantly slower across all conditions. Both groups were faster in discriminating statements than questions. Performance accuracy supports these findings, as questions were also harder to discriminate. The present results warrant further investigation of the linguistic factors influencing amusics' perception of intonation.

2.4.13 p.443

Jue Yu, Dafydd Gibbon, Katarzyna Klessa,

Computational annotation-mining of syllable durations in speech varieties

There are many techniques for modelling properties of speech duration patterns, including models of rhythm as oscillation, partial models of rhythm types as departures from isochrony, models of tempo acceleration and deceleration, and models of duration hierarchies and their relation to hierarchies in word and phrase structure. Except for oscillator modelling, many approaches use data extraction from speech annotations, often with mainly manual methods. We employ computational data-mining for phonetic research, as opposed to phonological research on the one hand or speech technological research on the other, and explore the potential of the computational annotation data-mining paradigm for improving efficiency and scope of analysis. We show consistent variation in syllable duration patterns in selected speech varieties in English, Chinese and Polish, chosen for their known different prosodic typological properties. Results include a possible limen of 50ms for relevant timing patterns. For data-mining we use the Time Group Analysis (TGA) methodology, directly in the TGA online tool and integrated into

the Annotation Pro+TGA desktop software.

2.4.14 p.448

Izabel Seara, Juan Manuel Sosa, Vanessa Nunes,
Sentence type and prenuclear contours in Brazilian Portuguese: production and perception
In this paper we examine how the interrogative sentence mode is encoded in some dialects of Brazilian Portuguese (BP) and how questions differ from their declarative counterparts. Our aim is to identify which specific prosodic features, including prenuclear pitch range values, are systematically associated with the interrogative mode of enunciation. In the interdialectal comparison, the speakers from Blumenau (SC) and Aracaju (SE) distinguish themselves from the speakers of the other varieties in their prenuclear patterns, significantly higher for the yes/no interrogatives than the declarative counterparts. This is not the case with other dialects in our study. Perception tests corroborated the production results.

2.4.15 p.453

Ji Young Kim,
Use of suprasegmental information in the perception of Spanish lexical stress by Spanish heritage speakers of different generations
The present study examines the perception of Spanish lexical stress by Spanish heritage speakers of different generations and compares their performance to that of Spanish native controls and English second language (L2) learners of Spanish. Previous studies have shown that English L2 learners experience great difficulty in perceiving Spanish lexical stress. Such difficulty is argued to be derived from English listeners using different strategies from Spanish listeners in the perception of stress. Given that Spanish heritage speakers share the same dominant language with English L2 learners (English), but differ from them with regard to the first language (Spanish), the present study intends to seek whether heritage speakers show similar or different patterns when compared with L2 learners. The present study also intends to account for the heterogeneity among heritage speakers by comparing heritage speakers of different generations. Using a forced-choice identification task with stressed minimal pairs of paroxytone and oxytone verbs, results showed that while 1st generation US-born heritage speakers pattern like Spanish native controls by paying more attention to the acoustic cues of the stimuli, 1.5/2nd generation US-born heritage speakers pattern like English L2 learners by showing bias towards paroxytone verbs.

2.4.16 p.457

David Escudero, Lourdes Aguilar, César González Ferreras, Valentín Cardeñoso, Yurena Gutierrez,
Applying a fuzzy classifier to generate Sp_ToBI annotation: preliminar results
One of the goals of the Glissando research project1 is to enrich a radio news corpus [1] with Sp_ToBI labels. In this paper we present the application of the automatic predictions of a fuzzy classifier to speed the labeling process. The strategy is proposed after completing the following steps: a) manual annotation of a part of the Glissando corpus with Sp_ToBI labels and checking of the coherence of the labels; b) training of the automatic system; c) validation or correction of the automatic system's predictions by a human expert. The automatic judgments of the classifier are enriched with confidence measures that are useful to represent uncertain situations concerning the label to be assigned. The main aim of the paper is to show that there exists a correspondence between the uncertain situations that are identified during an inter-transcriber experiment and the uncertain situations that the fuzzy classifier detects. Labeling time reduction encourages the use of this strategy.

2.4.17 p.462

Marie-Catherine Michaux, Sandrine Brognaux, George Christodoulides,
The production and perception of L1 and L2 Dutch stress.
This study aims at exploring the production and perception of Dutch word stress by Francophone learners of (Belgian) Dutch. For this purpose a production experiment was first carried out. In line with other studies, it was hypothesized that participants would show a tendency to stress the final syllable. Even though this hypothesis was confirmed, there was also a substantial lack of agreement between the five labellers who perceptually annotated the data for stress position. To further investigate this matter, acoustic measures were extracted. The data suggest that both groups of speakers do not use acoustic

correlates to signal prominence in the same way, the Dutch group using intensity, vocalic nucleus duration and pitch movement more, while the French group prefers duration and pitch movement. This study also led us to develop tools to phonetise, syllabify and facilitate the acoustic analysis of Dutch speech.

2.4.18 p.467

Takayuki Kagomiya, Seiji Nakagawa,

Evaluation of bone-conducted ultrasonic hearing-aid regarding transmission of speaker gender and age information

Human listeners can perceive speech signals in a voicemodulated ultrasonic carrier from a bone-conduction stimulator, even if the listeners are patients with sensorineural hearing loss. Considering this fact, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA). The purpose of this study was to assess the usefulness of the BCUHA in transmission of speakers' physical attributes: gender and age. The evaluation used gender and age-identification experiments. The experiments were also conduced under air-conduction (AC) and cochlear implant simulator (CIsim) conditions. The results showed that: the BCUHA can well transmit speakers' gender information; the BCUHA can transmit speaker age information better than CIsim.

2.4.19 p.472

Mara Breen, Sarah Weidman, Katharine Guarino,

Rhythm and Expression in The Cat in the Hat

In recent years, there has been increasing interest in whether rhythmic interventions support young children's literacy development [1]. To begin to explore this connection, we assessed several aspects of rhythmicity and expressivity of productions of the notably rhythmic and rhyming children's book, The Cat in The Hat by Dr. Seuss. Participants subjectively rated either the rhythmicity or expressivity of speech taken from recordings of the book read aloud. These perceptual ratings were correlated with acoustic measures of rhythmicity and expressivity. Moreover, we observed a surprising lack of consistency between perceptual ratings of rhythmicity and expressivity. However, we observed a consistent relationship between the perceptual ratings of the first couplet of verses and the second. These findings can inform our investigation of the role of rhythm in literacy development.

2.4.20 p.477

Laurence White, Sven Mattys, Linda Stefansdottir, Victoria Jones,

Lengthened Consonants are Interpreted as Word-Initial

Prosody facilitates listeners' segmentation of the speech stream into a sequence of words and phrases. With regard to speech timing, vowel lengthening is interpreted as a cue to an upcoming boundary, in accordance with the iambic-trochaic law. However, the impact of consonant lengthening on segmentation, in the absence of other boundary cues, has not been tested. In a series of artificial language learning experiments, we examined how durational variation affects listeners' extraction of novel trisyllables defined by transition probabilities. In line with previous research, syllables containing lengthened vowels were interpreted by listeners as word-final. However, syllables with lengthened onset consonants were interpreted as word-initial. Thus, the structural interpretation of durational variation depends upon localization: longer vowels cue a following boundary; longer consonants cue a preceding boundary.

2.4.21 p.482

Alexandra Markó, Mária Gósy, Tilda Neuberger,

Prosody patterns of feedback expressions in Hungarian spontaneous speech

Speech communication incorporates non-verbal signals and semi-lexical vocal phenomena as well as words used as the listener's responses to the speaker's message. They are most common in conversation with various functions regardless of language. A specific subcategory is feedback expressions (FEs) that can be found in the listener's production as well as in the current speaker's speech production when reacting to the former speaker's message. This paper reports on the temporal and intonational characteristics of four types of FEs identified in 20 interviews and conversations from the BEA Hungarian database. Altogether 262 occurrences were categorized into four discourse functions signaling 'attention', 'comprehension', 'agreement' and 'other attitude'. Durations showed statistically significant differences across

discourse functions. They were significantly longer in females than in males in all functions. The pitch range data revealed a statistically significant difference depending on discourse function and gender only in the case of the 'attention' function. The dominant frequency contour was a rise in the functions of 'attention' and 'agreement' (90%). The same contour was observed only in 75.5% of the 'comprehension' function. An integrated approach is proposed to analyze these phenomena in spontaneous speech.

2.4.22 p.487

Eduardo Patricio Velázquez Patiño,
Intonation Patterns of Morelos Nahuatl
There are still relatively few studies on the phonetics and phonology of the indigenous languages of Mexico, and just a minority of them deals with less explored areas like prosody or, specifically, intonation. This study reports a preliminary analysis of Nahuatl intonation, taking into account its phonological characteristics: a) trochaic binary rhythm; b) generation of secondary stress inside rhythmic structures; c) generation of rhythmic groups according to clause structures; d) phonetic syllable lengthening at the end of sentences; e) laryngealization or voicelessness at the end of utterances, and f) vowel lengthening. Data collected by means of different methods, developed in order to obtain authentic and spontaneous utterances, show that different sentence types tend to have specific intonation patterns with many typologically common features and some original characteristics.

2.4.23 p.497

Amelia Kimball, Jennifer Cole,
Avoidance of Stress Clash in Perception of American English
We examine stress clash in perception, asking to what degree listeners perceive speech as metrically regular. We assess metrical regularity through a stress perception task carried out by untrained listeners annotating transcripts of spontaneous conversation and sentences designed to be metrically regular. Results show listeners report perceiving fewer stress clashes than predicted by random placement of stresses or by concatenating the citation form stress patterns of each individual word in a given sentence. These results suggest that listeners perceive spontaneous conversational English as metrically regular.

2.4.24 p.502

Lenka Weingartová, Eliška Churaňová, Pavel Šturm,
Transitions, pauses and overlaps: Temporal characteristics of turn-taking in Czech
This study aims to describe temporal characteristics of pausing and turn-taking phenomena in conversation. The material comes from the VASST corpus of contemporary Czech and uses four spontaneous dialogues in the form of an informal interview. We describe both general and idiosyncratic effects found in our data and compare them with results from other languages. In our material, transitions with a silent gap, overlaps and back-channels all display notably similar durational distributions with the median around 360 ms and a marked skewing. The four dialogues did not differ in the proportion of turns belonging to the interviewer (58%) vs. interviewee (42%), which is hypothesized to characterize the experimental task. Despite a number of general tendencies, individual differences in pausing and turn-taking behaviour of the speakers were found as well. For instance, the ratio of pauses and gap transitions proved to be highly dialogue-specific. We also gathered evidence for a substantial change in the speech behaviour of the interviewer resulting from a change of her communication partner.

2.4.25 p.507

Riikka Ullakonoja, Mikko Kuronen, Pertti Hurme, Hannele Dufva,
Segment Duration in Finnish as Imitated by Russians
The paper reports findings of a study in which Russian speakers without any prior knowledge of the language imitated Finnish utterances, and, in particular, how they succeeded in imitating segmental duration. The data was analyzed using acoustic measurements of segment duration as well as auditory analysis by four judges. The results show that while Russian speakers faced difficulties in imitating some aspects of the Finnish quantity, many imitated words were judged as comprehensible.

2.4.26 p.512

Candide Simard, Claudia Wegener, Albert Lee, Faith Chiu, Connor Youngberg,
Savosavo word stress: a quantitative analysis
This paper presents a quantative analysis of stress in Savosavo (unclassified), an endangered language spoken on Savo Island, (Solomon Islands). Acoustic analyses comprise the measurements of F0, duration, and intensity for each syllable in a dataset carefully selected from elicited speech from one speaker only, aiming to test the effect of increasing morphological complexity on stress realization in a system that displays some variation. Statistically significant variation is found in all correlates between stressed and unstressed syllables, thus fitting with widely attested manifestations of stress cross-linguistically. Findings were further tested with a re-synthesis tool, to confirm our initial hypotheses. Our results demonstrate that the current annotation scheme is a reliable representation of the data, and that the qTA component embedded in PENTAtrainer is effective in modelling F0 contours, even with less controlled data as input. We will argue for the usefulness of instrumental phonetic investigations in describing lesser-known languages, to enhance our understanding of the characterization of the prosodic systems of the world's languages.

2.4.27 p.515

Mortaza Taheri-Ardali, Hamed Rahmani, Yi Xu,
The Perception of Prosodic Focus in Persian
In a previous production experiment, post-focus compression (PFC) of F0 and intensity were found to be present in Persian. It was also shown that F0 and duration were the main correlates of prosodic focus in Persian. However, the perceptual relevance of PFC in Persian was not yet clear. The present paper reports the findings of an experiment on focus perception in Persian. Native speakers of Persian listened to sentences produced with focus in different positions as well as the neutral-focus sentence, and judged the presence and location of focus. Results show that final focus is identified much less well than other types of focus, and most of its confusion is with neutral focus. This shows that the presence of PFC is a main factor in recognizing prosodic focus in Persian.

2.4.28 p.520

Catherine Lai,
Final Rises in Task-oriented and Conversational Dialogue
This paper examines the distribution of utterance final pitch rises in dialogues with different task structures. More specifically, we examine map-task and topical conversation dialogues of Southern Standard British English speakers in the IViE corpus. Overall, we find that the map-task dialogues contain more rising features, where these mainly arise from instructions and affirmatives. While rise features were somewhat predictive of turn-changes, these effects were swamped by task and role effects. Final rises were not predictive of affirmative responses. These findings indicate that while rises can be interpreted as indicating some sort of contingency, it is with respect to the higher level discourse structure rather than the specific utterance bearing the rise. We explore the relationship between rises and the need for co-ordination in dialogue, and hypothesize that the more speakers have to co-ordinate in a dialogue, the more rising features we will see on non-question utterances. In general, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data, and when we link them to speaker states such as uncertainty or submissiveness.

2.4.29 p.525

Philippe Martin,
Spontaneous speech corpus data validates prosodic constraints
In the Autosegmental-Metrical model, the prosodic structure is defined as a hierarchy of Accent Phrases (AP). Groups of AP form intermediate prosodic phrases ip, which in turn are grouped into Intonation Phrases IP, and finally sequences of IP form the sentence intonation unit. In this hierarchy several constraints affect the prosodic structure, such as the AP 7 syllables rule, the stress clash conditions, eurhythmicity and syntactic clash. These constraints have been established essentially from read sentences data. They lead to an experimental justification in the observed synchronization of AP's syllabic chunking by Delta brain waves. This paper investigates the validity of the prosodic structure constraints on spontaneous speech data in French, as well as the adequacy of the Delta waves characteristics to

synchronize AP data.

2.4.30 p.530

Michael Phelan,
Hearing the Structure of Math: Use and Limits of Prosodic Disambiguation for Mathematical Stimuli
Listeners use the prosodic cues of an utterance to help determine its syntactic structure, but how does this process happen in the specialized domain of mathematics? Mathematical expressions can contain deeply embedded structures, and listeners encounter read mathematical expressions (RMEs) far less frequently than other potentially ambiguous utterances. How does experience with listening to math affect our ability to hear the structure of an RME via its prosody? Are there limits to the amount of structure we can pull out of the prosody of an utterance? A perception experiment was conducted with subjects aged 7-59 to help answer these questions. Participants heard recordings of RMEs and attempted to determine which of two or more mathematical structures the reader intended. When subjects chose between two options for phrases like nine times A minus two, they chose the mathematical expression that had bracketing matching the prosody of the utterance. However, for more complex phrases like the square root of sixteen over A plus twelve, results were at chance. Age played a surprising role: subjects' performance increased dramatically from age 7 to 16, but adults' performance varied widely. This is attributed to variation in exposure to read mathematics.

2.4.31 p.534

Sujan Kumar Roy, Md. Khademul Islam Molla, Keikichi Hirose,
Robust Pitch Estimation using Ensemble Empirical Mode Decomposition
This paper presents an efficient pitch estimation algorithm for noisy speech signal using ensemble empirical mode decomposition (EEMD) based time domain filtering. The dominant harmonic of noisy speech is enhanced to make pitch period more prominent. The normalized autocorrelation function (NACF) of the modified signal is then decomposed into time varying subband signals using EEMD. In contrast to the ordinary EMD, it does not introduce any mode mixing during decomposition. The subbands containing pitch component are selected and separated yielding partially reconstructed signal. The pitch period is determined from thus separated signals. The experimental results show that the proposed algorithm performs better compared to other recently reported algorithms in noisy environment.

2.4.32 p.539

Mónica Domínguez, Mireia Farrús, Alicia Burga, Leo Wanner,
The Information StructureProsody Language Interface Revisited
Several grammar theories relate information structure and prosody, highlighting a major correspondence between theme and rheme, and intonation patterns. Although these theories have been successfully exploited in some specific speech synthesis applications, they are mainly based on short default-order sentences, which limits their expressiveness for real discourse with longer sentences and complex structures. This paper revises these theories, identifying cases in which they are valid, and providing a new proposal for cases in which a more complex model is needed. Specifically, our experiments performed on real discourse from the Wall Street Journal corpus show that we need a model that: (1) foresees a hierarchical theme/rheme structure, and (2) introduces, apart from the traditional theme and rheme, a new element—the specifier.

2.4.33 p.544

Bahia Guellai, Alan Langus, Marina Nespor,
Prosody is perceived in the gestures of the speaker
It has been suggested that speech and hand gestures could form a single system of communication that facilitates the interaction between the speaker and the listener. What kind of information do gestures carry? In the present study, two experiments test the possibility that spontaneous gestures accompanying speech carry prosodic information. Experiment 1 shows that gestures provide prosodic information as adults are able to perceive the congruency between a low-pass filtered thus unintelligible - speech stream and the gestures of the speaker. These results show that prosody is not a modality specific phenomenon

and can be perceived in spontaneous gestures that accompany speech.

2.4.34 p.548

Joseph Tyler,
Rising pitch and quoted speech in everyday American English
Phonetic variation in rising pitch has been analyzed for how it correlates with contextual factors like speaker gender, utterance type (questions vs. statements) and turn position (turn-medial vs. turn-final). This paper analyzes variation in terminal rising pitch between quoted and non-quoted speech, using data from the Santa Barbara Corpus of Spoken American English. Results show rises in quoted speech start and end higher, rise more overall, but are no different in duration. These results are gender-dependent, however, for while women produce 65% of all rises in the corpus sample, they produce 100% (n=23) of the quoted speech rises.

2.4.35 p.553

Daniel Aalto, Stina Ojala,
Fine temporal structure of Finnish sign language
Signs can be divided to syllables and further into transitions and nuclei based on the signing flow of the handshapes. Here, a mixed effects linear regression model is used to describe the variation in the duration of the syllable nuclei in a data set of 341 signs (474 syllables) produced by five native FinSL signers during a map task. The phonetic fixed variables are the duration of the adjacent transients and syllable nuclei; phonological fixed variables are the syllabic length of the sign, the syllable position within the sign, and the sign type (functional or content bearing). Both preceding and following nucleus had a significant effect on the nucleus duration, while an asymmetric effect was found for the transitions: only the postnuclear transition had a significant effect. The syllable structure had no effect. However, the nuclei were shorter in function signs. These results suggest that signs are produced in two stages where the first stage, preparatory transition, is merged with the production of the previous syllable, and the second stage consists of executing the sign.

2.4.36 p.558

Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno,
Pause insertion prediction using evaluation model of perceptual pause insertion naturalness
This paper describes a pause insertion prediction technique for generating more natural synthesized speech for text-to-speech (TTS) synthesis systems. A novel point of the proposed technique is the use of an evaluation model of perceptual pause insertion naturalness in addition to a prediction model based on machine learning. The evaluation model represents the relationship between several features related to pause insertion and the perceptual pause insertion naturalness obtained in a subjective evaluation. First, using a prediction model based on machine learning, we obtain the N-best sequences that indicate whether or not a pause is present at each phrase boundary. We then estimate pause insertion naturalness scores for each N-best sequence using the evaluation model and select the sequence with the highest naturalness score. Objective and subjective evaluation results show that the proposed technique gives better results than a conventional technique.

2.4.37 p.492

Sophie Herment, Nicolas Ballier, Elisabeth Delais-Roussarie, Anne Tortel,
Modelling interlanguage intonation: the case of questions
In this paper, we study the intonational patterns observed in learners' productions in order to evaluate what motivates the deviations observed: systemic differences between the learners' L1 and the L2, differences in phonetic implementation, etc. The analysis consists of a cross-comparison of the intonation of yes-no questions in French, English and English as an L2. It is based on five information-seeking yes-no questions that were extracted from the AixOx corpus, which contains a set of 40 texts that were read by 10 native French speakers, 10 Native English speakers and 20 French learners of English. The analysis of the data showed that the differences between native and non-native speakers do not affect

the form of the nuclear contour. It mostly shows that French speakers of English have a tendency to assign a rising pitch movement at the end of any prosodic words, which leads to a clear difference in rhythm.

2.4.38 p.563

Fabian Santiago, Paolo Mairano, Elisabeth Delais-Roussarie,
Non-native perception of final boundary tones in French interrogatives
In this paper, we report a perception experiment in which native and non-native listeners judged resynthesized questions varying in respect to two aspects: their morphosyntactic structure (presence/absence of an interrogative marker) and the form of their final tonal contour (falling, rising and extra rising). The goal of the experiment was to exanimate how non-native listeners of French did perceive the extra-rising final contour that was observed in learners' productions. Do they consider it as unmarked form during the acquisition process? By and large, the results of the experiment show that native listeners preferred rising contours over falling ones in all question types, whereas non-native listeners rated the extra rising contours higher than French natives in stimuli having a morphosyntactic structure that differs from the one used in their L1. These results suggest that rising contours represent a default tonal form associated with the interrogative modality at the beginning of the L2 acquisition process.

2.4.39 p.568

Katalin Mády, Ádám Szalontai,
Where do questions begin? – phrase-initial boundary tones in Hungarian polar questions
Hungarian prosody is left-headed, as suggested by the placement of the accent on the initial syllable on the level of prosodic words and the placement of the strongest pitch accent on the first accented word of the prosodic phrase. Earlier studies have pointed out that the left edge of the intonational phrase can bear a phrase-initial boundary tone that distinguishes between string-identical wh-interrogatives and wh-exclamatives. In this paper, two other string-identical sentence types, polar questions and declaratives, are investigated with respect to their prosodic features. Polar questions were characterised by a higher f0 maximum and a lower sentence-initial f0 than declaratives. The only pitch accent within the sentence was low, whereas declaratives had falling pitch accents. Sentence-final f0 and the pitch level of the accented syllable did not show a consistent pattern across speakers. It is concluded that low sentence-initial f0 together with the high tone on the penultimate syllable is a relevant marker of polar questions in Hungarian.

2.4.40 p.573

Xiaoming Jiang, Marc Pell,
Encoding and decoding confidence information in speech
This study aims to investigate the perceptual-acoustic correlates of vocal confidence. Statements with different communicative functions (e.g., stating facts, making judgments) were spoken in confident, close-to-confident, unconfident and neutral voices. Statements with preceding linguistic cues (e.g. I'm positive, Most likely, Maybe, etc.) or no linguistic cues were presented to sixty listeners in a perceptual study. The listeners were asked to judge whether statements conveyed some level of confidence, and if so, they were asked to evaluate the level of confidence of the speaker. The results demonstrated that the intended levels of confidence varied in a graded manner in the perceptual rating score; the more confident the statement intended to be, the higher the rating. In general, the neutral voice was judged to be more confident than the close-to-confident voice, but less than the confident voice. The presence of a linguistic cue tended to increase ratings of confident voices but decrease ratings of voices in the less confident voice conditions. To evaluate how specific prosodic cues are used to encode and decode confidence information, acoustic analyses were performed on the stimuli without the linguistic cue based on the mean perceptual rating of speaker confidence for each item. Results showed that statements rated as confident versus unconfident differed in the mean and the variance of fundamental frequency (f0) as well as speech rate, with confident statements exhibiting lower mean f0, smaller f0 variance, and faster speaking rate than unconfident statements. The perceived level of confidence was differentiated in the mean fundamental frequency in a parametric way, the lower the level of confidence, the higher the mean f0. Confident voices were also distinct from the other three conditions in terms of mean and range of amplitude (i.e., loudness). These findings shed light on how linguistic and paralinguistic cues reveal confidence-related information to listeners during speech.

2.4.41 p.577

Jane Kühn,
Aspects of Prosodic Phrasing in Turkish
This pilot study investigates the prosodic marking of contrastive in-situ focus in monolingual Turkish. The results of the production study are based on a phonological and phonetic analysis of information structure modified target sentences. The prosodic analyses reveal (i) feature that derive properties of prosodic phrasing which are inherent to phrase languages. It is shown that Turkish is a radical splitting language since each prosodic word (ω) forms its own phonological phrase ϕindicated by a high phrase tone (H-) aligned to ω- final syllables. The languages preference for radical splitting of simple SOV sentences is maintained in information structure modified targets by one speakers group, but modified by another group in favor of wrapping adjacent given constituents into one ϕ. The analyses reveal (ii) that prosodic cues are not crucial to mark in-situ focus in Turkish, but they may be used to contextualize information structure. If focused constituents are marked at all by prosodic means they do not show an increased pitch like most Germanic languages, but focused constituents are aligned to prosodic boundaries. The data motivate the claim that prosodic alignment is an adequate way to describe the prosodic realization of focus in Turkish.

2.4.42 p.582

Lehlohonolo Mohasi, Thomas Niesler, Hansjörg Mixdorff,
Perceptual evaluation of the effect of mismatched Fujisaki model commands and surface tone in Sesotho
Sesotho is a tonal Southern Bantu language which has so far received extremely little attention by the speech research community. We consider tone modelling for Sesotho using the Fujisaki model-based analysis with a view to the development of a text-to-speech (TTS) system. Fujisaki analysis can be used to indicate the tone associated with a syllable, but it often differs from the surface tone that would be available for TTS synthesis. We investigate instances in which the surface tone differs from the tone indicated by Fujisaki analysis, and determine the effect of these discrepancies on speech quality. The amplitude of Fujisaki tone commands is manipulated to match the surface tones, and the resulting resynthesized speech subsequently analysed by perceptual tests. We find that the effect of inserting tone commands at high surface tone syllables is more severe than matching the Fujisaki tone commands with low surface tone syllables, in terms of naturalness. Furthermore, some discrepancies can be attributed to errors in the surface tonal transcription. However, on average, all manipulations lead only to a mild degradation in speech quality. We conclude that the Fujisaki model is a feasible way to model tone in Sesotho even in the presence of limited and under-developed linguistic resources.

2.4.43 p.587

Suki Yiu,
Musical Intervals of Tones in Cantonese English
It has been shown that the relative pitch levels of Cantonese tones closely correspond to musical intervals (MIs). Given that an emerging tone language, Cantonese English, has developed tone under the substrate influence of Cantonese, this paper examines the correspondence between the newly emerged tones and MIs, and how the musical analogy relates to those established for Cantonese. The fundamental frequencies of the tones produced by six speakers of Cantonese English were extracted with Praat, then time-normalized across rhymes. The mean values of the interval points of two tones were expressed in terms of ratio, then matched with the closest MI on the musical scale. This paper demonstrates that the pitch levels of tones in Cantonese English correspond to MIs, given the converging ranges of MIs for different speakers and similar MIs of different tone pairs for different speakers. It also shows that the MIs of tones in Cantonese English are related to the corresponding tone pairs for Cantonese. The viability of MI as a means to understand the tonal system of non-tonal languages whose speakers' native language is tonal extends the link between the use of pitch in speech tones and music.

2.4.44 p.592

Wentao Gu, Keikichi Hirose,
Rhythmic Patterns in Native and Non-Native Mandarin Speech
Rhythm plays an important role in the naturalness of speech. This study compared rhythmic patterns

of Mandarin speech between native speakers and two groups of L2 speakers whose first languages were Cantonese and English, respectively. The study started from isolated words, but focused on continuous speech, for which eleven durational metrics were used as objective rhythm indicators. The results on continuous speech showed that nonnative Mandarin gave a quite similar rhythmic mode as native one in terms of rate-normalized/independent metrics, but shifted towards the stress-timed class in terms of raw metrics, regardless of the rhythmic class of the L1. This seems to conflict with the L1 transfer effect and the results for isolated words, but it coincides with auditory impression and can be explained by speech rate difference and the lengthening effects associated with the change in prosodic structure.

2.5 Wednesday Evening Session

2-5-evening - **Reviewers' Reception in Trinity College Long Room** (by invitation only)

3 Day Three

3.1 Thursday Session One

May 22nd, 9am - 10:30am : 3-1-plenary (1+3 presentations)

3.1.1 KeyNote 3 (p.598)

Jürgen Trouvain 30-min
Laughing, breathing, clicking the prosody of nonverbal vocalisations

When analysing human spoken communication the focus on the linguistic side lies on speech with its verbal message, whereas the focus on the non-linguistic side usually is on the visually transported information such as gestures and facial expression. However, speech, especially in talk-in-interaction, also features numerous nonverbal vocalisations including various forms of laughter and inhalation noises as their most frequent forms. Although nonverbal vocalisations are usually short in duration they may provide rich information on linguistic, paralinguistic and extralinguistic levels including prosodic phrasing, cognitive load, affective state or speaker identity. The talk provides an overview on the phonetic and prosodic structure and the timing of laughter and audible breathing. Special attention is put on conversational speech where we can frequently find situations in which interlocutors temporally overlap. An emphasis is given to apical click sounds that often occur with inhalation before upcoming articulation but also during word-finding difficulty.

3.1.2 p.603

Willemijn Heeren, Sarah Bibyk, Christine Gunlogson, Michael Tanenhaus,
Tuning in to whispered boundary tones
Very little is known about how listeners incorporate "intonational" information in whispered speech during online language processing. We present data showing that listeners can incorporate information about boundary tones in whispered speech rapidly, but this process is complicated by additional structural biases as well as by the fact that speakers do not produce cues to boundary tones consistently in whisper. Listeners, however, are able to adapt to these differences in order to correctly identify different boundary tones in whisper.

3.1.3 p.608

Oliver Niebuhr,
"A little more ironic" Voice quality and segmental reduction differences between sarcastic and neutral utterances
The presented production experiment analyzes the phonetic differences between neutral (i.e. sincere) and sarcastically ironic utterances in German. Results show in line with previous studies that sarcastic irony is expressed by longer utterance durations, lower and flatter F0 contours, and a lower intensity level. Moreover, extending previous findings, sarcastic irony is also characterized by a more variable (in tendency breathier) voice quality and a higher degree of segmental reduction, probably reflecting the speakers' dissociation from the wording of their utterances.

3.1.4 p.613

Amélie Rochet Capellan, Gérard Bailly, Susanne Fuchs,
Is breathing sensitive to the communication partner?
This paper investigates breathing profiles in eleven female speakers (subjects) when talking successively with the same two females (partners). Breathing kinematics of the two interlocutors was recorded synchronously by means of two Inductance Plethysmographs. In order to understand the implication of breathing in dialogue, we analyzed changes in breathing pauses according to the main dialogue events (listening, backchannels, turns start and turns continuation). Breathing and syllable rates were also compared among partners and subjects. The duration of inhalations and related pauses was reduced before a turn continuation in comparison to a turn start. The delay between speech offset in a breathing cycle and the onset of the next inhalation increased when a speaker and a listener swap roles as compared to a speaker who continued the turn. This was observed for both partners and subjects. The partners differed in their breathing and articulation rates but the two rates were not clearly correlated. In agreement with previous works, the current study shows that breathing kinematics is strongly linked to dialogue events. However, it doesn't show any clear effect of partner on speaker's breathing. This last result is discussed relative to methodological aspects.

3.2 Thursday Session Two - Poster

11am - 1pm : 3-2-poster (48 presentations)

- theoretical and linguistic prosody -

3.2.1 p.619

Carlos Gussenhoven, Lu Wang,
Yuhuan Wu tone and the role of sonorant onsets
Co-occurrence restrictions on tones and consonants in Yuhuan Wu Chinese syllables provide a powerful illustration of the phonetic basis of phonological contrasts, with sonorant contexts allowing more tone contrasts than other contexts. Interestingly, the language also reveals that the phonetic implementation of tones depends on the phonological contrast it is involved in. Such phonetic enhancement may be the opposite of what could be expected on the basis of speech ergonomics. Moreover, the language has two tone deletion rules that exempt tones in the context with the largest number of contrasts, showing a phonological version of enhancement.

3.2.2 p.623

Preethi Jyothi, Jennifer Cole, Mark Hasegawa-Johnson, Vandana Puri,
An Investigation of Prosody in Hindi Narrative Speech
This paper investigates how prosodic elements such as prominences and prosodic boundaries in Hindi are perceived. We approach this using data from three sources: (i) native speakers of Hindi without any linguistic expertise (ii) a linguistically trained expert in Hindi prosody and finally, (iii) automatic classifiers trained on English for prominence and boundary detection. We use speech from a corpus of Hindi narrative speech for our experiments. Our results indicate that non-expert transcribers do not have a consistent notion of prosodic prominences. However, they show considerable agreement regarding the placement of prosodic boundaries. Also, relative to the non-expert transcribers, there is higher agreement between the expert transcriber and the automatically derived labels for prominence (and prosodic boundaries); this suggests the possibility of using classifiers for automatic prediction of these prosodic events in Hindi.

3.2.3 p.628

Zenghui Liu, Aoju Chen, Hans Van de Velde,
Prosodic focus marking in Bai
This study investigates prosodic marking of focus in Bai, a Sino-Tibetan language spoken in the Southwest of China, by adopting a semi-spontaneous experimental approach. Our data have shown Bai speakers increase the duration of the focused constituent and reduce the duration of the post-focus constituent to encode focus. However, duration is not used in Bai to distinguish focus types differing in size and contrastivity. Further, pitch plays no role in signaling focus and differentiating focus types. The results thus

suggest that Bai uses prosody to mark focus, but to a lesser extent, compared to Mandarin Chinese, with which Bai has been in close contact for decades, and Cantonese, to which Bai is similar in the tonal system.

3.2.4 p.633

Maria Paola Bissiri, Margaret Zellers, Hongwei Ding,
Perception of Glottalization in Varying Pitch Contexts in Mandarin Chinese
Although glottalization has often been associated with low pitch, evidence from a number of sources supports the assertion that this association is not obligatory, and is likely to be language-specific. Following a previous study testing perception of glottalization by German, English, and Swedish listeners, the current research investigates the influence of pitch context on the perception of glottalization by native speakers of a tone language, Mandarin Chinese. Listeners heard AXB sets in which they were asked to match glottalized stimuli with pitch contours. We find that Mandarin listeners tend not to be influenced by the pitch context when judging the pitch of glottalized stretches of speech. These data lend support to the idea that the perception of glottalization varies in relation talinao language-specific prosodic structure.

3.2.5 p.638

Robert Bo Xu, Peggy Pik Ki Mok,
Cross-linguistic perception of Mandarin intonation
This study investigated how phonological knowledge and psychoacoustic mechanism interact in intonation perception. In the experiment, Mandarin and Cantonese listeners identified Mandarin statement and question in both unfiltered and low-pass filtered contexts. The results show that the importance of different perceptual factors varies depending on the perception materials. Language background plays an important role even in processing low-level psychoacoustic materials.

3.2.6 p.643

Wilbert Heeringa, Jörg Peters, Heike Schoormann,
Segmental and prosodic cues to vowel identification: The case of /I i i:/ and /U u u:/ in Saterland Frisian
Saterland Frisian has a complete set of closed short tense vowels. Together with the long tense vowels and the short lax vowels they constitute series of phonemes that differ by length and/or tenseness. We examined the cues that distinguish the front unrounded and the back rounded series of short lax and short and long tense vowels in triplets by eliciting 'normal speech' and 'clear speech' in a reading task from two speakers. Short and long vowels were distinguished by vowel duration, and lax and short vowels by their location in the F1-F2 space. The durational difference between short tense and long tense vowels, however, was largely restricted to the 'clear speech' condition. In 'clear speech', f0 excursion and centralization in the F1-F2 space were used as additional means to make short tense vowels more distinct from long tense vowels. These results suggest that length and tenseness are used as distinctive features, while f0 excursion and centralization in the F1-F2 space were optionally used to enhance the contrast between short and long tense vowels.

3.2.7 p.648

Marilisa Vitale, Philippe Boula de Mareüil, Anna De Meo,
An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based on yes/no questions
The present study investigates the prosody of yes/no questions (in comparison with statements) in Chinese learners and native speakers of Italian. Acoustic analyses and a perceptual test were performed, in order to identify the main trends in non-native productions. Results show the relevance of prosody, which differentiates elementary, intermediate and advanced Chinese learners of Italian. Listening tests based on prosody transplantation also suggest that non-native segments with a native Italian prosody are rated as less accented than are native Italian segments with a non-native prosody. Similar trends were found, overall, in terms of question/assertion discrimination, confirming the relative importance of prosody. These findings could be helpful for teachers and learners of Italian as a foreign language.

3.2.8 p.653

Wei Lai, Ya Li, Hao Che, Shanfeng Liu, Jianhua Tao, Xiaoying Xu,
Final Lowering Effect in Questions and Statements of Chinese Mandarin Based on a Large-scale Natural Dialogue Corpus Analysis
To support text-to-speech with detailed prosody rules and generate natural prosody, the paper studied the pitch variation near the end of sentences based on a Chinese natural dialogue corpus. An additional lowering effect on the last prosodic word was found in both questions and statements, and proved to be independent of tone influence. Nevertheless, this effect, which is referred to as final lowering in other languages, was claimed to be absent in Chinese by some previous experimental studies. The cause of such a contradiction is very likely to be the difference between experimental speech vs. natural speech. Based on this observation, this paper proposed a combination of the two methods in intonation studies, in which experimental speech serves as an entry point to develop new topics, and natural speech serves as a necessary extension to revise and apply prosody rules.

3.2.9 p.658

Vered Silber-Varod, Tal Levy,
Intonation Unit Size in Spontaneous Hebrew: Gender and Channel Differences
In this corpus-driven research, the question of whether there is a tempo at the Intonation Unit (IU) level, and whether defined IUs differ not only with regard to their pitch contour and boundary tones but also with respect to their phonological size. For this reason, the inventory of syllable size (in terms of segments (phonemes)) and word size (in terms of syllables) was examined, and then each IU category (mainly Terminal vs. Continuous) was measured with respect to the number of syllables and words it contains. Moreover, terminal IU size was also measured with regard to the amount of embedded continuous IUs. Results showed that terminal IUs in spontaneous Israeli Hebrew (IH) do not necessarily consist of embedded continuous IUs. This can be explained due to their massive use as short feedback units in spontaneous speech. Statistical measurements for gender and channel (Face-to-Face vs. telephone conversations) variables were carried with no significance for gender, but with statistically significance for several channel aspects. Last, estimated durational measurements of the IU size are presented.

3.2.10 p.663

Allison Benner, John Esling,
Acoustic Cues to Tone and Register in Bai: Adult Baseline Data
This paper presents the results of a study of the acoustic cues associated with the tense/lax distinction in Bai, a Tibeto-Burman register tone language spoken in Yunnan, China. The purpose of the paper is to provide baseline adult data for comparison with infant speech in an acoustic study of infants' acquisition of Bai register tones in the second and third years of life. The results show that among adults, F0, F1, and spectral tilt combine to create the tense/lax contrast in Bai. While these three cues tend to be correlated, individual speakers differ in their use, particularly spectral tilt. The patterns in this study suggest that as Bai infants acquire tones in the second and third years of life, their utterances are likely to become structured around these three acoustic cues in previously unattested ways that exemplify the complex interaction between universal physiological and developmental tendencies and the ambient phonological tone system of Bai.

3.2.11 p.668

José Hualde, Tomas Riad,
Word accent and intonation in Baltic
We examine the realization of word accent contrasts in Standard Latvian and East Aukštaitian Lithuanian across intonational contexts. In our Latvian data the contrast is manifested as level vs. falling pitch in most contexts, in addition to a durational difference. In Aukštaitian Lithuanian, instead, differences in vowel quality and duration cue the lexical contrast in the nuclei that we examine. While Latvian retains a tonal contrast, in Aukštaitian Lithuanian it has been replaced with a combined segmental/quantitative contrast, where the so-called circumflex tone corresponds to relatively shorter duration and, in the case of diphthongs, centralized quality in the first half. We discuss the implications of these findings for further typological work.

3.2.12 p.673

Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, Jiahong Yuan,
Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information
A deep neural network (DNN) classifier based only on 40 mel-frequency cepstral coefficients (MFCCs) achieved 29.99% frame error rate (FER) and 16.86% segment error rate (SER) in recognizing five tonal categories in Mandarin Chinese broadcast news. With the addition of sub- band autocorrelation change detection (SACD) pitch-class features, the classifier scored 27.58% FER and 15.56% SER. These results are substantially better than the best previously reported results on broadcast-news tone classification, and are also better than a human listener achieved in categorizing test stimuli created by amplitude- and frequency-modulating complex tones to match the extracted F0 and amplitude parameters. The same DNN architecture scored substantially worse when trained and tested with SACD pitch-class parameters alone: 39.22% FER and 24.89% SER. RAPT F0 estimates are worse yet: 44.37% FER and 27.28% SER. The 40 MFCC parameters do not encode F0 in any obvious way and attempts to predict SACD or other pitch features from them work badly. These surprising results raise difficult questions for theories of Chinese tone.

3.2.13 p.678

Marzena Zygis, Daniel Pape, Luis Jesus, Marek Jaskula,
Intended intonation of statements and polar questions in Polish in whispered, semi-whispered and normal speech modes.
This paper provides acoustic correlates of intonation in whispered, semi-whispered and normal speech modes. In particular, it investigates correlates of utterance-final rising intonation in polar questions and falling intonation in statements. The paper does not only examine properties of vowels but also properties of the following voiceless consonant clusters. For the purpose of this study 2592 items produced by 16 native speakers of Polish were analysed. The results point to differences in spectral properties of both utterance-final vowels and consonants where falling intonation in statements contrasts with rising intonation in polar questions. Regarding the consonants, questions are produced with higher peaks, intensity, COG and STD values as well as smaller skewness and kurtosis values. Some spectral differences of consonants, including spectral slopes, are more distinguishable for questions versus statements in the whispered speech mode than in other speech modes. The more pronounced role of these cues in whispered speech suggests their compensatory function for the fundamental frequency, which is present in phonated speech. In summary, the study shows that speakers produce intended intonation patterns by varying the choice of cues as well as their magnitude in dependence on both (i) speech modes and (ii) intonation patterns.

3.2.14 p.683

Alina Lausecker, Annika Brehm, Ingo Feldhausen,
Intonational Aspects of Imperatives in Mexican Spanish
This paper sheds new light on the intonation of imperatives in Mexican Spanish. Results from a production experiment based on scripted speech show that imperative sentences have two different nuclear configurations depending on the position of the imperative verb (VI): (i) (L+)H* L% with VI in sentence-final position, and (ii) L* L% with VI in non-final position. The pitch accent on VI in non-final position is characterized by a late peak (L+>H*). However, if the sentence is uttered with some sort of emphasis, the nuclear configuration in the non-final context can also be rising. While these results partly confirm claims made concerning the nuclear configuration in De-la-Mota et al. (2010), they contradict the findings in Willis (2002), who attested strong pitch accent variation on VI.

3.2.15 p.688

Joanne Jingwen Li, Peggy P.K. Mok,
The acquisition of English lexical stress by Cantonese-English bilingual children at 2;06 and 3;0
This study investigates the acquisition of English lexical stress by Cantonese-English bilingual children at the age of 2;06 and 3;0 respectively, comparing them with the English monolingual peers. Research on early bilingual phonological acquisition often focuses on segmental level. Few studies are available when it concerns prosodic features, especially in children speaking non-Indo-European languages. This study

examines an important prosodic feature, lexical stress, in Cantonese-English bilingual children. The results showed that there is delayed acquisition of English lexical stress among the bilingual children, as reflected in less contrastive syllable duration and peak F0, possibly due to a lack of lexical stress in Cantonese, a typical syllable-timed language. This study helps to understand the bilingual interaction of two distinctive prosodic systems, and broaden our knowledge about early bilingual prosodic development.

3.2.16 p.693

Adrian Leemann, Volker Dellwo, Marie José Kolly, Stephan Schmid,
Disentangling sources of rhythmic variability between dialects
Speech rhythm is highly variable. Previous studies reported variability between languages, dialects, speakers, and labelers. Research further revealed an effect of sentence in the rhythmic characteristics of speakers of the same language. In the present study we tested whether the effect of sentence material is constant across varieties of the same language. We addressed this question by an example of analyzing rhythmic variability between eight dialects of Swiss German in three different sentences. Results showed a significant interaction for dialect*sentence for most of the tested rhythm metrics. We take this as evidence that differences between dialects are contingent upon the sentences used in the experiment. We further investigated which sources in the sentence material caused between-dialect differences in rhythm scores to vary. W e found exemplary evidence that dialect-specific phonological and morphological phenomena contained in the individual sentences are the prime suspects. Implications for future speech rhythm research are discussed.

3.2.17 p.698

Maria Del Mar Vanrell, Olga Fernández Soriano,
Dialectal variation at the Prosody-Syntax interface: Evidence from Catalan and Spanish interrogatives
In this study we investigate how prosody interacts with word order in the expression of interrogativity in different varieties of two Ibero-Romance languages, Catalan and Spanish. We analyze a corpus obtained by means of the Discourse Completion Task Methodology. The collected data were prosodically and syntactically annotated and show that the absence of syntactic marking (wh-word, subject-verb inversion or subject dislocation) for questions tends to correspond to a more salient intonational marking. Thus, wh-questions favor general falling intonational patterns. By contrast, yes-no questions can be classified depending on the nuclear tone (with preference for low tones in Catalan and high tones in Spanish) and final tone (low for language varieties with subject inversion or dislocation, but optionally high for those that do not present syntactic marking in a mandatory way.

3.2.18 p.703

Mathieu Avanzi, George Christodoulides, Elisabeth Delais-Roussarie,
Prosodic Phrasing of SVO Sentences in French
In the literature on prosody/syntax interface, syntactic information is usually considered as playing an important role in deriving the prosodic phrasing of an utterance. NP subjects, for instance, have often been claimed to phrase independently from the VP. It has nevertheless been shown that metrical factors could have an impact on phrasing, and that NPs could be phrased in the same prosodic phrase as the VP, or that the verb could be phrased with the subject. Several methods were used to measure metrical weight: number of syllables, of prosodic words, syntactic branchingness, etc. In order to determine which factors are more important, and how the all interact, we evaluate the weight that different metrical predictors have on prosodic phrasing. This is done by analyzing the phrasing of SVO structures in 200 sentences extracted from various French corpora. From the observation of the data that were semi-automatically annotated, it appears that subjects can be phrased independently or in the same PP as the VP, and that objects are rarely isolated from the verb. The analysis reveals interesting results regarding the effect of articulation rate and number of syllables, whereas syntactic-branchingness didn't show any effect.

3.2.19 p.708

Michelina Savino, Andrea Bosco, Martine Grice,
Intonational cues to item position in lists: evidence from a serial recall task
Intonation can convey information about how lists are structured into groups, as well as about specific item positions within a group. In Bari Italian, this function is expressed by three different tunes a) a rising contour, signalling that the list has not yet been completed; b) a high-rising contour, marking the penultimate item, i.e. signalling that the end of the list is approaching; c) a falling contour, marking the last item, i.e. cueing the end of the sequence. In this paper we explore the effects of such intonational information on working memory. In particular, we demonstrate that when listeners are requested to recall spoken nine-digit sequences by strictly following their serial order, their performance is significantly better when lists are characterised by tunes of the type described above, compared to sequences whose items are marked by a neutral, peak accent and/or are grouped by inserting a silent pause. We also observed that recall of items marked by specific contours at positions 3, 6 and 9 is particularly enhanced at these positions, whereas in sequences also containing intonational cues to items in penultimate position (2, 5 and 8) recall of those items is not equally improved. Therefore, it appears that in serial recall of spoken sequences, even when a large number of specific intonational cues to serial positions are available, listeners can make use of only a selection of them.

3.2.20 p.713

Anqi Yang, Aoju Chen,
Prosodic focus-marking in Chinese four- and eight-year-olds
This study investigates how Mandarin Chinese speaking children use prosody to distinguish focus from non-focus, and focus types differing in size of constituent and contrastivity. SVO sentences were elicited from four- and eight-year-olds in a game setting. Sentence-medial verbs were acoustically analysed for both duration and pitch range in different focus conditions. The children started to use duration to differentiate focus from non-focus at the age of four. But their use of pitch range varied with age and depended on non-focus conditions (pre- vs. post-focus) and the lexical tones of the verbs. Further, the children in both age groups used pitch range but not duration to differentiate narrow focus from broad focus, and they did not differentiate contrastive narrow focus from non-contrastive narrow focus using duration or pitch range. The results indicated that Chinese children acquire the prosodic means (duration and pitch range) of marking focus in stages, and their acquisition of these two means appear to be early, compared to children speaking an intonation language, for example, Dutch.

3.2.21 p.718

Simone Graetzer, Janet Fletcher, John Hajek,
Prosodic effects on vowel spectra in three Australian languages
In this paper, the spectral properties of vowels in three Australian languages are examined with the aim of determining whether prosodic prominence and domain-edge effects on formant frequencies, formant variability and vowel space dispersion can be identified. It is shown that these vowel systems are sufficiently dispersed, with an anchoring of the system by the open central vowel. It is also shown that for Burarra but not for Gupapuyngu or Warlpiri there is some evidence of prosodically-driven hyper-articulation. Finally, the data indicate pre-boundary lengthening in all three languages, which in some cases appears to be associated with changes in vowel quality.

3.2.22 p.723

Xi Chen, Peggy Pik Ki Mok,
Rhythmic Correspondence between Music and Speech in English Vocal Music
This study aims to investigate the rhythmic structures of music and speech, and to find out the possible corresponding rhythmic patterns between the two domains in English vocal music. With fifteen English songs as samples, lexical stress of multi-syllabic words is compared with three musical dimensionsmetrical stress, duration, and pitch respectively. It is found that in the chosen English songs, there is a good mapping between the metrical stress of music and the lexical stress of lyrics. In addition, the duration and the pitch patterns not only generally match lexical stress patterns most of the time, but also serve to manifest the prominence of the primary lexical stress on one hand, and to reflect the weakness of the unstressed syllables on the other. Except a general good match in rhythm, this study also shows

matching differences within the three comparisons. Matching degrees vary according to different meter patterns. Moreover, pitch takes priority over duration in their respective matching with lexical stress of the lyrics. Finally, the primarily stressed syllables match duration and pitch patterns much better than the unstressed ones do. Index Terms: Musical rhythm, Speech rhythm, English songs

3.2.23 p.728

Christoph Gabriel, Elena Kireva,
Speech rhythm and vowel raising in Bulgarian Judeo-Spanish
The study investigates selected prosodic characteristics of (Sofian) Bulgarian Judeo-Spanish, a diaspora variety of Spanish spoken by descendants of the Jews expelled from Spain, all of them bilingual speakers with Bulgarian as their dominant language. While exhibiting some few relics from Old Spanish on the segmental level, Judeo-Spanish shows a puzzling similarity with Bulgarian with respect to speech rhythm and vowel raising. It is shown that the two languages spoken by the bilinguals, Bulgarian and Judeo-Spanish, pattern alike in displaying almost the same rhythmic values (except for %V) and that raising of unstressed /a/ and /o/ as is typical of the variety of Bulgarian spoken in Sofia also regularly occurs in the Judeo-Spanish data. Our findings show that Judeo-Spanish is crucially influenced by Bulgarian, thus suggesting that it has largely converged toward the surrounding language on the phonological level.

3.2.24 p.733

Sandra Schwab, Carla V. Jara Murillo,
The role of stress perception in the assignment of written accent in Spanish
The aim of this investigation is to examine whether the adults' difficulty in placing the written accent in Spanish words is related to their ability in perceiving stress. The following variables were also taken into account in this study: the participant's education level (academic and non-academic), the stimulus lexical status (words and non-words), accentual pattern (proparoxytone, paroxytone and oxytone words) and length (2, 3 and 4 syllables). Participants performed a stress identification task and a word spelling task. Besides the effects of lexical status, education level and accentual pattern, results show an effect of the stress perception in the assignment of the written accent: stimuli with a correctly identified stress were more likely to be correctly written (i.e. with or without written accent) than the incorrectly perceived stimuli. This finding reinforces the idea that there is a relationship between prosodic and written skills.

3.2.25 p.738

Uwe Reichel, Alexandra Markó, Katalin Mády,
Parameterization and automatic labeling of Hungarian intonation
In Hungarian intonation research a common framework developed by Varga (2002) is to categorize the intonation within the domain of accent groups by character contours. We propose a linear parameterization of a subset of these contours derived from polynomial stylization. These parameters were used to train classification trees and support vector machines for contour prediction. Parameter extraction and training was carried out on the original F0 contours of spontaneous speech data as well as on three differently normalized variants suppressing fundamental frequency level and range effects. The highest accuracies were obtained for classification trees and F0 residuals after midline subtraction, but the overall performances were rather poor. Nevertheless, a significant improvement of the results was achieved by a Hidden Markov model to predict the correct label sequence from the partly erroneous classification output.

3.2.26 p.743

Maciej Karpinski, Katarzyna Klessa, Agnieszka Czoska,
Local and global convergence in the temporal domain in Polish task-oriented dialogue
Conversational parties tend to mutually adapt their communicative behaviour in a number of dimensions, from the level of physical aspects of speech signal and gesture, utterance properties, up to the level of mental representations. In the present study, an attempt is made to track the process of convergence in the temporal domain both as a global tendency and a local phenomenon. The material under study consists of two sets of task-oriented dialogues recorded with or without eye contact (telephone conversations) between the speakers. All the recordings were segmented into syllables and analysed in terms of speech rate and nPVI for each speaker as well as for the correlations between the speakers in each pair. Global convergence tendencies were proven to be weak but some influence of dialogue settings and gender was

found. The results seem to support the hypotheses that the alignment-related processes remain under the influence of many factors related to the dialogue flow and cannot be modelled as simply incremental.

3.2.27 p.748

Beatriz Raposo de Medeiros, Fred Cummins,
Speech and song synchronization: A comparative study
Does synchronization among speakers or singers require the presence of a beat? Is an implied underlying pulse or meter relevant? We set out to explore synchronization among speakers and singers as they speak or sing a variety of texts. We compare metrically strong nursery rhymes with non-metered prose. We compare singing in genres with two very different types of rhythm (samba and rock), and we compare sung and spoken versions of texts. In each case, we ask whether the rhythmic qualities of the texts facilitate synchronization. The metrical structure of the nursery rhyme does not facilitate synchronization compared to prose, while the simple beat of rock music does help. Further comparisons are provided in the text.

3.2.28 p.752

Katalin Mády, Uwe D. Reichel, Štefan Beňuš,
Accentual phrases in Slovak and Hungarian
Languages with primarily delimitative function of word stress commonly make use of accentual phrases (APs) in their intonational phonology (e.g. Tamil or French). Slovak and Hungarian are genetically unrelated but geographically close languages with word-initial lexical stress. In this paper we compared the stylised f0 of single accent groups (AGs) with the f0 level pattern of the entire intonational phrase (IP) to test if AGs are relevant for the intonational phonology of Slovak and Hungarian. Steep f0 slopes with a recurring pattern (rising or falling) and large deviations from IP level patterns were interpreted as evidence for the autonomy of the AG in the given language. The results suggest that Hungarian is indeed a language in which accent groups form a unit on their own, however, such evidence was not found for Slovak.

3.2.29 p.757

Nicole Dehé,
Final devoicing of /l/ in Reykjavík Icelandic
Icelandic has a phonological process which devoices sonorants after voiced segments in domain-final position, but to date the category of the relevant domain and potential further factors affecting it have not been identified. The present paper studies final devoicing of /l/, by which /l/ is realized as the voiceless lateral fricative [ɬ] in domain-final position. It reports on the results of an experimental reading study designed to test the exact environments of this process and the implications for a prosodic hierarchy for Icelandic. The results suggest that devoicing of /l/ is bound by the prosodic utterance. All instances of /l/ were devoiced in utterance final position. Within the utterance, final devoicing is optional, but the frequency of its application reflects the syntactic and prosodic hierarchy such that it is most frequent at a clause/an IP-boundary, significantly less frequent at a syntactic XP-edge and it almost never occurs within a syntactic XP.

3.2.30 p.762

Scott Lee,
The Realiziation of French Rising Intonation by Native Speakers of American English
This study examines the acquisition of French intonational rises by adult native speakers of American English. Production data were gathered using a discourse completion task and a storytelling task from eight American college students beginning a semester-long study abroad program in Southern France. Results suggest that speakers struggled with two particular aspects of French intonation: the grouping of words into Accentual Phrases, and the phonetic realization of phrase-final rises. In particular, the probability distribution for the alignment of the late L elbow was bimodal for L2 speakers but unimodal for L1 speakers, suggesting the use in the learner speech of two distinct tonal patterns instead of the single French LH*. Mean values for overall pitch range and the scaling of continuative rises were significantly

lower and less variable than French L1 values as well.

3.2.31 p.767

Niamh Kelly, Rajka Smiljanic,
Monosyllabic Lexical Pitch Contrasts in Norwegian
This paper examines the lexical tonal accent contrast in monosyllabic words in the Trondersk dialect of Norwegian. The results of a production experiment in which speakers produced the unmarked accent and the circumflex accent showed that the tonal distinction is characterized by a difference in f0 maximum, f0 height at onset, f0 minimum and its timing, and height of the final Accent Phrase H tone. The presence of the tonal accent contrast on monosyllabic words is unusual among dialects of Norwegian and Swedish.

3.2.32 p.772

Yu Lun Hsieh, Ching-Ting Chuang, Feng Fan Hsieh, Yueh Chin Chang, Wen Lian Hsu,
Taiwanese Tone Recognition Using Fractionalized Curve-fitting of Prosodic Features
In this paper, we examined different methods of modeling prosodic features of tones, and their effects on a speaker-independent Taiwanese tone recognition system. Tones can be modeled either by plain or curve-fitted features. Plain features represent the original curve faithfully using pitch values, while curve-fitted features can be thought of as an approximation to the values using mathematical functions, such as a Legendre polynomial. In addition, durational information of tones was also proven effective in previous researches. Thus, we proposed a new approach of modeling Taiwanese tones using curve-fitted features extracted from fractions of the pitch curve, along with duration as an additional prosodic feature. Our experimental results showed that using these features in an SVM classifier could substantially improve the accuracy of tone recognition in Taiwanese. Besides, we provided an empirical perspective for theoretic studies on tonal neutralization.

3.2.33 p.776

Bistra Andreeva, Grazyna Demenko, Magdalena Wolska, Bernd Möbius, Frank Zimmerer, Jeanin Jügler, Magdalena Oleskowicz- Popiel, Jürgen Trouvain,
Comparison of Pitch Range and Pitch Variation in Slavic and Germanic Languages
This study presents the results of a large-scale comparison of various measures of pitch range and pitch variation in two Slavic (Bulgarian and Polish) and two Germanic(German and British English) languages. The productions of twenty two speakers per language (eleven male and eleven female) in two different tasks (read passages and number sets) are compared. Significant differences between the language groups have been found: German and English speakers use lower pitch maxima, narrower pitch span and generally less variable pitch than Bulgarian and Polish speakers. These findings support the hypothesis that particular linguistic communities tend to be characterized by particular pitch profiles.

3.2.34 p.781

Philippe Martin,
Silent reading and prosodic structure constraints
Silent reading of written texts involves necessarily a process of subvocalization, i.e. the presence of a voice reading the text in the head of the reader speaking to her/himself. This process includes not only the sequences of syllables corresponding to the written material, but also sentence intonation. Since subvocalization cannot be eliminated other than by changing the status of each word into a pictographic function (as it may be the case for a STOP road panel sign), it is argued here that sentence intonation is essential to language comprehension, and more specifically to the conversion of sequences of syllables into higher order linguistic units (corresponding to accent phrases AP in the Autosegmental-Metrical model). Consequently, reading and in particular silent reading is constrained by the same rules than the prosodic structure in general, and specifically to the minimal duration of accent phrases. This minimal value, occurring when AP's contain only one syllable, is about 250 ms, a value which corresponds to the minimal period value of Delta brain waves. Therefore this AP minimal duration limits also the maximal number of AP that could be processed in silent reading, i.e. about 240 per minute, which corresponds to the maximal number of words per minutes experts in fast reading can process while keeping a reasonable

level of comprehension, i.e. about 800 wpm.

3.2.35 p.785

Emma Valtersson, Francisco Torreira,
Rising intonation in spontaneous French: how well can continuation statements and polar questions be distinguished?
This study investigates whether a clear distinction can be made between the prosody of continuation statements and polar questions in conversational French, which are both typically produced with final rising intonation. We show that the two utterance types can be distinguished over chance level by several pitch, duration, and intensity cues. However, given the substantial amount of phonetic overlap and the nature of the observed differences between the two utterance types (i.e. overall F0 scaling, final intensity drop and degree of final lengthening), we propose that variability in the phonetic detail of intonation rises in French is due to the effects of interactional factors (e.g. turn-taking context, type of speech act) rather than to the existence of two distinct rising intonation contour types in this language.

3.2.36 p.790

Ludger Paschen,
Intonation and focus marking in Ulyap Kabardian
This paper presents a pilot study that aims at establishing a model for the intonation of Ulyap Kabardian in the ToBI framework. On the basis of data gathered during a field trip in 2012, it is suggested that four/three pitch accents and three boundary tones are needed to describe intonation in four communicative contexts. Additionally, it is shown that for focus marking in Ulyap Kabardian questions, a stress shifting rule dislocates word stress to a prosodically determined position. This shift rule is extraordinary in that it is insensitive to stress clashes. From a cross-linguistic perspective, the intonation system of Ulyap Kabardian bears a higher resemblance to the system of one of the Kabardian dialects spoken in Turkey than to Russian, the principal contact language.

3.2.37 p.795

Oliver Jokisch, Tristan Langenberg, Gabor Pinter,
Intonation-Based Classification of Language Proficiency Using FDA
State-of-the-art pronunciation tutoring (CAPT) systems are based on ASR technology. Consequently, they can provide a distinguished learning feedback which is focused on phonetic features and the positions of articulation errors. In contrast with the relative success with segmental errors, the acquisition and assessment of second language (L2) prosody is still a challenging problem. Although prosodic parameters like f0 contour or duration measures are usually displayed, the consequential evaluation components are generally missing. Considering the strong variation in speech data, functional data analysis (FDA) is a useful concept which statistically analyses interrelations between principal components (e.g., given accentuation) and their contribution to superimposed forms (e.g., resulting f0 contour). This article describes baseline processing and preliminary results of a pilot study on the intonation-based proficiency classification of German by using FDA methods. The experimental part contains the FDA-based classification results compared to a perceptual classification by German natives.

3.2.38 p.800

Kieu Phuong Ha, Martine Grice, Marc Brunelle,
Tonal allophony in Vietnamese: Evidence from task-oriented dialogues
In this paper we investigate the behaviour of the lexical rising tone (SAC) in disyllabic sequences in the Northern variety of Vietnamese. Results from task-oriented dialogues show that this rising tone (SAC), when occurring before the lexical high-level tone (NGANG), can be realised as low level or falling, resembling a different tone in the language (HUYEN). This is the case word-internally and within noun phrases. Two further observations give us an indication that a sandhi process could be developing: (a) this variation is not found in sequences across a larger juncture, and (b) the SAC tone does not undergo this change before other tones.

3.2.39 p.804

Nina Grønnum,
Laryngealization or Pitch Accent - the Case of Danish Stød
According to recent proposals Danish stød is the phonetic manifestation of a HL tonal pattern compressed within one syllable, making the stød/non-stød distinction a special case of the more general tonal word accent distinction in Swedish and Norwegian. This review of the relevant aspects of Danish stød and intonation demonstrates that (1) such a tonal representation of stød is contradicted by the phonetic reality. (2) Stød is distributed in words according to roughly the same principles across regional varieties of Danish, but tonal patterns are highly variable. (3) Word accents in Swedish and Norwegian are associated exclusively with stressed syllables, whereas stød occurs also in less than fully stressed syllables, devoid of autonomous pitch movements. (4) A word in Swedish and Norwegian can have one pitch accent only, but Danish words may have more than one stød.

3.2.40 p.809

Ann Bailey,
Intonational Phonology of Cuban Spanish: A Preliminary AM Model
The present study proposes a preliminary model of intonational phonology for Cuban Spanish in the framework of Autosegmental-Metrical phonology. Data from controlled and semi-spontaneous speech were used to establish the boundary tones and pitch accents which are contrastive in this variety of Spanish. It was found that Cuban Spanish shares various tonal categories with both the Pan Spanish ToBI (Tones and Break Indices) and other Caribbean Island Spanish dialects (Puerto Rican and Dominican), but differ from these dialects in how those pitch accents and boundary tones are used to convey meaning. Cuban Spanish shares its primary prenuclear pitch accents and nuclear contours for imperative statement and narrow focus with the Pan Sp_ToBI, but shares the nuclear contours for broad focus, vocative, and wh-questions with Puerto Rican Spanish. Similar to the other Caribbean Island Spanish varieties, the Cuban Spanish boundary tone inventory consists of a subset of the attested boundary tones found in the Pan Sp_ToBI, and all three Caribbean varieties share low boundary tones in non-wh questions, a marker of Caribbean Spanish speech.

3.2.41 p.814

Roberto Paternostro, Jean Philippe Goldman,
Modeling of a rise-fall intonation pattern in the language of young Paris speakers
Intonation seems to be one of the major cues for identifying youth language in the Paris region. As part of a large-scale corpus-based analysis, this paper attempts to model a high-low final prosodic pattern, considered to be representative of a Paris working-class suburbs accent. Comparison with the emphatic high-low prosodic pattern, well-known in general French, will provide the opportunity for sociolinguistic insights. The ethnic hypothesis is dismissed in favor of a context-bound and interaction-sensitive interpretation.

3.2.42 p.819

David Le Gac,
Topic and Focus Intonation in Argentinean Porteño
This paper investigates the intonation of topics and focus in Argentinean Porteño. We have found that whereas tonal alignment is phonetically conditioned, pitch height and duration constitute the main cues to express various types of focus in declarative and interrogative sentences; at least four intonational categories seem to be used by our speakers and the relevance of a register feature is discussed. As for topics, they are marked by special tunes that depend on the type of sentence; in particular, the topic tune in questions is the opposite of those found in declaratives.

3.2.43 p.824

Liang Zhang, Yuan Jia, Aijun Li,
Analysis of Prosodic and Rhetorical Structural Influence on Pause Duration in Chinese Reading Texts
This paper investigates factors that influence pause duration in Chinese reading texts through examining

the stress degree in pre-pausal and post-pausal positions and the rhetorical structure in discourse as a whole. The RSTTool is used in diagramming the rhetorical structures of the texts. The recordings, extracted from the ASCCD corpus, are further analyzed acoustically and statistically by applying Praat and R. The statistical analysis results show that the stress degree in both pre- and post-pausal positions has a significant impact on pause duration. Moreover, the nuclearity in both positions have also been shown to have a remarkable influence. Specifically, the nucleus in pre-pausal and satellite in post-pausal positions can significantly lengthen the pause duration.

3.2.44 p.829

Irina Nesterenko,
Statistical and temporal properties of prosodic phrasing in French conversational speech
Our study investigates prosodic phrasing in a corpus of French conversational speech. We looked at statistical and temporal properties of prosodic constituents, which were previously identified within laboratory phonology paradigm. Prosodic annotation of our corpus implements two-level hierarchical model distinguishing major prosodic units (Intonational Phrases, IPs) and minor prosodic units (Accentual Phrases, AP). Both temporal data and distribution of the number of APs in an IP evidence the global tendency to produce shorter units in conversation. Moreover, Intonational phrases containing no more than two Accentual phrases cover 80% of the data. We discuss the implication of these results for both phonological studies of the constraints on prosodic phrasing and oral document tagging.

3.2.45 p.833

Oyedeji Musiliyu, Miguel Oliveira,
Intonational Patterns of Telephone Numbers In Brazilian Portuguese
The main purpose of this study was to identify intonational patterns of a quite common type of numeric grouping in Brazilian Portuguese: the one associated with telephone numbers. To this aim, 30 samples of spoken telephone numbers, read aloud by 85 native speakers of Brazilian Portuguese were analysed. The description of their intonation contour was observed by using Momel/Intsint [1] and ProsodyPro [2] scripts for Praat (version 5.3.53) [3], through a semi-automatic analysis of pitch variations in numeric groupings that form the telephone numbers. The results show a pattern of intonation and numeric grouping strategy that are sufficient enough to prosodically characterize different types of spoken telephone numbers in Brazilian Portuguese.

3.2.46 p.838

Simone Falk, Elena Maslow,
Song and speech prosody influences VOT in stuttering and non-stuttering adolescents
Since a long time, it is known that singing helps persons who stutter to produce their utterances more fluently. The prosodic characteristics of spoken and sung utterances differ considerably in their rhythmic and tonal structure. Therefore, it has been proposed that song prosody helps stutterers to improve their rhythmic planning of verbal material [1]. In order to investigate this idea, we examined temporal aspects, namely Voice Onset Time (henceforth, VOT) of voiceless plosives, in sung and spoken utterances of young German stutterers and non-stuttering controls. VOT tends to be reduced in song compared to speech. We expected a more important reduction in the stuttering group as voice onset timing should be facilitated in song compared to speech. Eight stuttering adolescents and eight normal fluent peers read and sang an altered version of Happy Birthday with test words containing the three voiceless stops /p/, /t/, /k/. Results showed that stuttering as well as non-stuttering adolescents reduced VOT during singing compared to speech. In contrast, only adolescents who stutter were less variable in their VOT production in song compared to speech. Additional analyses indicated further group differences in vowel duration following the stop consonant. These findings suggest that young stutterers benefit from sung prosody in their timing abilities.

3.2.47 p.843

Stefanie Jannedy, Melanie Weirich,
Some aspects on individual speaking style features in Hood German
Multiethnic urban German (Hood German) as spoken by adolescents in Berlin differs in several significant ways from more standard varieties of Berlin German. It is characterized by a variety of morpho-syntactic

alternations and phonetic variants uncommon to the regional standard spoken in Berlin. Previous quantitative corpus analyses have shown that overall speakers of the multiethnic youth style German have a strong tendency to centralize /ɔɪ/ compared to speakers rendering the local regional standard. This paper now summarizes this centralization tendency and investigates auditory salient realizations of variation by individuals which show tendencies towards a hiatus in the diphthong /ɔɪ/, breaking the nucleus and the off glide. Moreover, there are other prosodic and segmental co-occurring features in the speech of some adolescents which are displayed since it is suspected that some of these may be(come) markers of Hood German.

3.2.48 p.848

Katarina Bartkova, Mathilde Dargnat,
Automatic extraction of prosodic patterns Cross linguistic study on laboratory data
The goal of our study is to use an automatic approach to extract the general prosodic tendency of the speech signal conveyed by the F0 pattern and the syllable duration. The speech signal is prosodically annotated by an automatic prosodic transcriber and then prosodic patterns are extracted from this annotation. The pertinence of the pattern extraction is tested here on laboratory data containing isolated sentences in French and English uttered by native and non-native speakers. An analysis of the extracted parameters allows observing how the prosody of the sentences is defined by their shared syntactic structures and to what extend are the prosodic features used by the two languages different or similar. It appears from the analyzed data that such an automatic prosodic parameter processing can yield relevant information for a cross-linguistic study of the prosody.

3.3 Thursday Session Three - Panel: Terminology in Prosody Research

2pm - 2:30pm : *Tribute to Sugito Miyoko sensei*

2:30pm - 3:30pm : Panel discussion: **Terminology in Prosody Research**
Organisers: Hiroya Fujisaki and Nick Campbell

1. Brief introduction - Hiroya Fujisaki
   Necessity of discussion and recommendation on common terminology in prosody research
2. Issues of teminology in several academic domains related to prosody research
   (a) acoustic/physical/physiological/psychological terms (Fujisaki)
   (b) linguistic/phonetic terms (Hirst)
   (c) semantic/pragmatic terms (Gibbon)
   (d) prosodic terms (Campbell)
3. lively discussion and debate from the floor

3.4 Thursday Session Four

4pm - 6pm : 3-4-oral (6 presentations)

- perception and production -

3.4.1 p.854

Sun-Ah Jun, Jason Bishop,
Implicit prosodic priming and autistic traits in relative clause attachment
Using the structural priming paradigm, the present study explores predictions made by the Implicit Prosody Hypothesis (Fodor 1998) by testing whether an implicit prosodic boundary generated from a silently read sentence influences attachment preference for a novel, subsequently read sentence. Results indicate that such priming does occur, although the patterns are highly dependent on individual differences in listeners' "autistic" traits.

3.4.2 p.859

Jennifer Cole, Tim Mahrt, José I. Hualde,
Listening for sound, listening for meaning: Task effects on prosodic transcription
The perception of prosodic structure (phrasal prominences and boundaries) may depend in part on acoustic information present in the signal and in part on meaning based on syntactic, semantic and pragmatic factors. Listeners may also be able to weigh acoustics and meaning to different degrees. We test naïve subjects' marking of prominences and boundaries in spontaneous American English under three different conditions, all of which involve listening to audio recordings and marking prominences and boundaries on a transcript. The three conditions differ in the instructions that transcribers were given. In one condition, subjects were instructed to transcribe prominence and boundaries based on meaning criteria, in a second condition they were told to transcribe based on criteria of acoustic salience. A third condition had more general instructions, without explicit reference to either meaning or acoustic perception. Our results show that subjects perform differently when focusing on meaning and on acoustics, especially for prominence marking, where many different words are selected as prominent under the two tasks. Boundary marking is more similar under the two instructions, with acoustic criteria resulting in a higher frequency of boundaries, but with boundaries marked largely on the same words in both tasks. When given non-specific instructions, performance was much more similar to that obtained under acoustic-based instructions. We report on agreement rates within and across conditions. This study has implications for models of prosody perception and the methodology of prosodic transcription.

3.4.3 p.864

Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, Jarek Krajewski,
Acoustic-Prosodic Characteristics of Sleepy Speech - Between Performance and Interpretation
When we address speaker states like sleepiness, two partly competing interests can be observed: within both applications and engineering approaches, we aim at utmost performance in terms of classification or regression accuracy - which normally means using a very large feature vector and a brute forcing approach. The other interest is interpretation: we want to know what tells apart atypical (here: sleepy) speech from typical (here: non-sleepy) speech, i.e., their respective feature characteristics. Both interests cannot be served at the same time. In this paper, we preselect a small number of easily interpretable acoustic-prosodic features modelling spectrum and prosody, based on the literature and on the general idea of sleepiness being characterised by relaxation. Performance obtained with these single features and this small feature vector is compared with the performance obtained with a very large feature vector; moreover, we discuss to which extent the features chosen model relaxation as sleepiness characteristic.

3.4.4 p.869

Juraj Šimko, Štefan Beňuš, Martti Vainio,
Hyperarticulation in Lombard speech: A preliminary study
Over the last century researchers collected a considerable amount of data reflecting properties of the Lombard speech, i.e., speech in a loud environment. The documented phenomena include effects on intensity, fundamental frequency, spectral tilt, speech rate and articulationm. Relatively little attention has been paid to the effects on relative extent of movement of individual articulators. In an attempt to fill in this gap we present a preliminary analysis of EMA data collected in increasing levels of babble noise. We introduce HH-index as a measure of overall relative activity of articulators. Our results indicate a non-linearity of the effect of noise on articulatory movement and quantitatively different effects on the movement extent for different groups of articulators. The effects of noise are compared with those brought out by other techniques for eliciting articulatory variation. We also discuss possible application of Lombard speech as an elicitation paradigm for studies of hyperarticulation.

3.4.5 p.874

Susanne Schötz, Joost van de Weijer,
A Study of Human Perception of Intonation in Domestic Cat Meows
This study examined human listeners' ability to classify domestic cat vocalisations (meows) recorded in two different contexts; during feeding time (food related meows) and while waiting to visit a veterinarian

(vet related meows). A pitch analysis showed a tendency for food related meows to have rising F0 contours, while vet related meows tended to have more falling F0 contours. 30 listeners judged twelve meows (six of each context) in a perception test. Classification accuracy was significantly above chance, and listeners who had reported previous experience with cats performed significantly better than inexperienced listeners. Moreover, the two food related meows with the highest classification accuracy showed clear rising F0 contours, while clear falling F0 contours characterised the two vet related meows that received the highest classification accuracy. Listeners also reported that some meows were very easy to classify, while others were more difficult. Taken together, these results suggest that cats may use different intonation patterns in their vocal interaction with humans, and that humans are able to identify the vocalisations based on intonation.

3.4.6 p.879

Chunyue Zhu, Toshiyuki Sadanobu,
Observation of so-called "pursed-lip" and "curled-lip" utterances in Japanese, using video and MRI images
The Japanese language includes utterances described by the idioms "speaking with pursed lips" and "speaking with curled lips." This study employs video and MRI imaging to examine the articulatory characteristics of these utterances ("utterances P" and "utterances C", respectively) by comparing their articulation with that of "unmarked" utterances ("utterances U"). Through doing so, we arrive at the following four conclusions: (1) For the articulation of utterance P, the lips are projected outward, and rounded by expanding in the vertical direction and narrowing in the horizontal direction. (2) For the articulation of utterance C, curling the lips is not an absolute requirement. The articulation of utterance C is similar with that of utterance P in that the lips are projected outward and rounded. (3) Utterances P and C differ in two points: (a) Lips projection accompanies the lower jaw projection only in utterances P; (b) Lips in utterance P is wider than those in utterance C. (4) The shapes the lips make in utterance P, utterance C, and utterance U can be described as a circle, a horizontal rectangle, and a horizontal oval, respectively. (5) There are many facts that contradict the accepted theory that "Rounding the lips causes both lips to project outward. In reaction to this movement, the surface of the tongue is pushed toward the rear" (Koizumi 1989).

3.5 Banquet - May 22nd

7pm - 9:30pm : banquet - **Trinity College Old Dining Hall** - ALL WELCOME!

4 Day Four - May 23rd

4.1 Friday Session One

9am - 10:30am : 4-1-poster (48 presentations)

- intonation and speaking style -

4.1.1 p.885

Page Piccinini, Marc Garellek,
Prosodic Cues to Monolingual versus Code-switching Sentences in English and Spanish
Code-switching offers an interesting methodology to examine what happens when two linguistic systems come into contact. In the present study two experiments were conducted to see if (1) listeners are able to anticipate code-switches in speech-in-noise, and (2) prosodic cues are present in the signal to warn of an upcoming code-switch. A speech-in-noise perception experiment with early Spanish-English bilinguals found that listeners are able to accurately identify words in code-switching sentences with the same accuracy as in monolingual sentences, even in highly degraded listening conditions. We then analyzed the stimuli used in the perception experiment, and found that the speaker does use different prosodic contours for code-switching productions as compared to monolingual productions. We propose that listeners use these code-switching specific prosodic contours to anticipate code-switches, and thus ease processing costs in word identification.

4.1.2 p.890

Simon Ritter, Timo B. Roettger,
Speakers modulate noise-induced pitch according to intonational context
Recent studies have shown that speakers systematically modulate properties of voiceless segments according to intonational context. More specifically, in the absence of fundamental frequency (F0), speakers appear to adjust the Center of Gravity (CoG) and the intensity of voiceless fricatives to convey the impression of pitch. In line with these findings, the present production study extends earlier work and investigates noise-induced properties of fricatives, modulated by the intonation context. It is shown for German that the mean CoG and intensity of intended contours with a high boundary tone are higher than those produced for intended contours with a low boundary tone. Furthermore, looking at the development of CoG and intensity over the time course of the fricative, the trajectories corresponding to the boundary tones differ in intercept and slope, i.e. reveal a steeper fall in case of a corresponding falling tone.

4.1.3 p.895

Albert Rilliard, Donna Erickson, Takaaki Shochi, Joõ Moraes,
US English attitudinal prosody performances in L1 and L2 speakers
Expressive behavior linked to paralinguistic meanings finds grounds in codes proposed as universals, as well as in culture-specific conventions. This study observes performances in such kinds of attitudinal prosody for USA English, produced by L1 and L2 speakers. The results show that the observed variance is linked to individual competence, to the linguistic context, and to the cultural background of the speakers. They also show that the code used to express a given speech act, code learned in the L1 language by L2 speakers of English, may be used in their L2 language. For some of these expressions, L2 speakers received higher scores than L1 speakers, suggesting that expressions conventionalized in a foreign language, are adequately fulfilling not-conventionalized expressions in the L1 culture.

4.1.4 p.900

András Beke, György Szaszák, Viola Váradi,
An Automatic Hierarchical Multiple Level Phrase segmentation approach for Spontaneous speech
The present paper investigates automatic prosodic phrasing of spontaneous speech: a two-step segmentation technique is presented, based on unsupervised learning. In the first step, the Intonational Phrases (IP) are detected automatically based on speech energy, spectral centroid and a double-thresholding technique. In the second step, Phonological Phrases (PP) are identified within the IPs. As acoustic features, F0, overall energy and vowel duration are investigated. An adaptive thresholding method is used based on Kullback-Leibler divergence computed in an autocorrelative manner for the feature streams. For Hungarian spontaneous speech, a phrasing accuracy of over 80% can be reached when comparing to a hand-labelled reference phrasing. It is found that in Hungarian sponatenous speech, F0 and energy play an essential role in IP level phrasing, whereas PP level phrasing is most effective using F0 related features alone. Vowel durations are shown not to contribute to prosodic phrasing in Hungarian. Although the evaluation targets the Hungarian language, the applied method is universal and can be easily adapted for other languages. Index Terms: speech synthesis, unit selection, joint costs.

4.1.5 p.905

Katelyn Eng, Beverly Hannah, Keith Leung, Yue Wang,
Effects of auditory, visual and gestural input on the perceptual learning of tones
Research has shown that audio-visual speech information facilitates second language (L2) speech learning, yet multiple input modalities including co-speech gestures show mixed results. While L2 learners may benefit from additional channels of input for processing challenging L2 sounds, multiple resources may also be inhibitory if learners experience excessive cognitive load. The present study examines the use of metaphoric hand gestures in training English perceivers to identify Mandarin tones. Native Mandarin speakers produced tonal stimuli with simultaneous hand gestures mimicking pitch contours in space. The English participants were trained to identify Mandarin tones in one of four modalities: audio-only, (AO), audio-visual (AV, speaker voice and face), audio-gesture (AG, speaker voice and hand gestures) and audio-visual-gesture (AVG). Results show significant improvements in tone identification from pre-

to post-training tests across all four training groups, demonstrating that gestural as well as visual articulatory information may facilitate tone perception. However, further analyses with individual tones reveal some group differences. Most noticeably, the AVG group had a slower learning curve during training compared to the other trainee groups for Tone 4, the least accurately identified tone, indicating a negative effect of multiple input modalities on the perception of difficult L2 sounds. In contrast, for Tones 2 and 3, the AG group revealed slower learning effects compared to the AV group, presumably because of the similar gestural trajectories for these two tones, which made the gestural input less distinct. Overall, the results suggest a positive role of gestures in tone identification, one that may also be constrained by phonetic and cognitive demands.

4.1.6 p.910

Céline De Looze, Daniel Hirst,
The OMe (Octave-Median) scale: a natural scale for speech melody.
Fundamental frequency, the primary acoustic correlate of speech melody, is generally analysed and displayed using a linear scale (Hertz) or a logarithmic one, generally in semitones and usually offset to an arbitrary reference level such as 100 Hz. In this paper we argue that a more natural scale for analysing speech is the OME (Octave-MEdian) scale, using the octave (o) as the basic unit, offset to the median value of the speaker's range. We present results showing that a reasonable estimate of a speaker's pitch range can be obtained directly from the median.

4.1.7 p.915

Nigel Ward,
Automatic Discovery of Simply-Composable Prosodic Element
As a way to discover the elements of prosody, Principal Component Analysis was applied to several dozen contextual prosodic features computed at 600,000 timepoints in dialog data. The results suggest that English has at least several dozen prosodic patterns, each with its own communicative function.

4.1.8 p.920

Jan Volín, Eliška Churaňová, Pavel Šturm,
P-centre Position in Natural Two-Syllable Czech Words
The ability to lock motor activity oscillator with external acoustic events is typical of various forms of human behaviour. Previous research showed that the beginning of an action is not necessarily the beginning of the rhythmic phase and led to the concept of p-centres. We present an experiment with 18 natural two-syllable Czech words spoken in synchrony with metronome beats by 18 subjects. Complexity of the consonantal onset and the type of coda together with distinctive phonological vowel length were carefully controlled to reveal a complex but comprehensible relationship between the word structure and phase locking.

4.1.9 p.925

Hyun Kyung Hwang, Satoshi Ito,
Correlation between prosody and epistemic bias of negative polar interrogatives in Japanese
The study investigates the correlation between prosodic patterns and epistemic bias observed in Japanese negative polar interrogatives, with special attention given to the perceptual and functional aspects of the correlation. The result of a naturalness rating test and a comprehension test demonstrate that listeners perceive the matching interrogative-answer pairs more natural, compared to the conflicting pairs. Also, it is revealed that the prosodic patterns successfully guide listeners to identify the epistemic bias of negative polar interrogatives.

4.1.10 p.929

Einar Meister, Lya Meister,
L2 production of Estonian quantity degrees
The Estonian quantity system involves three contrastive patterns referred to as short (Q1), long (Q2)

and overlong (Q3) quantity degrees. Our previous studies have shown that for L2 learners the distinction between Q2 and Q3 is a difficult task in both production and perception. While Q1 and Q2 structures are always distinguished in the orthography, this is not the case in most Q2 and Q3 words excluding the words with plosives between first and second syllable vowels. Thus, the orthography might be the reason for the use of the same production pattern for both Q2 and Q3. The current paper studies the role of L2 orthographic input on the L2 production of Estonian quantity degrees by two groups of subjects with different language backgrounds: Finnish and Russian. The material used in the study involves word structures with and without orthographic manifestation of quantity contrasts. The results confirm the role of Estonian orthography on the L2 pronunciation, however, the two L2 subject groups show different prosodic patterns.

4.1.11 p.934

Rachel Steindel Burdin,
Variation in list intonation in American Jewish English
Yiddish-influenced intonation has been previously noted as a potential defining characteristic of American Jewish English, and list intonation was identified as a possible area of differentiation. However, apart from remarks in general descriptions of Standard American English (SAE) prosody, a systematic study of list intonation has not been conducted in SAE. In this study, lists were defined, and extracted from sociolinguistic interviews with Jewish women with varying degrees of exposure to Yiddish. The lists were then ToBI annotated. Speakers from different language backgrounds differed significantly in their use of contours, boundary tones and pitch accents on list items, with speakers with less exposure to Yiddish using more of the standard English contour (H* H-L%) than speakers with more exposure to Yiddish. Yiddish bilinguals were more likely to use a rise fall contour (L+H* L-L%), fewer H-L% boundary tones and H* pitch accents, and more rising pitch accents (L+H* and L*+H) than non-bilinguals. In addition, speakers of all language backgrounds used a variety of list intonations, showing the need for more systematic study into the uses and meanings, social and otherwise, of list intonations.

4.1.12 p.939

Bogdan Ludusan, Stefan Ziegler, Guillaume Gravier,
Is Syllable Stress Information Robust for ASR in Adverse Conditions?
This paper presents a study on the robustness of stress information for automatic speech recognition in the presence of noise. The syllable stress, extracted from the speech signal, was integrated in the recognition process by means of a previously proposed decoding method. Experiments were conducted for several signal-to-noise ratio conditions and the results show that stress information is robust in the presence of medium to low noise. This was found to be true both when syllable boundary information was used for stress detection and when this information was not available. Furthermore, the obtained relative improvement increased with a decrease in signal quality, indicating that the stressed parts of the signal can be considered islands of reliability.

4.1.13 p.944

John Dalton, John Kane, Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl,
GlóRí - the Glottal Research Instrument
This papers presents GlóRí - the glottal research instrument. GlóRí is a speech analysis interface which offers a exibility and multiplicity of approaches to voice analysis. The system allows for fully automatic processing, for instance for analysis of large corpora. However, for more ne-grained studies, which may require precise voice source measurements, the systems facilitates manual optimisation of parameter settings. The present paper highlights the main features of the GlóRí system and provides illustrations of the usefulness of this approach.

4.1.14 p.949

Bettina Braun, Muna Pohl, Katharina Zahner,
Speech segmentation is modulated by peak alignment: Evidence from German 10-month-olds
In two headturn preference experiments, we tested whether German infants' segmentation strategies are

sensitive to the position of a pitch peak relative to the stressed syllable. Specifically, we compared target words with early-peak accents (where the pitch peak is early with respect to the stressed syllable, i.e. H+L* accents) and medial-peak accent (where the pitch peak is aligned with the stressed syllable, i.e. H* accents). Such differences in accent type signal mostly pragmatic distinctions, such as the difference between new and recoverable information. We familiarized infants with target words with one of the two intonation conditions that were embedded in sentences. We measured looking times to lists of trochaic part-words that were embedded in target words or were novel to them. Results showed an effect of familiarity only in the medial-peak condition, suggesting that infants at 10 months of age are very sensitive to pitch information for segmenting running speech.

4.1.15 p.954

Hae-Sung Jeon,
The Perception of Korean Boundary Tones by First and Second Language Speakers
This paper reports an experiment which investigated the perception of prosody in Korean or non-word utterances by native Korean speakers and English learners of Korean. Listeners rated the degrees of positivity and excitement of resynthesized utterances with different pitch ranges and durations. The results revealed no significant differences between the two groups of listeners. The variations in pitch range and duration had systematic effects on the ratings. However, the interactions between various factors suggest that the mapping between prosodic shapes and their paralinguistic meaning is not straightforward.

4.1.16 p.959

Irena Yanushevskaya, John Kane, Céline De Looze, Ailbhe Ní Chasaide,
The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps
This paper describes the distribution of pitch patterns and communicative types in the interpausal units (IPUs) preceding pause or gap silences extracted from a corpus of spontaneous speech as part of our work towards automatic prediction of turn-taking in dialogue interaction. IPUs preceding speaker change ('Gaps') and IPUs preceding silence where the same speaker continues talking ('Pauses') were selected in the course of automatic extraction of pause/gap silences in dyadic dialogue interactions. A listening test was conducted to estab-lish 'human predictable' pause/gap data sets which were sub-sequently manually annotated in terms of pitch patterns and communicative types. Overall, the Gaps and Pauses subsets show differentiation in terms of both their communicative types and pitch tunes. Declaratives and Questions are mainly found in Gaps, whereas in Pauses we mainly find Hesitations and Incomplete Declaratives. Gaps are generally characterised by falling or rising pitch patterns, whereas in Pauses a large proportion of speech samples are realised with level pitch. Classification experiments reveal strong discrimination of pauses and gaps for both prosodic and functional annotation labels.

4.1.17 p.964

Holly S.H. Fung, Peggy P.K. Mok,
Realization of Narrow Focus in Hong Kong English declarativesa Pilot Study
Narrow focus, i.e., focus on one word, is realized differently in native English and Cantonese. While it is signaled primarily by on-focus F0 changes such as F0 range expansion in English, it is marked essentially by lengthening of duration in Cantonese. Another difference is the pitch of the post-focus elements. While native English demonstrates post-focus F0 compression, Cantonese shows no significant post-focus pitch change. To investigate how narrow focus is realized in Hong Kong English (HKE), an emergent variety of English spoken by native speakers of Cantonese in Hong Kong, a controlled production experiment was conducted with 8 HKE speakers. Results showed that while the HKE speakers did realize foci with significant on-focus F0 range expansion, they exhibited no post-focus compression.

4.1.18 p.969

David Abelman, Robert Clark,
Altering speech synthesis prosody through real time natural gestural control
This paper investigates the usage of natural gestural controls to alter synthesised speech prosody in

real time (for example, recognising a one-handed beat as a cue to emphasise a certain word in a synthesised sentence). A user's gestures are recognised using a Microsoft Kinect sensor, and synthesised speech prosody is altered through a series of hand-crafted rules running through a modified HTS engine (pHTS, developed at Université de Mons). Two sets of preliminary experiments are carried out. Firstly, it is shown that users can control the device to a moderate level of accuracy, though this is projected to improve further as the system is refined. Secondly, it is shown that the prosody of the altered output is significantly preferred to that of the baseline pHTS synthesis. Future work is recommended to focus on learning gestural and prosodic rules from data, and in using an updated version of the underlying pHTS engine. The reader is encouraged to watch a short video demonstration of the work at http://tinyurl.com/gesture-prosody.

4.1.19 p.974

Xiaoluan Liu, Yi Xu,

Body size projection by voice quality in emotional speechEvidence from Mandarin Chinese

This study attempts to extend the line of research on using body size projection theory to account for emotional speech. It is predicted by the theory that anger is expressed by projecting a large body size with low pitch, rough voice and long vocal tract; happiness is expressed by projecting a small body size with high pitch, breathy voice and short vocal tract. Ten native speakers of Mandarin with drama training background recorded sentences in happy, angry, disgust and neutral emotions. We used multiple measurements to assess voice quality, formant dispersion (as an indicator of vocal tract length) and pitch. The results show clear support for the body size projection theory in voice quality, with anger and disgust associated with pressed and rough voice while happiness with breathy voice. But the results of formant dispersion and pitch demonstrate no clear directions. While the study is the first to show clear speech production support for the body size projection theory with voice quality data, the equivocal results of formant and pitch call for improvement in method of emotion elicitation in the laboratory.

4.1.20 p.978

Jan Michalsky,

Scaling of Final Rises in German Questions and Statements

Although certain intonation contours occur more frequently with German questions than with German statements, there is evidence that the semantics of intonational phonology operates on a more abstract level [1][2][3][4]. Hence, it is unlikely that there are pitch patterns in German that are exclusively used in interrogatives. Rather, intonational signaling of interrogativity can be regarded as resulting from the interaction between tonal and phonetic features. The tonal structure provides abstract semantic features, which are modified by paralinguistic features through phonetic realization [5]. This paper deals with the question which phonetic features may serve as cues to interrogativity in German. We report a reading task that was designed to elicit utterances that have phonologically identical nuclear rising pitch contours but differed by pragmatic function, serving either as a question or a statement. The observed absolute and relative scaling of nuclear and prenuclear tonal targets suggests that questions differ from statements by larger f0 excursions of nuclear rising contours, whereas the scaling of prenuclear accents does not substantially contribute to the expression of interrogativity. We conclude that phonetic cues to interrogativity in German are mainly realized through scaling and are restricted to the nuclear part of the intonational phrase.

4.1.21 p.983

Grace Kuo,

Processing Prosodic Boundaries in Natural and Filtered Speech

The prosody of an utterance can carry information that is critically important to understand the meaning of a sentence. In addition, previous studies have shown that listeners are able to detect major prosodic boundaries in their native language in stimuli whose segmental information has been removed, such as low-pass filtered [1][2] and hummed speech [2][3][5]. The present boundary strength rating study is conducted on native and non-native speakers to Taiwanese and Swedish, in an attempt to observe native and non-native speakers' accuracy in judging the upcoming boundary size in natural and filtered speech. 36 Taiwanese and American English speakers were recruited for the rating task whose stimuli consisted of Taiwanese and Swedish utterances from three prosodic boundary types (word boundary, phrase/tone sandhi group boundary, and Intonation Phrase boundary). In Experiment 1, participants rated the

upcoming boundary strength on a slider for filtered speech stimuli. In Experiment 2, they rated the boundary strength for natural speech stimuli. The results show that both non-native speakers could accurately predict the upcoming prosodic boundary type in both natural and filtered speech. The acoustic analyses of duration, f0 range, f0 median, spectral tilt, and harmonics-to-noise ratio reveal that non-native speakers use these prosodic cues to make their judgment, however, they put different emphasis on different cues when they were presented with stimuli of different qualities (natural vs. filtered) and lengths.

4.1.22 p.987

Malin Svensson Lundmark,
Constant Tonal Alignment in Swedish Word Accent II
Studies on accentual tonal alignment of intonation languages suggest that L in rising (LH) pre-nuclear accents anchors with a specific point in the segmental string, while the timing of H varies. This study investigates if lexical accents, too, exhibit a constant alignment by testing the South Swedish word Accent II. When under the strain of tempo variability the L-target was found not to be anchored with syllable onset. The results were not fully conclusive regarding H, but no clear evidence was found against anchoring of H, which could mean that H is an important phonological event in Accent II, while L is not.

4.1.23 p.992

Antonio Origlia, Francesco Cutugno,
A simplified version of the OpS algorithm for pitch stylization
In this work we present a new version of our previously published Optimal Stylization (OpS) algorithm for pitch stylization. Here we give a better perceptual representation of the pitch curve for linguistics research. While the OpS algorithm produced good stylizations for naive listeners, when deployed in a prosodic analysis tool, we observed that, under specific conditions, important details were missed in the stylized curve to an expert's ear. Changes introduced in the dynamic tonal perception model to solve these problems resulted in a simpler and more robust model. We show how the new version of the OpS algorithm is able to recover these situations while not significantly altering the original OpS curves.

4.1.24 p.997

Hiroaki Hatano, Carlos Ishi, Miyako Kiso,
Interpersonal factors affecting tones of question-type utterances in Japanese
The purpose of this paper is to clarify the interpersonal factors affecting phrase final tones of question-type utterances in Japanese daily conversations. We extracted question-type utterances ending with final particles from our dialogue speech database and classified them into two categories according to the degree of information request. Prosodic features were then analyzed by focusing on phrase final F0 movement and pitch reset. Analysis results indicated that F0 rising and falling degrees increase when the speaker express a close attitude to the dialogue partner, such as in conversations among family members and infant-directed speech. In addition, the presence of pitch reset in the phrase final was found to have functions of relieving the speaker's tension, when the dialogue partners have distant relationship.

4.1.25 p.1002

George Christodoulides, Cédric Lenglet,
Prosodic correlates of perceived quality and fluency in simultaneous interpreting
This study explores the relationship between prosodic features specific to simultaneous interpreting and listeners' perception of the fluency and accuracy of interpreting, as well as their comprehension of the source speech. Two groups of participants (47 subject experts and 40 non-experts) listened to a 20-minute lecture in German, along with its interpretation into French under two conditions (the actual interpretation, or a read-aloud rendition of the same text by the same interpreter) and answered comprehension and rating questions. The prosodic features of the two conditions were analysed, confirming differences regarding the temporal organisation of speech, disfluencies, pitch register and the interface between prosody and syntax. Our results suggest that interpreting-specific prosodic features affect the perception of fluency, which in turn affects the perception of accuracy; however the impact on listeners who enjoy relevant contextual knowledge is less pronounced.

4.1.26 p.1007

Ghania Droua-Hamdani, Sid-Ahmed Selouani, Yousef A. Alotaibi,
Rhythm analysis in Arabic L2 speech
This paper investigates rhythm speech metrics in Modern Standard Arabic. The corpus -West Point- includes recordings of native and non-native (L2) speakers. The experiment examines the rhythm metric tendencies of L2 speech using PVI and IM models. The study describes also the application of CCI (Control/Compensation Index) to the corpus. Variation in rhythm metrics by focusing on between-speaker differences such as gender of speakers is also studied.

4.1.27 p.1012

Rasmus Dall, Junichi Yamagishi, Simon King,
Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation
In this paper we present evidence that speech produced spontaneously in a conversation is considered more natural than read prompts. We also explore the relationship between participant's expectations of the speech style under evaluation and their actual ratings. In successive listening tests subjects are presented with either spontaneously produced, read aloud or written sentences, and are asked to rate the naturalness of each sentence with either instructions toward conversational, reading or general naturalness. It was found that, when presented with spontaneous or read aloud speech, participants consistently rated spontaneous speech more natural - even when asked to rate naturalness in the reading case. Presented with only text, participants generally preferred transcriptions of spontaneous utterances, except when asked to evaluate naturalness in terms of reading aloud. This has implications for the application of MOS-scale naturalness ratings in Speech Synthesis, and potentially on the type of data suitable for use both in general TTS, dialogue systems and specifically in Conversational TTS, in which the goal is to reproduce speech as it is produced in a spontaneous conversational setting.

4.1.28 p.1017

Hao Liu, Yi Xu,
A Simplified Method of Learning Underlying Articulatory Pitch Target
Previous research has shown that parameters of the quantitative Target Approximation model (qTA) proposed by Prom-on and Xu can be directly extracted from natural speech with high accuracy through analysis-by-synthesis implemented in PENTAtrainers. While this may raise the possibility that PENTAtrainers actually simulate natural acquisition of prosody production, it is questionable that the human brain actually replicates the full articulatory mechanics represented by qTA in order to learn and control prosody production. In this paper we explore if a much simpler function can be used to extract at least some of the qTA parameters. We first managed to reduce the number of qTA parameters from three to two by evaluating their relative sensitivity. We then tested a pursuit function that learns only pitch target height and slope. Using a corpus of Mandarin utterances varying in lexical tone and focus, we show that parameters learned by the pursuit function can be used in qTA synthesis to generate F0 contours closely resembling those generated with parameters learned with qTA-based analysis-by-synthesis, with the advantage of having a much simpler learning algorithm. These results suggest that it is possible to learn articulatory control parameters for prosody without fully replicating the mechanical process itself.

4.1.29 p.1022

Maria Del Mar Vanrell, Meghan E. Armstrong, Pilar Prieto,
The role of prosody in the encoding of evidentiality
The overarching goal of this paper is to advance on the understanding of how evidential meanings are expressed in natural languages. Specifically, we aimed to investigate what type of meaning was encoded in yes-no questions through the combination of the question particle (QP) que 'that' and the nuclear intonational pattern L+H* L% in Majorcan Catalan yes-no questions (i.e., Que és un llibre? L+H* L% 'QP-It's a book?'), and to understand any temporal information that might be encoded through this construction. Several complementary research methods were used to address our question: the Discourse Completion Task, an acceptability task and a multiple-choice questionnaire. The results show that three types of information are encoded in QP que_L+H* L% questions: sentence modality, inference based on direct evidence and immediacy of the evidence.

4.1.30 p.1027

Pierre-Edouard Honnet, Alexandros Lazaridis, Jean-Philippe Goldman, Philip N. Garner,
Prosody in Swiss French Accents: Investigation using Analysis by Synthesis
It is very common for a language to have different dialects or accents. The different pronunciations of the same words is one of the reasons for the different accents, in the same language. Swiss French accents have similar pronunciation to standard French, but noticeable differences in prosody. In this paper we investigate the use of standard French synthetic acoustic parameters combined with Swiss French prosody in order to evaluate the importance of prosody in modelling Swiss French accents. We use speech synthesis techniques to produce standard French pronunciation with Swiss French duration and intonation. Subjective evaluation to rate the degree of Swiss accent was conducted and showed that prosody modification alone reduces perceived difference between original Swiss accented speech and standard French coupled with original duration and intonation by 29%.

4.1.31 p.1032

Ya Li, Jianhua Tao, Keikichi Hirose, Wei Lai, Xiaoying Xu,
Hierarchical stress generation with Fujisaki model in expressive speech synthesis
This paper introduces a hierarchical stress generation for expressive speech synthesis. In the previous study, we proposed a novel hierarchical Mandarin stress modeling method, and the text-based stress prediction experiments demonstrates a reliable stress assignment can be obtained from textual features. However, the stress model should be further verified to be an effective and efficient prosody model in a Text-to-Speech system. In this work, Fujisaki model known as an ideal global representation of prosody is adopted to construct the pitch contours. To illustrate the effect of stress model, the Fujisaki model parameters are automatically predicted by the textural feature with and without stress information. The synthetic speech sounds more natural than that without stress modeling. The RMSE of the pitch contour and the feature importance analysis also show stress information can improve the pitch modeling. This work offers a promising method to accurate pitch modeling for Mandarin expressive speech synthesis.

4.1.32 p.1037

Frank Zimmerer, Jeanin Jügler, Bistra Andreeva, Bernd Möbius, Jürgen Trouvain,
Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers.
This article presents preliminary results indicating that speakers have a different pitch range when they speak a foreign language compared to the pitch variation that occurs when they speak their native language. To this end, a learner corpus with French and German speakers was analyzed. Results suggest that speakers indeed produce a smaller pitch range in the respective L2. This is true for both groups of native speakers. A possible explanation for this finding is that speakers are less confident in their productions, therefore, they concentrate more on segments and words and subsequently refrain from realizing pitch range more native-like. For language teaching, the results suggest that learners should be trained extensively on the more pronounced use of pitch in the foreign language.

4.1.33 p.1042

Tomoyuki Mizukami, Hiroya Hashimoto, Keikichi Hirose, Daisuke Saito, Nobuaki Minematsu,
Selection of Training Data for HMM-based Speech Synthesis from Prosodic Features - Use of Generation Process Model of Fundamental Frequency Contours
-Generation process model of fundamental frequency (F0) contours is ideal to represent global movements of F0's keeping a clear relation with back-grounding linguistic information of utterances. Using the model, improvements of HMM-based speech synthesis are expected. A new method is developed to cope with erroneous F0's of utterances included in HMM training corpus. F0 extraction errors not only cause wrong F0's, but also degrade segmental features of synthetic speech, since they affect the over-all accuracy of speech analysis. The method is to exclude speech segments from HMM training, where extracted F0's are largely different from those generated by the generation process model. Experiments on speech synthesis showed a clear improvement in synthetic speech quality when phoneme-base exclusion is conducted with a properly selected threshold.

4.1.34 p.1047

Alexandros Lazaridis, Pierre-Edouard Honnet, Philip N. Garner,
SVR vs MLP for Phone Duration Modelling in HMM-based Speech Synthesis
In this paper we investigate external phone duration models (PDMs) for improving the quality of synthetic speech in hidden Markov model (HMM)-based speech synthesis. Support Vector Regression (SVR) and Multilayer Perceptron (MLP) were used for this task. SVR and MLP PDMs were compared with the explicit duration modelling of hidden semi-Markov models (HSMMs). Experiments done on an American English database showed the SVR outperforming the MLP and HSMM duration modelling on objective and subjective evaluation. In the objective test, SVR managed to outperform MLP and HSMM models achieving 15.3% and 25.09% relative improvement in terms of root mean square error (RMSE) respectively. Moreover, in the subjective evaluation test, on synthesized speech, the SVR model was preferred over the MLP and HSMMmodels, achieving a preference score of 35.93% and 56.30%, respectively.

4.1.35 p.1057

Decha Moungsri, Tomoki Koriyama, Takashi Nose, Takao Kobayashi,
Tone Modeling Using Stress Information for HMM-Based Thai Speech Synthesis
This paper describes a modeling technique of Thai tones for HMM-based speech synthesis. Tones are important prosodic features for tonal language including Thai because the phonetically same words but with different tones give different meanings. Although there have been several approaches to improving tone correctness of synthetic speech by considering tone types, another significant factor, stress, was not used explicitly for prosody modeling. We incorporate stress/unstress information into the framework of the HMM-based speech synthesis. Objective and subjective evaluation results show that the use of stress information improves the performance in Thai tone modeling.

4.1.36 p.1062

D. Gomathi, P. Gangamohan, B. Yegnanarayana,
Understanding the significance of different components of mimicry speech
Voice conversion systems aim at finding a transformation function using statistical models. Mimicry/ Voice imitation is a natural voice transformation technique which sounds convincing to the listeners. It thus seems advisable to study the transformation used by human beings who perform mimicry. The objective of this study is to examine the various components of speech that are modified during voice imitation. To transform a given speech utterance to sound like that of a target utterance, the process needs to be understood at both production and perception level. In this paper the importance of source and system parameters and also the significance of different components of speech that contribute to the perception of imitation are studied. A flexible analysis-synthesis tool is used to modify the features of natural utterance and convert it to imitated utterance. Perceptual studies are carried out to understand if the modified features contribute to imitation. The results show that a combination of features is varied by the imitator to achieve imitation and they vary depending on the target speaker.

4.1.37 p.1067

Hansjörg Mixdorff, Angelika Hönemann, Jeesun Kim, Chris Davis, Grégory Zelic,
The Cartoon Task Exploring Auditory-Visual Prosody in Dialogs
This paper introduces and evaluates a collaborative task designed to elicit auditory-visual dialogs. The task was based on the viewing of two versions of the same cartoon film that was edited so that in order to reconstruct the story information from two incomplete versions must be shared in a consecutive fashion. The aim of this design was to elicit a relatively balanced dialog between the two participants as the story is pieced together from the beginning to the end. The current paper describes the production of a corpus consisting of audio, video and motion capture data from 22 pairs of Australian English speaking participants, and presents results on turn-distribution and raw prosodic features. Our analysis showed that the task could produce relatively balanced dialogs although this was not the case for all pairs. Analysis of raw prosodic features did not suggest that convergence occurred over the conversation, but replicated earlier findings of similarity between partners as compared to others.

4.1.38 p.1072

Peggy P.K. Mok, Holly S.H. Fung, Jingwen Li,
A preliminary study on the prosody of broadcast news in Hong Kong Cantonese
Broadcast news is a distinctive register. Previous studies only provided some general descriptions of the prosodic features in broadcast news but with few concrete data. Most of them were also on English news. This study investigated the prosodic features of Cantonese TV broadcast news using acoustic data. Speech using the same materials from two groups was compared: eight Hong Kong professional TV news anchors, and a control group consisting of eight university students. The results show clear differences between the two groups in terms of speech rate, pitch range and variability of syllable duration (speech rhythm). It was found that the news anchors spoke significantly faster than the control group, also with an enlarged pitch range. They also produced more variability in syllable duration. There is clearly more prosodic variation in the news register than ordinary speech. Finally, we provide some possible reasons for these features, as well as directions for future studies.

4.1.39 p.1052

Elisabeth Delais-Roussarie, Ingo Feldhausen,
Variation in Prosodic Boundary Strength: a study on dislocated XPs in French
Three independently motivated types of information are usually assumed to influence prosodic boundary placement and to play a role in their relative strength: the morpho-syntactic structure, the information structure and the metrical complexity. The phonetic realization associated with the different boundary types (in particular IP and ip) is also assumed to vary. Based on data of clitic left-dislocations in French, we argue here that differences in the relative strength of the prosodic boundary occurring at the end of the dislocated XP (i.e. an intermediate (ip) or an intonational phrase (IP) boundary) cannot be derived in a straightforward manner from these three types of information. In a production experiment, where the syntactic and information structure were controlled, while the metrical complexity was varied, it appeared that the strength of the boundary occurring at the right edge of the dislocated object NP displayed a high degree of variability. In addition, the results indicate a lack of correlation between metrical complexity and boundary strength. The results lead us to argue that a sort of phonological neutralization occurs in certain textual contexts. This neutralization does not allow for distinguishing between intermediate and intonational phrase boundaries in all cases.

4.1.40 p.1076

Sébastien Le Maguer, Elisabeth Delais-Roussarie, Nelly Barbot, Mathieu Avanzi, Olivier Rosec, Damien Lolive,
Prosodic chunking algorithm for dictation with the use of speech synthesis
The aim of this paper is to present an algorithm that automatically segment a text in prosodic chunks for a dictation by conforming to the rules and procedures used in real settings to dictate a text to primary school children. A better understanding and modeling of these rules and procedures is crucial to develop robust automatic tools that could be used in autonomy by children to improve their spelling skills through dictation with the use of speech synthesis. The different steps used to derive the prosodic chunks from a given text will be explained through concrete examples. The proposal made here relies on the analysis of a corpus of 10 dictations given to children in French and French Canadian elementary schools, and more precisely during their first three years in elementary school (i.e., cycle 2 in the French school system). The phrasing observed in the data is described. It is thus simplified in order to develop an algorithm that automatically generates prosodic chunks from texts.

4.1.41 p.1081

Jitka Vaňková, Radek Skarnitzl,
Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality
This study focuses on short-term acoustic correlates of voice quality. It assesses the within-speaker stability (across different speaking styles) and between-speaker variability of measurements which compare the amplitudes of various spectral events H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2* and H1*-A3*. Although speakers do differ with regard to the compactness of the parameters in read and spontaneous speaking styles, the parameters H1*-H2*, H1*-A1* and H1*-A2* appear both considerably stable for one speaker

in different speaking styles and efficient in between-speaker comparisons. Though not directly applicable in forensic settings, these glottal parameters outperformed vowel formants in classification using LDA.

4.1.42 p.1086

Bénédicte Grandon, Hiyon Yoo,
Do Korean L2 learners have a "foreign accent" when they speak French? Production and perception experiments on rhythm and intonation
French and Korean are two languages with similar prosodic characteristics as far as rhythm and intonation are concerned. In this paper, we present the results of production and perception tests where we describe the prosodic characteristics of Korean L2 learners of French. The aim is to analyze the impression of "foreign accent" for two prosodic components (intonation and rhythm) of speech produced by Korean L2 learners of French and the perception of this "accent" by native listeners of French (L1). We show that the productions of Korean learners and French native speakers present minor differences but that they do no translate into cues for determining clearly the presence of a "foreign accent".

4.1.43 p.1091

Linda Garami, Anett Ragó, Ferenc Honbolygó, Valéria Csépe,
Prosodic processing in the first year of life: an ERP study
From early months of life prosody has a prominent contribution to segmentation: prosodic boundaries overlap with syntactic ones and facilitate the extraction of syntactic regularities both at word and at phrase level. Therefore, the long-term representation of rhythmic features of the native language, especially the stress templates derived from regularities are assumed to play a particular role in pre-lexical processing. We examined the nature of early stress representation in a language with a fixed stress pattern in an electrophysiological experiment (acoustic passive odd-ball paradigm, 10 month-olds: 28 infants; 6 month-olds: 21 infants, 400 items, deviant: p=20%) using bi-syllabic Hungarian pseudo-words to follow how prosodic features contribute to processing saliency and how word stress templates based on regularities may emerge. We used legally and illegally stressed stimulus both in standard and deviant positions in separate conditions. In the legal standard condition two mismatch responses (MMRs) temporally synchronized to each syllable could be recorded. On the contrary, in the illegal standard condition no significant response was found. It seems that language environment influences the processing of speech prosody and the MMR correlates of word stress processing are related both to saliency and to stress templates emerging during the first year of life.

4.1.44 p.1095

Lei He,
The Inadequacy of Rhythm Metrics to Quantify L2 Suprasegmental Characteristics
The study investigated the L2 speech rhythm of Chinese English speakers (L1 = Mandarin) using the metrics of deltaV, deltaC, %V, VarcoV, VarcoC, rPVI-C and nPVI-V. Five native speakers of American English and Mandarin were recruited to record five sentences in English. In addition, the Chinese speakers also recorded five Mandarin sentences. One-way ANOVAs were conducted to see if significant differences exist on each of the metrics among L1 English, L2 English and L1 Mandarin. Results show that the two L1's are categorically distinct on all metrics, conforming to the perceptually distinct rhythmicities of English and Mandarin. However, no significant differences were found between L1 and L2 English which have different intuitive rhythmicities, suggesting that the metrics are inadequate to capture the suprasegmental details that give the final make-up of speech rhythm.Finally, new directions of speech rhythm research and new applications of the rhythm metrics are sketched.

4.1.45 p.1100

Céline De Looze, Irena Yanushevskaya, John Kane, Ailbhe Ní Chasaide,
Pitch range declination and reset in turn-taking organisation
In this paper, we investigate how pitch range declination and reset contribute to turn-taking organisation. We (i) investigate the effect of the unit position in a turn on its pitch range as well as (ii) compare the difference in pitch range between consecutive units that are separated by a gap vs. a pause. We also (iii) test the effect of the number of speech units in a turn as well as the turn duration on the peak height

at the beginning of the turn. Our results suggest a pitch range declination trend between the Initial and Median speech units of a turn but a violation of this declination for the Final units of the turn. Consequently, the difference in two consecutive units' pitch range is found larger at pauses than at gaps. Our results also show that the higher the number of speech units in a turn or the longer the turn, the higher the peak height. Our findings particularly reveal that the distance between the pitch range level and its upper limit may be a salient cue in projecting the end of a turn. We discuss our findings along the debate on Projection and Reaction theories and that of Hard vs. Soft pre-planning of speech production, and address how these findings may be useful for human-machine interactions.

4.1.46 p.1105

Monica Dominguez, Mireia Farrús, Alicia Burga, Leo Wanner,
Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis
This paper deals with the adaptation of AuToBI annotation for speech synthesis purposes. AuToBI is a tool that automatically detects and classifies the standard ToBI labels for American English. AuToBI annotation is performed word-by-word. However, a labeling of intonation patterns at the intonational phrase level is essential for the detection of the correlation between theme/rheme (thematicity) and prosody and also much more appropriate for speech synthesis applications that use various layers of linguistic annotation (syntax, semantic, information, and prosody structures), such that if used in speech synthesis applications, AuToBI's output would require a post-processing stage of the extracted labels. We present a procedure that includes an initial AuToBI annotation and the adaptation of the AuToBI output to a phrase-based annotation, following a set of determined rules. A further analysis of the correspondence between prosodic patterns and themacity structures is used to validate the results.

4.1.47 p.1110

Vasilisa Verkhodanova, Vladimir Shapranov,
Automatic Detection of Filled Pauses and Lengthenings in the Spontaneous Russian Speech
During automatic speech processing a number of problems appear, and among them there are such as speech variation and different kinds of speech disfluences. In this article an algorithm for automatic detection of the most frequent of them (filled pauses and sound lengthenings) based on the analysis of their acoustical parameters is presented. The method of formant analysis was used to detect voiced hesitation phenomena and a method of band-filtering was used to detect unvoiced hesitation phenomena. For the experiments on filled pauses and lengthenings detection a specially collected corpus of spontaneous Russian map-task and appointment-task dialogs was used. The accuracy of voiced filled pauses and lengthening detection was 82%. And accuracy of detection of unvoiced fricative lengthening was 66%.

4.1.48 p.1115

Aline Pessoa-Almeida, Alexsandro Meireles, Sandra Madureira, Zuleica Camargo,
Prosodic analysis of the speech of a child with cochlear implant
According to previous studies [1,6], acoustic and perceptual analysis can be considered useful clinical tools to investigate the speech characteristics of hearing impaired children (HIC). This research aimed at describing the perceptual and acoustic correlates of the speech samples from a HI and user of CI child (within the chronological age range of 5 years and 1 month and 7 years and 1 month), through the vocal quality and voice dynamics descriptions in two different moments. The speech samples were collected during speech therapy sessions. The perceptual analysis of the vocal quality was based on the Vocal Profile Analysis Scheme for Brazilian Portuguese (BP-VPAS - Camargo &Madureira, 2008). The recorded corpus was analyzed through the ExpressionEvaluator script (Barbosa, 2009) ran by Praat software v5.2.10. The measures, which were automatically extracted, comprised the fundamental frequency f0, first f0 derivative, intensity, spectral slope and long-term mean spectrum. The correlations found between the acoustic and perceptual data proved relevant for rehabilitation processes.

4.1.49 p.1119

Tatiana Luchkina, Jennifer Cole,
Structural and Prosodic Correlates of Prominence in Free Word Order Language Discourse
Production and perception experiments with native speakers of Russian, a free word order language, show

that prosody and change in word order are used to mark discourse-prominent constituents. Concurrent application of these cues to prominence is possible, as evident from distinctively higher f0 and intensity maxima, and duration values associated with ex- situ words, as well as their higher visibility in discourse. Distinctive acoustic-prosodic realization of ex-situ words may cue their relatively high informational load and discourse prominence, as well as (redundantly) signal that the word is left- or right-dislocated.

4.2 Friday Session Two

11am - 1pm : 4-2-oral (6 presentations)

- intonation -

4.2.1 p.1125

Jonathan Barnes, Alejna Brugos, Nanette Veilleux, Stefanie Shattuck Hufnagel
Segmental Influences on the Perception of Pitch Accent Scaling in English
In both tone and intonation systems, segmental context is known to influence production and perception of target F0 contours in various ways. Many languages, for example, prefer to realize critical F0 events during maximally sonorous intervals, either by varying the timing of pitch movements, or by virtue of distributional limitations on certain contour types. Current analytic practice, by contrast, routinely ignores segmental backdrop when estimating the perceptual efficacy of putative cues, such as F0 turning points, to tone scaling and timing patterns. Results of the perception study presented here argue that pitch accent scaling is best modeled using a weighted average of F0 sampled over a defined region of interest, and that individual sample weights are determined in part by the sonority of the segments from which they are taken. That is, samples from lower sonority segments contribute less to integrated scaling percepts than those from higher sonority segments. This model, called TCoG-F(requency), accounts for crosslinguistic tonal timing and distribution patterns in the literature, and underscores the danger of analyzing tonal phenomena completely apart from the segments that express them.

4.2.2 p.1130

Kristine M. Yu, Sameer Ud Dowla Khan, Megha Sundara,
Intonational phonology in Bengali and English infant-directed speech
We examined the phonetics and phonology of intonation of infant-directed speech (IDS) and non-IDS in story-reading in two typologically-divergent languages, English and Bengali. In addition to finding an increase in f0 range and variability in IDS, replicating previous work on IDS prosody, we found novel evidence that f0 manipulations in IDS are constrained by intonational phonology. Speakers in both languages used an increased proportion of tonal elements with higher tonal targets and more turning points in IDS, within the language-specific intonational grammar. The tonal elements showing increased use in IDS also were associated with marking topic and focus. Thus, phonetic changes in IDS may in part be induced by speakers' choices of phonological tonal elements, which in turn may be connected with choices about marking discourse structure.

4.2.3 p.1135

Eszter Varga, Zsuzsanna Schnell, Gabor Perlaki, Gergely Orsi, Mihály Aradi, Tibor Auer, Flora John, Tamás Dóczi, Samuel Komoly, Norbert Kovács, Attila Schwarz, Tamás Tényi, Róbert Herold, József Janszky, Réka Horváth,
Hemispheric lateralization of sentence intonation in left handed subjects with typical and atypical language lateralization: an fMRI study
Introduction: Prosody (as the melody of speech) is an important component of human social interactions. More specifically, linguistic prosody conveys meaning of speech through syllable, word, or sentence level stress and intonation. In the modern neuroimaging era the hemispheric representation of sentence intonation is widely investigated. Most of these studies suggest bilateral activations predominantly in the perisylvian language areas and in the subdominant homologues. However, there are some inconsistencies about the hemispheric representation and lateralization of linguistic prosody. These inconsistencies could be due to the lack of attention on the language lateralization of the subjects. Aims: The present study aims to investigate the hemispheric representation and lateralization of linguistic prosody with a

sentence intonation task in two groups of left handed subjects with typical and atypical language lateralization. Functional MRI was used to test the assumption that - according to the functional lateralization hypothesis - the representation of sentence intonation is predominantly lateralized within the language dominant hemisphere and the lateralization of sentence intonation is associated with language lateralization in both groups. Methods: Left handers were examined to create two groups of subjects with typical and atypical language lateralization. In all, 32 healthy subjects were evaluated with a standard verbal fluency task with fMRI in order to assess functional hemispheric language lateralization. In our final investigation the atypical group consisted of 8 subjects with right hemispheric language dominance (LI<-0.2) and the typical group also consisted of 8 subjects with left hemispheric language dominance (LI>0.2). Sentence intonation task was utilized to test linguistic prosody skills with fMRI. 49 pairs of sentences (18 pairs of neutral-neutral sentences, 10 pairs of interrogative-interrogative sentences, and 1 pair of interrogative-neutral sentence) were presented with an event-related design. Sentences were matched in terms of syntactic structure, semantic complexity and length and all were affectively neutral. In the fMRI data analysis interrogative pairs were compared to neutral pairs. Results: One of the main findings of our study is that subjects with both typical and atypical language lateralization activated the middle temporal gyrus (MTG) on the right side. The activation of the MTG on the right hemisphere is classically associated with the encoding of prosodic information. Furthermore, both groups recruited the frontal language areas only in the language-dominant hemisphere. Moreover, between-group comparison showed significantly stronger activations in subjects with typical language lateralization only in left sided language areas: pars triangularis of the inferior frontal gyrus, the superior frontal gyrus and the inferior parietal lobule. Conclusion: This finding is in accordance with the functional lateralization hypothesis of prosody, and suggests a correlation between linguistic prosody lateralization and language lateralization.

4.2.4 p.1139

Meghan Armstrong, Núria Esteve-Gibert, Pilar Prieto,

The acquisition of multimodal cues to disbelief

In this study, we examine how 3-, 4-, and 5-year-old Catalan-acquiring children are able to make use of the audio (intonational) and visual (facial gesture) modalities in the comprehension of speaker disbelief, as well as the role of a child's developing Theory of Mind. Our results suggest that in this case, facial gesture provides children with scaffolding for linguistic meaning. In addition, those children that passed a false belief task tended to perform better on the comprehension task in general. We discuss the implications of these findings for the study of intonational development.

4.2.5 p.1144

Amalia Arvaniti, Mary Baltazani, Stella Gryllia,

The pragmatic interpretation of intonation in Greek wh-questions

We experimentally investigated the pragmatics of two melodies commonly used with Greek wh-questions, L*H L-!H%, described as the default, and LH* L-L% considered less frequent and polite. We tested two hypotheses: (a) the !H%-ending melody is associated with information-seeking questions, while the L%-ending melody is pragmatically more flexible and thus appropriate also for non-information-seeking wh-questions expressing bias; (b) the !H%-ending melody, being more polite, is more appropriate for female talkers, all else being equal. In Experiment 1, comprehenders rated !H-ending and L%-ending versions of the same questions for politeness and appropriateness for the context in which they were heard (which favored either information-seeking or "biased" wh-questions). In Experiment 2, comprehenders heard the same questions and chose between two follow-up responses, one providing information, the other addressing the bias of the wh-question. Comprehenders rated !H%-ending questions more appropriate than L%-ending questions and judged the !H%-ending questions of female talkers more polite. They also chose information-providing answers more frequently after !H%- than L%-ending questions, but the preference was higher for female talkers and depended on comprehender gender. The results argue in favor of a compositional view of intonational meaning which depends not only on the tune but also on context, broadly construed.

4.2.6 p.1149

Pablo Arantes, Anders Eriksson,

Temporal stability of long term measures of fundamental frequency

We investigated long-term mean, median and base value of F0 to estimate how long it takes for their variability to stabilize. Change point analysis was used to locate stabilization points. In one experiment

stabilization points were calculated in recordings of the same text spoken in 26 languages. Average stabilization points are 5 seconds for base value and 10 seconds for mean and median. Variance after the stabilization point was reduced around 40 times for mean and median and more than 100 times for the base value. In other experiment, four speakers read each two different texts. Stabilization points for the same speaker across the texts do not exactly coincide as would be ideally expected. Average point dislocation is 2.5 seconds for the base value, 3.4 for the median and 9.5 for the mean. After stabilization, individual differences in the three measures obtained from the two texts are on average 2% on average. Present results show that stabilization points in long-term measures of F0 occur earlier than suggested in the previous literature..

4.3 Friday Session Three

2pm - 3:30pm : 4-3-plenary (1+3 presentations)

4.3.1 KeyNote 4

Anne Cutler - Invited Keynote:

Aspects of the suprasegmental structure of speech are famously subject to speaker choice. There is no obligatory location for accent in a sentence such as "She didn't run home"; speakers may say "SHE didn't run home" or "She DIDN'T run home" or "She didn't RUN home" or "She didn't run HOME", with different resulting inferences in each case. But do listeners also have any degree of choice in the auditory processing of this dimension of speech? This presentation will argue that they do, and support the argument with evidence from laboratory studies of spoken-word recognition, of semantic structure computation in spoken sentences, and of the processing of delexicalised prosodic signals.

4.3.2 p.1154

Meredith Brown, Laura Dilley, Michael Tanenhaus,
Probabilistic prosody: Effects of relative speech rate on perception of (a) word(s) several syllables earlier
Speech perception depends on the ability to rapidly accommodate considerable variability in speech rate. We present results from two eye-tracking experiments indicating that listeners use context speech rate to generate, maintain, and update probabilistic hypotheses about the timing and number of constituents in upcoming speech. Participants heard utterances containing polysyllabic nouns preceded by indefinite articles and followed by [s]-initial words (e.g. ...saw a raccoon slowly...). We altered the speech rate of the indefinite article and of the [s] with respect to surrounding context, manipulating the likelihood that the item would be perceived as singular (a raccoon) vs. plural (raccoons). Shorter indefinite articles elicited higher proportions of fixations to plural target pictures than longer articles both before and after the processing of [s], demonstrating that listeners made rapid use of prosodic cues to the presence or absence of the article. Importantly, fixations were also influenced by the duration of [s] relative to context speech rate. These findings suggest that listeners maintain and update provisional speech-rate hypotheses across multiple morphophonemic units. We interpret these results with respect to probabilistic approaches to spoken language understanding.

4.3.3 p.1159

Jill C. Thorson, James L. Morgan,
The role of intonation in early word recognition and learning
The motivation for our study is to investigate how English-acquiring toddlers are guided by the mapping between intonation and information structure during on-line reference resolution and in novel word learning tasks. We ask whether specific pitch movements (deaccented, monotonal, bitonal) more systematically predict patterns of attention and subsequent novel word learning abilities depending on the referring or learning condition (new, given, contrastive). Experiment 1 examines the attentional patterns of 18-month-old toddlers when referents are either new or given in the discourse, and carry one of the three pitch accent types. Contrary to previous work, results show increased attention to the target in the deaccented condition if the referent is new to the discourse. Also, both monotonal and bitonal pitch movements direct attention to the target even when the target is given. Thus, pitch type interacts with information structure in directing toddler attention. Experiment 2 tests two-year-olds in a novel word

learning task, varying pitch type and contrastiveness during learning. Preliminary results show that learning is aided when the novel word is introduced in contrast to a previous referent. Together, these two experiments demonstrate the role of pitch type and information structure in guiding attention and aiding early word learning.

4.3.4 p.1164

Rory Turnbull, Adam J. Royer, Kiwako Ito, Shari R. Speer,
Prominence perception in and out of context
The perception of prosodic prominence is known to be influenced by several distinct factors. In this study, we investigated the role of context, both global and local, in the prominence judgements of naïve listeners. Monolingual English listeners marked where they heard prominence on pairs of two-word phrases (e.g. *blue ball*, *green drum*). Stimuli varied in whether or not the first phrase implied a contrastive focus on the second phrase. We found clear evidence of a hierarchy of prominence across pitch accent types: L+H* >H* >!H* >unaccented. Additionally, we found that contrast status only affected prominence markings when the participants were made explicitly aware of the discourse context and were instructed to imagine themselves physically present to observe the conversation. This effect of global context suggests that information structure cannot be reliably interpreted in the absence of an established discourse context. Taken together, these results suggest that naive listeners are sensitive to prominence differences at levels corresponding to categorical annotations. Perception of a word's relative prominence was consistently influenced by phonetic and phonological factors, while pragmatic factors (such as contrast-evoking context) required more elaborate plausibility manipulations in order to affect prominence perception.

4.4 Friday Closing Session

Firewall Speeches . . .

Full Texts

5 Tuesday 1

Cue-based analysis of speech: implications for prosodic labelling systems

Over the past few decades it has become clear that an adequate account of systematic context-driven variation in word forms requires representations below the level of the abstract symbolic phoneme or even the allophone. One proposal for this sub-allophonic level of description is in terms of feature cues, such as the cues to articulator-free features and articulator-bound features proposed by Halle (1992) and by Stevens (2002), also assumed in the concept of enhancing cues in Stevens and Keyser (2010), Keyser and Stevens (2006) and Stevens, Keyser and Kawasaki (1986). This proposal of a level of representation of discrete feature cues, along with continuous-valued cue parameters, has the potential to bridge the gap between abstract symbolic categories of the phonology and the concrete spatial and temporal specifications that drive the articulatory-acoustic implementation of word forms in continuous communicative speech. Such an approach suggests that phonetic transcription might benefit from a focus on capturing the individual cues to feature contrasts that are realized in the speech signal. Does this approach to understanding phonetic variation in word forms have implications for prosodic labelling? We will explore this possibility, taking as our point of departure Arbisi-Kelm's (2006) proposal for labelling the separate correlates of prosodic disfluency in stuttered speech, and adapted by Brugos and Shattuck-Hufnagel (2012) for prosodic disfluencies in utterances produced by typical speakers. Our hypothesis is that variation in cue selection and cue parameter values is systematically governed by context, and that cue-level transcription may be needed to capture systematicity in the phonetic implementation of prosodic phonology as well as of lexical phonology.

Arbisi-Kelm, T. (2006). Intonation structure and disfluency detection in stuttering. Paper presented at LabPhon10, Paris

Brugos, A. and Shattuck-Hufnagel, S. (2012), A proposal for labelling prosodic disfluencies in ToBI. Presented at the Workshop for Advancing Prosodic Labelling, in conjunction with LabPhon 13, Stuttgart, July 2012

Halle, M (1992). Features. In Bright, W. (Ed.), Oxford International Encyclopedia of Linguistics 3, 207-212. New York: Oxford University Press.

Keyser, S.J. and Stevens, K.N. (2006) Enhancement and overlap in the speech chain. Language 82, 33–62.

Stevens, K.N. (2002), Toward a model for lexical access based on acoustic landmarks and distinctive features. JASA 111, 1872-1891

Stevens, K.N. and Keyser, S.J. (2010), Quantal theory, enhancement and overlap. Journal of Phonetics 38, 10–19

Stevens, K.N., Keyser, S.J. and Kawasaki, H. (1986), Toward a phonetic theory of redundant features. In J. Perkell and D.H. Klatt (Eds.), A symposium on invariance and variability of speech processes. Hillsdale, N.J.: Lawrence Erlbaum Associates

Prosodic Entrainment in Mandarin and English: A Cross-Linguistic Comparison

Zhihua Xia[1], *Rivka Levitan*[2], *Julia Hirschberg*[2]

[1]Jiangsu Normal University, Jiangsu, and Tongji University, Shanghai, P. R. China
[2]Computer Science Department, Columbia University, New York, USA
xzhlf@163.com, rlevitan@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Entrainment is the propensity of speakers to begin behaving like one another in conversation. We identify evidence of entrainment in a number of acoustic and prosodic dimensions in conversational speech of Standard American English speakers and Mandarin Chinese speakers. We compare entrainment in the Columbia Games Corpus and the Tongji Games Corpus and find similar patterns of global and local entrainment in both. Differences appear primarily in global convergence.
Index Terms: prosody, entrainment, discourse

1. Introduction

Entrainment — also known as *adaptation*, *accommodation*, or *alignment* — occurs in many dimensions of human-human conversation as people begin to act similarly to one another. This process is critical to humans' assessment of dialogue success and overall quality and to their evaluation of conversational partners [1, 2]. In a study of entrainment on gesture and facial expression, [3] found that subjects displayed strong unintentional entrainment and that greater entrainment led them to report liking their partner more and believing the interaction was progressing more smoothly. [4] found that degree of entrainment in lexical and syntactic repetitions occurring in just the first five minutes of a dialogue significantly predicted task success in studies of the HCRC Map Task Corpus.

In previous research on acoustic-prosodic indicators of entrainment ([5, 6, 7]), we found considerable evidence of entrainment in the Columbia Games Corpus. In this paper we examine cross-language and cross-cultural entrainment.[1] We compare results from our previous experiments on Standard American English (SAE) conversations to entrainment in Mandarin Chinese (MC) conversations collected in a similar setting in the Tongji Games Corpus. We compare entrainment in pitch, loudness, and speaking rate over all speakers and in female, male, and mixed-gender dialogue pairs.

Entrainment has been studied at the conversation level or at the turn level. At the conversation level, there may be evidence of an overall coordination of behavior despite local variation; at the turn level there may be turn-by-turn coordination, in which speakers closely match their partner's previous turn. Entrainment may also be measured in different ways, in terms of *similarity* over either level, *synchrony*, as behavior varies in tandem, although absolute values of features may be different, or *convergence*, as behaviors become more similar over time.

[1]Our previous findings on the Columbia Games Corpus are presented here for comparison and contrast with the new findings on the Tongji Games Corpus.

Our goal is to identify where subjects from these different language groups show evidence of each of these aspects of entrainment at the global and local levels. In Section 2 we describe the corpora we compare. In Section 3 we describe the acoustic and prosodic features we examined. In Section 4 we compare MC and SAE speaker entrainment at the global or conversational level. In Section 5 we make similar comparisons at the local or turn-by-turn level. In Section 7 we discuss our results and describe future research.

2. Corpora

2.1. Columbia Games Corpus

The SAE experiments in this work were conducted on the Columbia Games Corpus [8], a corpus of spontaneous, task-oriented speech between pairs of strangers. The corpus comprises twelve dyadic conversations elicited from thirteen native speakers of SAE (six female, seven male). Each pair of subjects played a set of computer games that required them to cooperate to achieve a mutual goal. Subjects were recorded in a sound-proof booth on laptops with a curtain between them. Neither could see the other's screen. In the Cards games, one speaker (the information *giver*) described the cards she saw on her screen, and her partner (the *follower*) attempted to match them to the cards on his screen. In the Objects games, one speaker (the *giver*) described the location of an object on her screen, and her partner (the *follower*) attempted to place the corresponding object in exactly the same location on his own screen. For each game, participants received points based on how exact a match was; they later were paid for each point. Each of the twelve sessions consists of two Cards games and one Objects game. Each session, on average, contains 45 minutes of dialogue. On average, each Cards game took 7.7 minutes, and each Objects game took 21.5 minutes. In total, the corpus consists of approximately nine hours of recorded dialogue. It has been orthographically transcribed and annotated with prosodic and turn-taking labels.

2.2. Tongji Games Corpus

The MC experiments in this study were conducted on theTongji Games Corpus. The corpus contains approximately 12 hours of spontaneous, task-oriented conversations between pairs of subjects comprising 115 conversations averaging 6 minutes between 70 pairs of speakers (40 female, 30 male). Subjects were randomly selected from university students with a National Mandarin Test Certificate level 2, with a grade of A or above to increase the likelihood that the Mandarin spoken in the corpus is standard. Recordings were made in a sound-proof

booth on laptops with a curtain between participants so that neither could see the other's screen. Two games were used to elicit spontaneous speech in the collection of the corpus. In the Picture Ordering game, one subject, the information *giver*, gave the other, the *follower*, instructions for ordering a set of 18 cards. When the task was completed, the same pair switched roles and repeated the task. In the Picture Classifying game, each pair worked together to classify 18 pictures into several categories by discussing each picture. Seventeen pairs played the Picture Ordering game; 39 pairs played the Picture Classification game; and 14 pairs played both games. The corpus was segmented automatically using SPPAS [9]. The automatic segments were manually checked and orthographically transcribed. Turns were identified by two PhD students specializing in Conversation Analysis.

3. Features and units of analysis

The smallest unit of analysis in this work is the *inter-pausal unit*, or IPU, defined as a pause-free segment of speech from a single speaker. A *turn* is defined as a maximal sequence of IPUs from a single speaker. For the SAE speakers 50ms was used as the minimal pause length and 80ms was used for the MC speakers, based upon the average length of stop gaps in each corpus. Each game conversation in the Tongji Corpus is also divided into 18 tasks, each of which involves the placing or classification of a single card. For our analysis, we include one randomly chosen conversation from each of the 70 speaker pairs for a total of 70 conversations among 99 speakers, since some speakers participated more than once with different partners. We compare seven acoustic-prosodic features in our comparison: intensity min, intensity mean, intensity max, f0 min, f0 mean, f0 max, and speaking rate (syllables/second). All seven were extracted from each IPU using Praat ([10]). We compare results from the MC subjects with our previous experiments on SAE speakers ([6, 7]), in which we looked at intensity mean, intensity max, f0 mean, f0 max, jitter, shimmer, noise-to-harmonics ratio (NHR), and speaking rate.

4. Global entrainment

We begin our analysis by considering entrainment globally, to see whether speakers are similar with respect to a given feature at the conversation level. We first look for evidence of global similarity, using a method proposed in [6]. For each speaker, we compute a *partner* similarity and a *non-partner* similarity. The first is the negated absolute difference between the two partners' values. The second is the negated absolute difference between a speaker and the averaged values for the non-partner speakers in the corpus. For the MC study, the non-partners are restricted to those of the same gender and conversational role (information giver or receiver) as the partner. If partner similarities are larger than the non-partner similarities for a given feature, we conclude that the speakers entrain on that feature. Using this method, we previously showed ([6]) that speakers of SAE showed evidence of *global* entrainment for intensity mean, intensity max, f0 max, and speaking rate. That is, for these four features, speakers were more similar to their partners than to the speakers in the corpus with whom they were never paired. For intensity mean and max, they were also more similar to their partners' speech than they were to the speech that they produced themselves in conversation with a different interlocutor.

Our comparison shows (Table 1) the same pattern for MC speakers as for SAE speakers. Speakers in the Tongji Games Corpus were significantly more similar to their partners than to their non-partners in intensity mean, intensity max, f0 max, and speaking rate. That is, for all four features that show evidence of entrainment in SAE, speakers of MC show evidence of entrainment as well. As in the SAE study, we found no evidence of global entrainment on f0 mean. The MC subjects also showed no evidence of entrainment for intensity min or f0 min, which the SAE study did not consider.

Table 1: *T-tests for global similarity in Mandarin Chinese.*

Feature	t	df	p	MC	SAE
Intensity mean	-5.05	98	0.0	✓*	✓
Intensity max	-5.13	98	0.0	✓	✓
Intensity min	-1.16	98	0.25	x	–
F0 mean	0.67	98	0.51	x	x
F0 max	-3.44	98	0.001	✓	✓
F0 min	0.45	98	0.65	x	–
Speaking rate	-7.99	98	0.0	✓	✓

* A checkmark indicates differences are significant for a language.

4.1. Global convergence

While global similarity takes a static view of entrainment, global convergence measures entrainment dynamically, to see whether speakers increase in similarity as the conversation progresses. In our study of entrainment in SAE, we divided each conversation into two parts. If for a given feature the similarity between speaker averages in the second part was greater than their similarity in the first part, we concluded that the speakers displayed convergence on that feature. We further experimented by splitting the first *game* in each session in half (each session in the Games Corpus consists of three games) and then with splitting the entire session in half. We found that intensity mean, shimmer, and NHR were more similar in the second half of the first game than in the first, and that shimmer and NHR, which do not show evidence of entrainment when computed over an entire session, are more similar between partner than between non-partners when computed over the second half alone ($p < 0.0001$). When we compared similarity features across halves of an entire session, we found that pitch mean and jitter were more similar in the second half. We found no evidence of convergence on f0 max or speaking rate in SAE.

In the Tongji Games Corpus, each conversation consists of 18 sections, each of which involves the placement or description of a single card. We compared partner differences over the first nine sections with those in the second nine. We also compared partner differences in the first section with those in the last. The analysis in this section is over 66 conversations; four were omitted because they were missing speech from one of the interlocutors for one or more of the 18 sections. We found that intensity mean and max were significantly more *different* in the second halves of the conversations; no other significant differences were found. We therefore cannot conclude that MC speakers globally converge on any of the features we examined.

5. Local entrainment

Our domain of entrainment in the previous section is global – across the whole conversation. That is, overall, we can say that speakers are similar to each other – more similar than they are to people with whom they are not speaking – but they may not be similar locally: a pair of speakers may fluctuate around similar

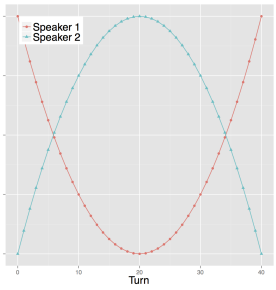


Figure 1: *Speakers entraining locally but not globally.*

means for a feature but diverge widely at any given point. Figure 1 illustrates such a case. The speakers' means are identical, but the distance between them at most points in time is large.

We test for local similarity entrainment by comparing the distance between adjacent IPUs at turn exchanges. If, for a given feature, adjacent IPUs are more similar than non-adjacent IPUs, we conclude that speakers entrain *locally* on that feature, irrespective of their *global* similarity. Consider a conversation comprised of turns $0...n$. For turn i, uttered by speaker A, we calculate the *adjacent* distance as the distance between the initial IPU in turn i, and the final IPU in turn $i-1$, spoken by B. The *non-adjacent* distance is the distance between the turn i's initial IPU and the final IPUs in ten other randomly chosen turns uttered by B. In our study of entrainment in the Columbia Games Corpus, we found evidence of *local* entrainment for every feature we examined using this method. Table 2 shows the results of our comparisons between the adjacent and non-adjacent distances for MC compared with previous results for SAE; t-test values are for MC; checkmarks indicate where the difference is significant for the language. Our results show

Table 2: *T-tests for local similarity in Mandarin Chinese.*

Feature	t	df	p	MC	SAE
Intensity mean	-3.72	69	0.001	✓*	✓
Intensity max	-4.16	69	0.001	✓	✓
Intensity min	0.75	69	0.458	x	–
F0 mean	1.28	69	0.205	x	✓
F0 max	0.81	69	0.419	x	✓
F0 min	-0.17	69	0.986	x	–
Speaking rate	-3.61	69	0.001	✓	✓

* A checkmark indicates differences are significant for a language.

that in MC conversations showing evidence of *global* entrainment, speakers tend to entrain *locally* on intensity mean, intensity max, and speaking rate. This pattern is consistent with our results for *global* entrainment, which was evident in our data for those three features but also for f0 max. However, in contrast to our results for global entrainment, which showed entrainment on the same features as SAE, we did not see evidence of local entrainment on f0 mean or max, while all features displayed local entrainment in SAE.

5.1. Local synchrony

Another way of measuring entrainment is by studying whether speakers' behavior changes in synchrony. Measuring the Pearson's correlation between two sets of values, proposed by [11], captures the dynamics of turn-by-turn synchronous matching between interlocutors to see whether speakers' values vary together even if they are not similar. In our SAE study we found significant correlations between adjacent IPUs for all features ($p \approx 0$). However, correlations for most features were weak ($\gamma < 3$). Intensity mean and max were moderately correlated between adjacent IPUs ($\gamma = 0.50, 0.47$).

To reduce the degree of computation for local synchrony and convergence, we computed the MC correlations over only 30 conversations out of the 70 for which we examined global entrainment, randomly selecting ten conversations each from female-female, male-male, and mixed-gender pairs. We found much stronger correlations between adjacent IPUs in the MC conversations (Table 3). The most noticeable difference between the two languages is that, in SAE, correlations between f0 features in adjacent IPUs are weak, while in MC, they are among the strongest. This may reflect that fact that pitch plays a dual role in a tonal language, conveying both lexical and pragmatic information. Additionally, it is interesting to note the pattern similarities: the correlations for means, both intensity and f0, are slightly stronger than the correlations for maximums, and the correlation for speaking rate is the lowest correlation in both languages, though far more so in MC.

Table 3: *Pearson's correlation between adjacent IPUs. ($p \approx 0$ for all results except MC speaking rate)*

Feature	γ MC	γ SAE
Intensity mean	0.63	0.50
Intensity max	0.55	0.47
Intensity min	0.31	–
F0 mean	0.66	0.28
F0 max	0.61	0.18
F0 min	0.63	–
Speaking rate	-0.048	0.15

5.2. Local convergence

For another view of local entrainment, we look at whether entrainment increases over time: whether speakers *converge* locally. As before, we calculate the difference between adjacent IPUs at turn exchanges (the *adjacent* distance). We then correlate this distance with time. A negative correlation constitutes evidence that the distance between partners at turn exchanges decreases with time. In our SAE study, we found negative correlations between adjacent distance and time for pitch mean and max ($\gamma = -0.06, -0.05; p = 4.6e-11, 4.9e-08$). However, these correlations, although highly significant, are also extremely low.

Using the same 30 MC conversations as in the previous section, we numbered all the turns in each conversation and correlated the adjacent distances for each feature with the turn indices (Table 4). For our MC corpus, the negative correlations between adjacent differences and time are about four times as strong as correlations for the SAE corpus, although still only moderate. As with SAE, we see significant local convergence for pitch mean and max, as well as for intensity min and f0 min, which the SAE study did not consider. Speaking rate, however, shows *divergence*.

6. Entrainment and gender

Several theories of entrainment predict that females will entrain to a greater degree than males. The male dominance hypothesis asserts that differences in speech between males and females can be attributed to women's subordinate social status. Speech

Table 4: *Pearson's correlation between adjacent differences and turn index in Mandarin Chinese.*

Feature	γ	p	MC	SAE
Intensity mean	0.028	0295	x	x
Intensity max	0.022	0.418	x	x
Intensity min	-0.086	0.001	✓*	–
F0 mean	-0.218	0.0	✓	✓
F0 max	-0.238	0.0	✓	✓
F0 min	-0.193	0.0	✓	–
Speaking rate	0.128	0.0	x**	x

* A checkmark indicates differences are significant for a language.
** Displays divergence.

Accommodation Theory ([12]) claims that when a power imbalance exists between interlocutors, the less dominant or powerful speaker will converge more. However, [13] found that these theories failed to explain results of their observations of convergence and divergence in same- and mixed-gender dyads. Alternatively, [3] posits that the perception-behavior link is the mechanism behind entrainment. Thus, women should entrain more than men, regardless of the gender of their conversational partner, because women are known to have greater perceptual sensitivity to vocal characteristics. [14] explained their finding that female speakers were perceived to accommodate more in a shadowing task than male speakers in this way. [15], on the other hand, found that female pairs were *less* similar to each other than male pairs, and concluded that functions outside the domain of perception appear to be influencing the degree of phonetic convergence.

For SAE, we found ([7]) that female-male pairs entrained on every feature examined; in addition, the degree of entrainment on intensity mean and max was greatest for female-male pairs. Male pairs showed the least evidence of entrainment, entraining only on intensity mean, intensity max, and syllables per second, supporting the hypothesis that entrainment is less prevalent among males. Their degree of entrainment on these features was also lower than that displayed by female or mixed-gender pairs. Female pairs entrained on all features except pitch mean, pitch max, and jitter.

The Tongji Games Corpus includes 23 female-female conversations, 17 male-male, and 30 mixed-gender. As in Section 4, and as in [7], we compared *partner* differences – the differences in feature values between interlocutors – with *non-partner* differences – differences in feature values between each speaker and the averaged values with all speakers of her partner's gender and role with whom she is never partnered. Our results are shown in Table 5. Again, the similarity in pattern to

Table 5: *Evidence of global entrainment by gender group.*

Feature	FF		MM		MF	
	MC	SAE	MC	SAE	MC	SAE
Intensity mean	✓	✓	x	✓	✓	✓
Intensity max	✓	✓	x	✓	✓	✓
Intensity min	x	–	x	–	x	–
F0 mean	x	x	x	x	✓	✓
F0 max	x	x	x	x	✓	✓
F0 min	x	–	x	–	x	–
Speaking rate	✓	✓	✓	✓	✓	✓

our results for SAE is striking. We find that mixed-gender pairs entrain on the greatest number of features, and male pairs on the least. As for SAE, the most consistent results are for intensity mean, intensity max, and speaking rate, although all gender groups entrained on these in SAE, and male pairs entrain only on speaking rate in MC.

In addition to the number of features entrained on, we are also interested in the degree of entrainment exhibited by each gender group. We compare each group's *partner* similarities, normalized by the *non-partner* similarities to control for the overall within-group similarity. We compare the strength of entrainment on intensity mean, intensity max, and speaking rate, the three features that show the most evidence of entrainment among all three gender groups. For SAE, we found that entrainment on intensity mean and max was strongest for mixed-gender pairs and weakest for male pairs; the strength of entrainment on speaking rate followed this pattern but the differences only approached significance ($p = 0.08$). For MC, the differences in entrainment strength were significant between all three groups for all three features (Intensity mean: $F = 3.13, p = 0.048$; intensity max: $F = 3.73, p = 0.028$; speaking rate: $F = 5.10, p = 0.008$). A post-hoc test revealed that entrainment on intensity mean and max was weakest for male pairs, while entrainment on speaking rate was weakest for mixed-gender pairs. While for SAE, we concluded that entrainment is both strongest and most prevalent in mixed-gender pairs, for MC we can only conclude that it is most prevalent in mixed-gender pairs, but not necessarily strongest.

7. Discussion and Future Research

The truly striking finding presented in this paper that entrainment in pitch, intensity and speaking rate appears to be generally very similar in SAE and in MC. We have presented evidence that MC speakers entrain globally in similarity of values for the three main aspects of prosody: duration, pitch and intensity. However, unlike SAE speakers, they show no evidence of global convergence on any feature. Locally, they entrain in similarity of values on intensity and speaking rate and in synchrony on intensity and pitch. They converge locally on intensity min and all f0 features, and diverge on pitch. The prominence of intensity among these results – it is the only feature for which there is evidence of entrainment for all three local measures – is something we observed in SAE as well.

When we examine entrainment behavior among different gender groups, we find that, as for SAE, entrainment is most prevalent in mixed-gender pairs and least prevalent among male pairs. Also as for SAE, all gender groups entrain most consistently on intensity and speaking rate. However, we did not find that entrainment was strongest among MC mixed-gender pairs, as we did for SAE.

The similarity of our findings for Columbia and the Tongji Games Corpora supports not only the view that entrainment is a cross-cultural phenomenon but provides evidence that members of different language groups entrain in similar ways. In future work, we will focus on individual differences in entrainment behavior, analyzing the patterns of which features speakers entrain and converge on, both globally and locally. We will also examine how conversational role affects entrainment behavior.

Acknowledgements

This work was supported in part by NSF IIS-0803148 and by the Humanities and Social Sciences Foundation of China Education Ministry (11YJA740113). The authors thank Štefan Beňuš, Agustìn Gravano, Daniel Hirst, Qiuwu Ma, and Fengying Mu for useful comments, and Yuanyuan Zhang, Jin Peng, Chenxiao Zou, Yalu Chen, Li Jiao, who labeled the Tongji Games Corpus.

8. References

[1] M. J. Pickering and S. Garrod, "Towards a mechanistic psychology of dialogue," *Behavioral and Brain Sciences*, vol. 27, pp. 169–226, 2004.

[2] D. Goleman, *Social Intelligence: The New Science of Human Relationships*. Bantam, 2006.

[3] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[4] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[5] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[6] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011.

[7] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 11–19. [Online]. Available: http://www.aclweb.org/anthology/N12-1002

[8] A. Gravano, "Turn-taking and affirmative cue words in task-oriented dialogue," Ph.D. dissertation, Columbia University, 2009.

[9] B. Bigi and D. Hirst, "Speech phonetization alignment and syllabification (sppas): a tool for the automatic analysis of speech prosody," in *Speech Prosody*. Tongji University Press, 2012, pp. 19–22.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

[11] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proceedings of Interspeech*, 2009.

[12] H. Giles, A. Mulac, J. Bradac, and P. Johnson, *Speech accommodation theory: the first decade and beyond*. Beverly Hills, CA: Sage, 1987.

[13] F. R. Bilous and R. M. Krauss, "Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads," *Language & Communication*, vol. 8, no. 3/4, pp. 183–194, 1988.

[14] L. L. Namy, L. C. Nygaard, and D. Sauerteig, "Gender differences in vocal accommodation: the role pf perception," *Journal of Personality and Social Psychology*, vol. 21, no. 4, pp. 422–432, 2002.

[15] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustic Society of America*, vol. 19, no. 4, 2006.

Speaker Movement Correlates with Prosodic Indicators of Engagement

Rob Voigt, Robert J. Podesva, Dan Jurafsky

Linguistics Department, Stanford University, Stanford, CA

robvoigt@stanford.edu, podesva@stanford.edu, jurafsky@stanford.edu

Abstract

Recent research on multimodal prosody has begun to identify associations between discrete body movements and categorical acoustic prosodic events such as pitch accents and boundaries. We propose to generalize this work to understand more about continuous prosodic phenomena distributed over a phrase - like those indicative of speaker engagement - and how they covary with bodily movements. We introduce *movement amplitude*, a new vision-based metric for estimating continuous body movements over time from video by quantifying frame-to-frame visual changes. Application of this automatic metric to a collection of video monologues demonstrates that speakers move more during phrases in which their pitch and intensity are higher and more variable. These findings offer further evidence for the relationship between acoustic and visual prosody, and suggest a previously unreported quantitative connection between raw bodily movement and speaker engagement.

Index Terms: acoustic prosody, visual prosody, movement, gesture, speech-gesture interface, automatic methods

1. Introduction

Gesture and movement are fundamental and ubiquitous parts of the human communicative system, but are traditionally understudied phenomena in linguistics. In recent years, interest in the study of multi-modal communication and the connection between speech prosody and "visual prosody" has increased, and empirical evidence has begun to convincingly demonstrate the co-articulatory nature of "gestures and language [as] one system" (McNeill 1992).

For instance, Jannedy and Mendoza-Denton (2006) investigate gesture's role in structuring spoken discourse, showing evidence for the co-occurrence of pitch accents and gestural apices. Krahmer and Swerts (2007) show that even independent of pitch accents, the production of "visual beats" has an effect on the prosodic realization and prominence of the co-occurring speech. Gibbon (2011) demonstrates rhythmic matching between the acoustic and physical beat rhythms in drum-accompanied storytelling in the Ega language. Loehr (2012) confirms findings of temporal synchrony, showing that gestural phrases align with intermediate phrases.

Beyond overt gestures, substantial evidence supports the connection between other kinds of "visual prosody" and speech. Guaïtella et al. (2009) track rapid eyebrow movements in dialogue, and demonstrate their connection with turn-taking in interaction. Cvejic et al. (2010) use facial optical markers in motion capture recordings; in their data, speakers exhibit greater movement during prosodically focused words, even for non-articulatory features such as eyebrow and head movement. Walker (2012) explores "trail-off" conjunctions, showing that speakers and listeners in interaction use visible features such as dropped gaze to signal "disengagement."

Studies in speech perception further demonstrate the important communicative functions of gesture and movement. Munhall et al. (2004) record and recreate 3D models of head movement in talk, finding that subjects are able to correctly identify more syllables when the speech is accompanied by 3D models of natural head movements as compared with distorted or absent models. Scarborough et al. (2009) obtain forced-choice judgments of lexical and phrasal stress from subjects shown video data with the audio track removed; they show that phrasal stress is more easily perceived by subjects than lexical stress. Rilliard et al. (2009) give results for French and Japanese suggesting visual information helps listeners disambiguate "social affects" that are less clear in the audio signal alone.

In spite of this progress, major methodological impediments remain. A principled analysis of gesture in experimental settings requires complex and time-consuming human annotation, commonly based on one of several existing annotation schemes. These include interval annotation of "gesture units" and "gesture phrases" based on written transcripts (Kendon 2004); expressivity annotations on parameters such as "fluidity," "spatial expansion," and "repetitivity" (Chafai et al. 2006); and keyframe-based manual posing of animated 3D characters (Kipp et al. 2007). Though their descriptive power is high, these schemes share the property that they require huge amounts of time and effort from highly-trained human annotators. As a result, existing linguistic studies of gestural prosody necessarily operate on extremely small data sets: for example, Jannedy and Mendoza-Denton (2006) perform their analyses on 130 seconds of videorecorded speech data; Loehr (2012) on speech events from four separate speakers totalling 164 seconds of data.

Analyses of movement pose their own unique difficulties: mainly, raw movement is difficult or impossible to annotate by hand; tools for facial or motion tracking must be used, and this equipment is likely to be expensive or invasive. We therefore know little about the theoretically important relation between affective measures of speaker engagement and the embodied expression that takes form in movement.

This work is aimed at addressing these problems. We introduce a new method for automatically measuring movement magnitude and variance from video data, and apply it to a corpus of single-speaker YouTube videos, extracting acoustic and movement measurements for each phrase in the data. We then investigate the relationship between our proposed movement measure and several prosodic features indicative of engagement including pitch, pitch variance, intensity, and intensity variance (Liscombe, Venditti, and Hirschberg 2003; Mairesse et al. 2007; Gravano et al. 2011; McFarland et al. 2013).

We hypothesize that increased movement amplitude will be predictive of higher values in these acoustic categories. That is, during phrases in which speakers are engaged, excited, and moving more, they will use a correspondingly higher pitch and intensity as well as greater variance in their pitch and intensity.

2. Methods

In this work, we propose a pipeline of fully automatic, replicable annotation for the analysis of visual and acoustic prosody on single-speaker videorecorded data.

2.1. Data

Web-based video streaming services are an increasingly culturally significant tool for communication. YouTube alone reports more than 100 hours of video uploaded per minute[1], of which a significant portion is certainly linguistic in nature - conversations, vlogs, lectures, and so on. Existing work has begun to use YouTube to investigate multi-modal sentiment (Wöllmer et al. 2013) and sociolinguistic aspects of identity construction (Chun 2013), but this data source remains vastly underutilized.

As we describe in Section 2.4, our new proposed "movement amplitude" measure operates on raw video data, without the need for additional equipment at recording time. This makes it useful for the analysis of movement in YouTube data, which has the significant benefit of replicability: since users who post videos explicitly open them to the public, researchers can apply new methods and test new hypotheses on existing datasets without confronting issues of subject confidentiality.

In this study, we focus on the connection between acoustic and visual prosody in a single, narrowly-defined genre of YouTube videos: the "First Day of School" vlog. In such videos students speak into their cameras to describe their experiences in their first day starting or returning to school. This is a productive genre, with a search for "first day of school vlog" on YouTube returning nearly 1.3 million results.

We collect 14 such videos from different speakers, resulting in a total of 95 minutes of footage. The speakers all fit the most commonly represented demographic in such videos: female high-school-aged students from the United States. We use *pafy*[2] to download 360p-quality mp4-encoded versions of each video. Each video consists of a single speaker seated against a static background. We cut each video to a section of continuous speech, removing introductory and closing title cards.

A first application of our methodology to this dataset is interesting and appropriate for several reasons. The videos were found "in the wild," naturally uploaded by the speaker outside of an experimental context. They are linguistically and gesturally interesting; the speakers are in general very animated and performative, communicating directly to their peer group in discussing social expectations, classes, relationships, and so on.

2.2. Automatic Identification of PBUs

We need to extract prosodic units for our analysis; we use pause-bounded units (PBUs), automatically identified with a simple heuristic algorithm using the silence detection function in *Praat* (Boersma and Weenink 2013).[3]

We write a *Praat* script that runs an intensity analysis on the audio track of a given videorecording, then identifies silent and sounding intervals with a minimum duration of one-tenth of a second. We begin with a silence threshold of -30.0 dB and calculate the average length of the phrases thus identified. If the phrases have an average length of greater than two seconds, we increase our silence threshold by +3.0 dB and re-extract, continuing to do so until they average below two seconds in length. Our two-second length threshold is derived from an average length calculation on hand-annotated PBUs from a separate set of interactional data.

2.3. Acoustic Features

Following prior work on prosodic engagement, we extract fundamental frequency (henceforth "pitch") and intensity features for each automatically-identified pause-bounded unit in our dataset. We use a *Praat* script to calculate eight acoustic variables for each PBU: a maximum, mean, minimum, and standard deviation for both pitch and intensity.

2.4. Movement Amplitude

Here we propose a new measure for the analysis of movement in videorecorded data. The intuition behind this measure is simple. Video data consists of a series of frames, which are fundamentally images, played back at a high speed to simulate movement. In circumstances where the video camera is stationary and the background of the recorded images is relatively static, speaker movement can be quantified by measuring the difference between successive frames.

In practice, we propose a measure obtained by finding the average difference in RGB values between a given pixel and the corresponding pixel in the preceding frame, summing these values across all pixels in the image, and taking the natural log of the total. We extract video frames as uncompressed png images using the ffmpeg software package[4] and compute frame differences using the ImageChops python module from the Python Imaging Library.[5] Formally, we define the movement amplitude (MA) measured at time t in (1), with the current frame number n, an image size of width x and height y, and a function $pix_{x,y}(i)$ that returns the red, green, and blue values of a given pixel at an arbitrary frame number i:

$$\mathrm{MA}(t) = \ln \sum_{x,y} avg(|pix_{x,y}(n) - pix_{x,y}(n-1)|) \qquad (1)$$

Graphical observation of density and quartile plots of our data confirm the intuition that movement amplitude must be computed in log space. Large movements and gestures are relatively sparse compared with the continuous, generalized movements of speech, and the large variance in the number of pixels they comprise necessitates log space calculation.

Such a measure has many attractive properties. First, it is fully automatic, allowing replicable quantification of visual prosody without painstaking hand-annotation. This means it can be scaled up to provide annotation for datasets of arbitrary size with little additional effort or expense.

Secondly, like measures of acoustic prosody, it is functionally continuous, albeit at a much coarser granularity than most audio recordings. Standard audio sampling rates are 44,100 and 48,000 Hz, while standard video frame rates include 24 and 30 frames per second (FPS). We can calculate one measurement per frame, so these frame rates would allow us to extract movement amplitude samples at 24 and 30 Hz, respectively. In the 30 FPS case this provides one measurement each 33ms. In this study we show that samples extracted at this frequency are sufficient and offer meaningful data on visual prosody; however,

[1] http://www.youtube.com/yt/press/statistics.html

[2] https://pypi.python.org/pypi/pafy

[3] While PBUs provide a meaningful and computationally tractable approximate prosodic unit for this analysis, it remains an interesting task for future work to determine how they might correlate with or be related to a more theoretically principled unit such as the intonational phrase (Pierrehumbert 1980).

[4] http://ffmpeg.org

[5] http://www.pythonware.com/products/pil/

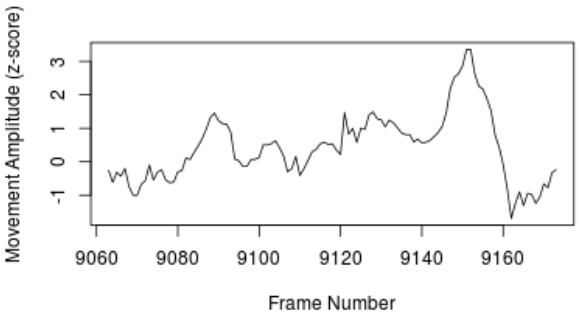


Figure 1: *Visualization of four seconds (120 frames at 30 FPS) of movement.*

the sampling rate of the movement amplitude measure could be arbitrarily scaled in proportion to the quality of the camera available to the experimenter.

Figure 1 demonstrates this continuity, showing movement amplitudes extracted from a four-second clip in our data. The peak near frame 9090, representing an MA measurement one standard deviation above the mean, encodes the combination of a speaker opening her eyes, turning her head towards the camera, and opening her mouth to say "Oh!" in recognition after a moment of thought. The significantly higher peak near frame 9155, three standard deviations above the mean, is primarily the result of a large one-handed swiping gesture.

This plot makes clear another important property of the movement amplitude measure in its current form: it encodes any and all movement, including that of the eyes, mouth, head, body, and so on. Standing up from a seated position, for example, would be recorded as a dramatic peak in MA. It also necessarily compresses 3D movement to a 2D representation in the camera's visual plane. In these ways it is substantially more coarse than any of the measures used in prior studies; however, MA quantifies overall movement in a reasonable way, as the movement of larger objects is given more weight than that of small ones. With computer vision tools such as accurate face detection, in future work this measure could also be applied to submovements to separate out, for example, facial movement as compared to body movement.

Our measure is limited by several required conditions that a recording must meet. The background of the video must be static: the functional result of this is that the contribution to the MA measurement for all pixels that show only background is negligible or absent. Additionally, all speakers in a video must be visually separable. The present study is concerned only with single-speaker video data, but in continuing work we have applied this measure to multi-speaker interactions by defining a rectangular pixel bounding region for each speaker and calculating MA for each speaker only within their region.

These conditions are limiting insofar as other techniques such as hand annotation of gesture could be applied to video footage with variable camera angles and non-static backgrounds. As discussed in Section 2.1, however, data meeting these conditions is reasonable to collect experimentally.

In processing our data we also convert MA measurements to z-scores per speaker to allow for comparable measurements in spite of differences across recording conditions such as speaker distance from the camera, color of the background and speaker clothing, and so on. As with our acoustic variables, for each video we extract an MA maximum, mean, minimum, and standard deviation for each PBU.

Figure 2: *Five-frame composite visualization of the speaker's head and facial movements as captured by movement amplitude across the first peak shown in Figure 1.*

2.5. Statistical Analysis

Automatic extraction of PBUs from our 14 videos results in 2172 observed pause-bounded units, and for each we extract prosodic features for speech and movement as described above.

To model the behavior of this data we use linear mixed-effects models as implemented in the lme4 package in R (Bates et al. 2013). We model speakers as random effects in a series of regressions predicting acoustic variables from our new movement amplitude measure, including log PBU duration in the model as a control variable. The four MA measurements (max, mean, min, and std) are highly collinear, so we use principal component analysis (PCA) for dimensionality reduction. Statistical significance is based on p-values calculated using the lmerTest package in R (Kuznetsova et al. 2013) with degrees of freedom estimated using Satterthwate's approximations.

3. Results

Dimensionality reduction with PCA on our MA measurements reveals two orthogonal components that together explain 96% of the variance in the MA data. The loadings for these two factors are seen in Table 1. Factor 1 is interpretable as overall movement, and factor 2 as variance in movement.

	Overall Movement	Movement Variance
MA Measure	*Factor 1*	*Factor 2*
max	42.1	76.4
min	73.9	-54.3
mean	52.2	19.1
std	-6.9	29.0
variance explained	64.8%	31.2%

Table 1: *Loadings of orthogonal components for movement amplitude calculated from principal components analysis.*

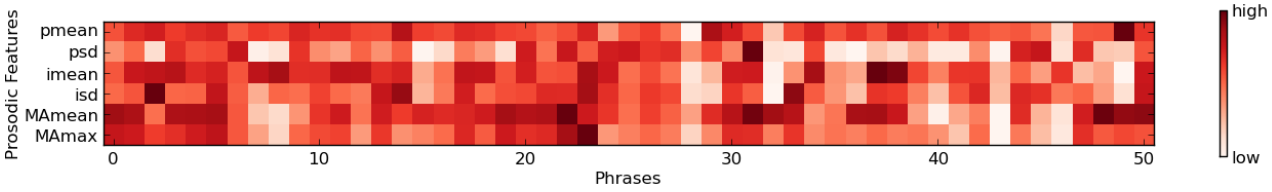


Figure 3: *Visualization of prosodic features including movement across 51 consecutive phrases from one speaker. Each vertical line of boxes represents one spoken phrase, where more deeply-shaded boxes represent higher feature values. Notice light and dark vertical banding; phrases with higher movement amplitude also have higher values for measures of acoustic prosody (and vice versa).*

The results in Table 2 show that high movement amplitude (Factor 1) has a statistically significant positive relationship with all of our acoustic variables except for intensity minimum and pitch minimum. That is, during phrases in which speakers are moving more, we can predict an increase in pitch maximum, mean, and standard deviation as well as intensity maximum, mean, and standard deviation. High movement variance (Factor 2) was not predictive of any of the acoustic features we measured. Though we ran a series of models, the results are highly significant and would survive a Bonferroni correction or any related control for multiple comparisons.

	Overall Movement	Movement Variance
Pitch	*Effect Size (Hz)*	
pmax	4.889 ***	—
pmin	—	—
pmean	2.797 ***	—
psd	0.875 **	—
Intensity	*Effect Size (dB)*	
imax	0.280 ***	—
imin	—	—
imean	0.152 **	—
isd	0.082 ***	—

Table 2: *Effect sizes for acoustic features predicted at a statistically significant level by movement amplitude in a series of mixed-effects regressions.*
$*$ *indicates* $p < 0.05$, $**$ *is* $p < 0.01$, *and* $***$ *is* $p < 0.001$.
— *indicates no significant relationship.*

To confirm these results, we also run mixed-effects regressions using speaker-scaled pitch max and min, where pitch measurements are converted to a 0-1 scale based on a speaker's overall max and min. We also calculate pitch range features per PBU, which are similarly a value between 0 and 1 calculated by subtracting the scaled pitch max from min. These models show a similar statistically significant trend: an increase of one standard deviation in our overall movement factor predicts use of 1% more of a speaker's pitch range within a given phrase.

4. Discussion

Our results build upon prior work to provide further empirical evidence for a strong connection between speech prosody and the "visual prosody" of movement and gesture. In our dataset, phrases with more overall movement were likely to have higher and more variable pitch as well as louder and more variable intensity, confirming our initial hypotheses. This finding is novel in that it adds a dimension of quantity: whereas the previous literature has found primarily temporal synchrony (i.e., the timing of pitch accents aligns with gestural apices), our results demonstrate that more movement is indicative of increased excitement in these prosodic categories.

On the other hand, movement variance - the extent to which a particular phrase had both large movements and periods of little to no movement - was not predictive of any of our acoustic features. This finding is particularly interesting in that a high movement variance would encode some well-defined fully semantic gestures. Consider, for example, a definitive pointing gesture with a pause at its apex. According to the movement amplitude measure, such a gesture would have a high MA value during the stroke and a very low value at the apex, resulting in high MA variation in the PBU during which it occurred. While the movement amplitude measure is not sufficiently fine-grained to make definitive claims in this regard, our findings are suggestive that the synchrony of gestural apices and pitch accents is predominantly temporal and local: that is, a speaker's most "extreme" gestural apices may accurately predict the timing of the immediately adjacent pitch accent, but not necessarily its magnitude with respect to that speaker's global pitch range. This hypothesis remains to be tested in detail in future work.

Additionally, this study makes several valuable methodological contributions to the multi-modal analysis of speech prosody. The newly proposed movement amplitude measure provides a replicable estimation of overall movement from raw video footage, without the need for expensive or invasive equipment at the time of filming or time-consuming annotation effort after the fact. These properties make this measure particularly attractive for the analysis of videos collected "in the wild," such as from internet video sharing sites like YouTube.

The fact that this measure, when combined with automatic extraction of approximate pause-bounded phrases as presented in this paper, is completely automatic for single speakers makes it tractable for empirically-driven sociolinguistic analyses of video data in a way that is simply infeasible by means of hand annotation alone. This paper presented statistically significant results on a narrowly-defined dataset of "First Day of School" vlog posts in order to most directly control for prosodic differences across sociolinguistic groups, but we aim to continue and expand this research. Future work will consider the possible influence of genre effects, social meaning and contextual factors such as performativity, differences in interactional or conversational speech as compared with monologues, and the influence of sociolinguistic variables such as age, gender, and dialect.

5. Acknowledgements

Thanks to the NSF (award IIS-1216875). Thanks also to Jeremy Calder, Patrick Callier, Annette D'Onofrio, Katherine Hilton, Teresa Pratt, and Janneke Van Hofwegen for their helpful advice and thought-provoking discussion.

6. References

[1] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1-0.

[2] Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. Version 5.3. 39.

[3] Chafai, N. E., Pelachaud, C., Pel, D., & Breton, G. (2006). Gesture expressivity modulations in an ECA application. Intelligent Virtual Agents, 181-192.

[4] Chun, E. W. (2013). Ironic Blackness as Masculine Cool: Asian American Language and Authenticity on YouTube. Applied Linguistics, 34(5), 592-612.

[5] Cvejic, E., Kim, J., Davis, C., & Gibert, G. (2010). Prosody for the eyes: quantifying visual prosody using guided principal component analysis. Proceedings of INTERSPEECH 2010, 1433-1436.

[6] Gibbon, D. (2011). Modelling gesture as speech: a linguistic approach. Poznan Studies in Contemporary Linguistics, 47, 470.

[7] Gravano, A., Levitan, R., Willson, L., Benus, S., Hirschberg, J., & Nenkova, A. (2011). Acoustic and Prosodic Correlates of Social Behavior. Proceedings of INTERSPEECH 2011, 97-100.

[8] Guaïtella, I., Santi, S., Lagrue, B., & Cav, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. Language and Speech, 52(2-3), 207-222.

[9] Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. Interdisciplinary Studies on Information Structure, 3, 199-244.

[10] Kipp, M., Neff, M., & Albrecht, I. (2007). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. Language Resources and Evaluation, 41(3-4), 325-339.

[11] Krahmer, E., & Swerts, M. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. Journal of Memory and Language 57.3 (2007): 396-414.

[12] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. (2013). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).

[13] Liscombe, J., Venditti, J., & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. Proceedings of Eurospeech 2003.

[14] Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology, 3(1), 71-89.

[15] Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artifial Intelligence Research, 30, 457-500.

[16] McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013). Making the Connection: Social Bonding in Courtship Situations. American Journal of Sociology, 118(6), 1596-1649.

[17] McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago Press.

[18] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychological Science, 15(2), 133-137.

[19] Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. Doctoral dissertation, Massachusetts Institute of Technology.

[20] Rilliard, A., Shochi, T., Martin, J. C., Erickson, D., & Auberg, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. Language and Speech, 52(2-3), 223-243.

[21] Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. Language and Speech, 52(2-3), 135-175.

[22] Walker, G. (2012). Coordination and interpretation of vocal and visible resources:Trail-offconjunctions. Language and Speech, 55(1), 141-163.

[23] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L. (2013). YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context. Intelligent Systems, IEEE, 28(3), 46-53.

Prosody, voice assimilation, and conversational fillers

Štefan Beňuš[1,2] *Marián Trnka*[2]

[1] Department of English and American Studies, Constantine the Philosopher University, Nitra, Slovakia

[2] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

sbenus@ukf.sk, trnka@savba.sk

Abstract

Conversational fillers (CFs), commonly transcribed as *uh*, *um*, or *er*, typically start with a schwa-like vowel, and signal multiple social, interactive, meta-cognitive, and pragmatic functions. They also co-occur with prosodic boundaries, increase saliency of inter-word disjunctures, and participate thus in coding the prosodic structure. Contrary to these functions, CFs are assumed not to participate in the phonological system of a language. This paper uses two types of Slovak conversational speech corpora for investigating the prosodic and phonological behavior of CFs. In Slovak, the vowel inventory does not include a schwa, and word-final obstruents undergo voice assimilation that is triggered by word-initial vowels but interacts with the strength of the prosodic boundary between the two words. Our data show the propensity of CFs to neutralize word-final voicing, and function thus as prosodic breaks, but also non-negligible number of cases of CFs triggering voicing of word-final obstruents, supporting their relevance for cognitive phonology.

Index Terms: conversational fillers, Slovak, non-verbal vocalizations

1. Introduction

1.1. Conversational fillers

Conversational fillers (CFs) are vocalizations such as *uh*, *um*, or *err* that are ubiquitous in everyday speaking, and if considered words, their rates reach 2-5% of all words (e.g. [1], [2]). They convey numerous para-linguistic functions ranging from social approval [3], turn-taking management [4] and meta-cognitive processing [5], to co-creating pragmatic and discourse structures in interactions [6]. CFs are thus necessary for successful, smooth, and socially natural spontaneous conversations ([7], [8]). For illustration, CFs participate in establishing conceptual alignment between interlocutors by signaling given-new distinctions, problems with parsing or understanding preceding speech, or focusing attention on the upcoming material [9]. In this paper we focus on turn-internal CFs that most commonly signal turn-holding hesitation associated with cognitive planning and packaging information in upcoming speech material.

In addition to the frequencies and distributions of CFs, their prosody also provides useful information. For example, in turn-medial position, F0 of CFs was found to systematically correlate with the intonation of the preceding material [10]. CFs are commonly delimited by silent pauses and form thus a separate intonational phrase [2]. However, their role in prosodic chunking when integrally linked to the surrounding material is less clear. On the one hand they are perceived as increasing the salience of the disjuncture between the word that precedes and follows a CF. On the other hand, the pitch re-set, one of the typical signals of prosodic chunking between adjacent units, is very often absent between the units flanking a CF. Hence, it is not clear what role, if any, these prosodically integrated CFs play in prosodic structuring.

1.2. Slovak

Slovak is a West-Slavic language with a vowel inventory containing 5 basic vowel qualities [i, e, a, o, u], phonemic vowel duration, left-most words stress, and minimal reduction of vowel quality in unstressed syllables [11]. Fig. 1 illustrates prosodically conditioned (quantity & stress) Slovak vocalic qualities.

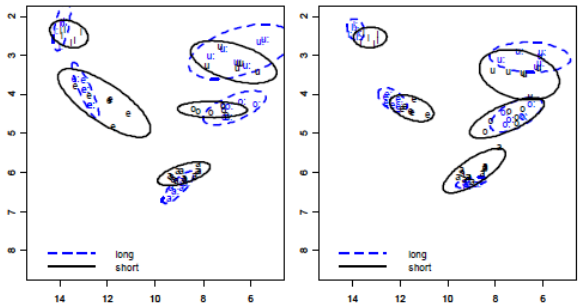


Figure 1: *Slovak vowel quality based on F1 (y-axis) and F2 (x-axis) in Bark for short (solid) and long (dashed) stressed (left) and unstressed (right) monophthongs in CVCa nonsense words from one speaker; adapted from [11].*

Slovak consonantal inventory includes 8 plosives, 4 affricates, and 8 fricatives paired for the underlying phonemic voiced-voiceless contrast. These obstruents are targets for voicing assimilation in the following way. Word-final *voiced* obstruents are *devoiced* if the following word-initial consonant is voiced or if there is a major prosodic break between the words, commonly including a silent pause. Word-final *voiceless* obstruents are *voiced* if the following word starts with a voiced segment, including a vowel, and no silent pause occurs between the two words. While consonant-triggered voicing neutralization is common, vowel-induced voicing of obstruents is rather rare, reported for example for Cracow Polish [12], a variety of another West-Slavic language. Phonetically, voicing neutralization has been found incomplete for languages such as German [13], Polish [14], or Catalan [15], but limited available experimental evidence for Slovak supports complete consonant-induced voicing neutralization across words separated by weak prosodic breaks [16].

1.3. Research questions

The majority of Slovak CFs begin with a schwa-like vowel [17]. This, together with the nature of Slovak voicing assimilation targeting word-final obstruents, offers a unique

testing ground for the role of CFs in the sound system of a language. If CFs trigger voicing of word-final voiceless obstruents, or block de-voicing of word-final voiced ones, CFs might be considered as regular words participating in the phonology of Slovak. If CFs trigger de-voicing of voiced obstruents, or block voicing of the voiceless ones, this would support their analysis as boundary signals, similarly to pauses, participating in the establishment of prosodic structuring. In this sense, CFs might be an optional modal voice boundary marker similar to pre-boundary lengthening or pitch-reset. Alternatively, CFs might be invisible and filtered out for voicing assimilations or prosodic boundary marking. Finally, we want to test if the phonetic realization of fillers, such as their initial glottalization or their duration, affects their behavior in voicing assimilation. A positive answer would be consistent with their analysis as a prosodic boundary marker.

The above questions concern the nature of CFs and their role in the Slovak sound system. A possibility of CF-triggered voicing assimilation might also shed light on the nature of the assimilation process itself. Some analyses treat word-final voicing of obstruents as a phonological phenomenon accounted for in a discrete fashion (i.e. alternations with +/- voice irrespective of the formal framework applied, e.g. [18], [19]). Other approaches employ dynamic modeling to account for both discrete and continuous nature of word-final voicing, especially in light of incomplete neutralization discussed above, e.g. [20]. Finally, some assume that voicing of word-final pre-sonorant obstruents is a purely phonetic phenomenon, e.g. [12]. Hence, if CFs trigger (or do not block) voicing of word-final obstruents preceding them, then CFs must be specified for [+voice], which then spreads to, or licences the voicing of, the preceding obstruent. Alternatively, CFs might be treated as non-linguistic entities incapable of receiving a [+voice] specification and their participation in voicing alternations would be taken as support for the phonetic nature of pre-sonorant word-final obstruent voicing.

2. Methodology

2.1. Corpora

Two corpora of conversational Slovak are analyzed in this paper. The first is a corpus of task-oriented dyadic collaborative conversations accompanying the interactive game designed to elicit dialogues and adapted from the OBJECT Games of Columbia Games Corpus [21], [17]. A pair of subjects saw images on their computer screens and without seeing each other, had to agree on the location of the target object with respect to other objects. One of the subjects (Placer) then dragged this image with the mouse on the location matching as closely as possible the location of this image on the other subject's (Describer) screen. Each session consisted of 14 such placement tasks in which the roles of the Placer and Describer were equally divided between the two players. We will refer to this corpus as *SK-Games*.

In this paper we analyze a subset of *SK-Games* including 6 sessions and totaling almost 4 hours of speech (3h, 54m) from 7 subjects (3 females 4 males; 5 subjects played the game twice with a different partner and 2 male subjects played only one game). There are 21773 words in total, out of which 763 are conversational fillers (labeled as *uh* or *mm*) [17].

The second corpus contains recordings of a courthouse TV show in which semi-professional actors play attorneys and plaintiffs and plead their cases in front of a judge. To match approximately the number of CFs of interest in first corpus (156, see 3.1), we randomly selected 188 conversational fillers following words ending in obstruents. We disregarded the cases in which the disjuncture between the word and the filler corresponded to a #4 ToBI break, that is a silent pause in most cases, focusing thus on the core cases meeting the environment of voicing assimilation. Finally, the extensive size of this corpus allowed for rough balancing of underlying voicing of obstruents. This was not done in SK-Games corpus and resulted in a strong bias toward underlyingly voiceless obstruents (110 vs. 46; see Table 2), which is partially lessened in this corpus (107 vs. 81; see Table 3). We will refer to this second corpus as *Court-TV*.

2.2. Data processing and labeling

Speech was manually transcribed including the transcription of conversational fillers. Transcripts with the audio signal were used for automatic forced alignment using the SPHINX toolkit adjusted for Slovak [22]. Three dimensions, described in Table 1, represented the core labeling effort: characterizing the articulatory activity of the vocal cords initiating the filler, voicing of the obstruent ending the word preceding the filler, and the degree of disjuncture between the filler and the preceding word within the ToBI framework [23].

Table 1. *Labeling scheme for voice assimilation types, "UR" stands for underlying representation*

Label	Dimension	Function description
M	Vocal cords	Modal voice; smooth amplitude increase
G		Glottalization
B		Burst; glottal stop with a observable burst
D	Voice assimilation	Devoicing: UR [+v] obstruent →[-v]
N		No application: UR [+v] obstruent remains [+v]
V		Voicing: UR [-v] obstruent →[+v]
K		Blocking of voicing: UR [-v] obstruent remains [-v]
1	Prosodic disjuncture	No perceivable break
2		Perceived lengthening of word-final rhyme
3		Minor break, commonly associated with tonal marking and/or small pause
4		Major break with significant silent pause between the word and the filler

Additionally, to assess the hypothesis that CFs might be invisible for the purposes of voice assimilation, for all cases with possible voicing (disjunctures 1-3) in *Court-TV*, we noted if the voicing of word-final obstruents respects the generalization described in Section 1.2 should the intervening filler be disregarded and treated as transparent. In other words, we checked the voicing agreement between the final obstruent before the CF and the initial sound of the word following the CF and labeled as respecting (1) if agreement was present or if word-final obstruent was voiceless and a silent pause followed the CF, and as not respecting (0) in the complementary cases.

Finally, the only continuous feature analyzed in this paper is CF duration that is assumed to correlate positively with the boundary strength: longer CF signal stronger prosodic boundaries. Given the uncontrolled nature of the speech in the corpora and questionable reliability of word alignment to the

signal, we employ raw duration and leave normalization effort for subsequent research.

3. Results

3.1. SK-Games corpus

The distribution of voice assimilation types in the corpus of 763 fillers is shown in Table 2. The first four rows correspond to the assimilation types based on the second dimension of Table 1, and the last row is a 'catch all' in which CFs were either preceded by sonorants or a major silent pause inhibiting the application of voicing assimilation. The second and third columns represent the underlying representation (UR) and surface form (SF) of the voicing feature respectively, and the following four columns represent ToBI's boundary indices.

Table 2. *Distribution of voice assimilation types in SK-Game corpus*

Type	UR	SF	1	2	3	4	Total
D	V+	V-	21	1	1	0	23
N	V+	V+	22	1	0	0	23
V	V-	V+	9	0	0	0	9
K	V-	V-	81	15	5	0	101
N-rest			59	40	28	480	607
Total			192	57	34	480	763

If we take the 4 most relevant categories, there are 156 cases in which a word-final obstruent is followed by a filler with a disjuncture lower than ToBI's #4 break. These cases represent the core data for evaluating the questions set in Section 1.3. There are several observations that can be made from the table. First, all possible situations can be observed (with frequency counts of more than 5), and thus any process responsible for voicing alternations potentially triggered by conversation fillers is optional and/or variable. Second, the distribution of underlyingly voiced and voiceless obstruents in this corpus (46 vs. 110) significantly deviates from the expected equal split given the same number of voiced and voiceless obstruents in Slovak ($p < 0.001$ with exact binomial calculation). The same applies to surface forms that are more likely to be voiceless than voiced (124 vs. 32). Third, the 2x2 contingency table for voiced/voiceless obstruents underlyingly and on the surface gives a significant deviation from the expected values, X^2 [1] = 32.9, $p < 0.001$. The analysis of residuals shows that D and V types are significantly under-represented while N is significantly over-represented. Seen in this way, the absence of CF-triggered voicing (K) is to be expected and retention of UR voicing occurs significantly more often than chance. However, changes in underlying specification of voicing (D & V) occur significantly less often.

The preference of voiceless obstruents before the fillers, might be triggered by the voiceless glottal stop initiating the filler. The labeling of the glottal activity shows the presence of burst (B) as extremely rare (9 cases, i.e. 1.2%) and modal voice (M) and glottalization (G) as the most frequent ones. The frequencies of M and G types varied significantly with the boundary strength; X^2 [3] = 55.6, $p < 0.001$. The analysis of residuals showed that their frequencies in the #4 break did not differ significantly but they did for the other 3 break strengths: G-tokens were more frequent for #2 and #3 breaks and less frequent for #1 breaks than M-tokens. This supports previous research in that glottalizations signal prosodic boundary.

Although the type of glottal activity varied significantly for the four assimilation types; X^2 [3] = 13.47, $p < 0.01$, only one cell – glottalization in the blocking of voicing (K) – was significantly different from the expected frequency. Surprisingly, this frequency was significantly lower than expected. This suggests that the phonetic realization of the filler (initiated with a glottal closure or without it) does not predict whether the filler triggers voicing assimilation. Additionally, modal voice was more frequent for all assimilation types, but glottalization appeared for each type as well. These observations suggest that the phonetic realization of fillers and triggering voice assimilation are not linked.

Finally, we examine CF duration as the only continuous feature of this analysis. Fig. 2 shows the data. We observe a tendency for the CFs following surface voiceless obstruents (D, K) to be longer than those following the voiced ones. A mixed models test [24] showed a weak overall effect of voicing type on CF duration (F = 3.02) and Monte-Carlo simulations revealed that the only significant pair-wise difference was between the K and N voicing types ($p = 0.003$).

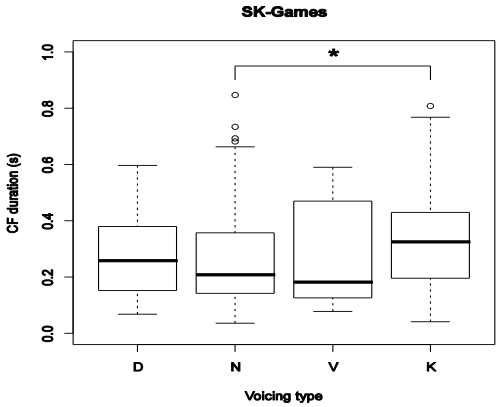


Figure 2: *Duration of conversational fillers for the four voicing types from Table 1.*

3.2. Court-TV corpus

Table 3 summarizes the behavior of fillers with respect to word-final voicing; see text above Table2 for the explanation of labels. The effect of the prosodic strength of the disjuncture corroborates the tendency of fillers to act as prosodic breaks (and either induce or not block word-final devoicing).

Table 3. *Voice assimilation types in Court-TV corpus*

Type	UR	SF	1	2	3	4	Total
D	V+	V-	59	7	4	NA	70
N	V+	V+	11	0	0	NA	11
V	V-	V+	14	0	0	NA	14
K	V-	V-	77	13	3	NA	93
Total			161	20	7	0	188

Contrary to *SK-Games*, the 2x2 contingency table for voiced/voiceless obstruents underlyingly and on the surface shows no significant deviation from the expected values; X^2 [1] = 0.1, $p = 0.92$. This corroborates a strong bias for SF [-voice] obstruents preceding a CF irrespective of their UR.

Regarding the possible interaction between word-final voicing alternations and the vocal cord activity initiating the

fillers, data in this corpus show no token with the burst, very few cases of glottalizations (N=24, i.e. 12.8%), and thus predominantly the initiation of fillers with modal voice. Moreover, all cases of glottalization occurred in tokens with surface voiceless word-final obstruents (11 for D and 13 for K types). Hence, in this corpus, the phonetic realization of the CFs reflected their phonological behavior: glottalization, assumed to reflect CF-initial voiceless glottal stop, co-occurred with devoicing of word-final obstruents resulting thus in the voicing agreement in these 24 tokens. Nevertheless, surface voiceless word-final obstruents were significantly more likely to be followed by CF initiated with regular modal voice than with glottalization (139 vs. 24). This provides additional support for CFs functioning as prosodic breaks rather than participating in phonetic voicing assimilations. Finally, the Fisher's exact test shows that the observed frequency of glottalizations for #2 and #3 breaks (combined to prevent low cell counts) is significantly greater than the expected one. This supports the observation from *SK-Games* and literature that initial glottalizations for vowel-initial words serves as one of the signals for prosodic boundaries.

As discussed in Section 2.2, the annotation of this corpus included binary coding that disregarded CFs and marked the voicing of word-final obstruents as either consistent or inconsistent with the general assimilation processes as described in Section 1.2. This served to test the hypothesis that fillers are transparent and function as neither prosodic breaks that induce devoicing nor as vowel-initial words triggering voicing. Data from this labeling suggest the rejection of this hypothesis. Disregarding CFs shows a slightly greater frequency of cases respecting the assimilation processes than those not respecting them (104 vs. 72), and exact binomial calculation shows that this split is significantly different from 88-88 split ($p = 0.019$). However, by the same token, 72 cases (41%) in which the surface word-final voicing with omitted CF did not respect typical Slovak voicing alternation is sufficient for rejecting the hypothesis. Moreover, the distribution of these two categories did not vary significantly in the four major assimilation types; $X^2[3] = 0.3$, $p = 0.96$.

Finally, Fig. 3 completes the analysis with CF durations in the similar way to Fig. 2 for *SK-Games*. Although the tendencies in the two figures are similar, a mixed-models test showed neither any overall significant effect nor any significant pair-wise comparison.

4. Discussion

The data from both corpora revealed an overwhelming tendency for surface devoicing of word-final obstruents. While *SK-Games* showed this only for underlyingly voiceless obstruents, *Court-TV* showed this trend irrespective of the underlying voicing of the word-final obstruents preceding the filler. This result supports the hypothesis that CFs function as prosodic boundary markers even if the disjuncture between a word and the following CF is minimal (ToBI's #1 break).

Regarding a possible phonetically-based explanation for the observed voicing assimilations preceding CFs, the results from both corpora reject the hypothesis that the phonetic realization of the filler, i.e. whether it is initiated with a glottal closure or without it, predicts whether the filler supports surface voicing of preceding word-final obstruents. Moreover, the data in both corpora agree with the expected function of glottalization in vowel-initial of increasing saliency of disjunctures: In both corpora, CFs initiated with glottal voice

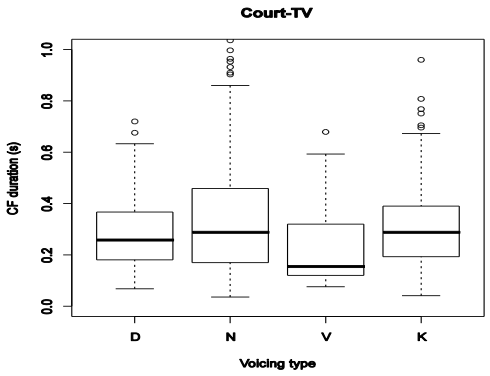


Figure 3: *Duration of conversational fillers for the four voicing types from Table 1.*

were more frequent in #2 and #3 breaks than those in #1 breaks. Interestingly, in *SK-Games*, there was no difference in the major #4 breaks, which suggests that glottalization may function as an additional boundary signal in Slovak for disjunctures of lower boundary strengths.

In spite of the overwhelming tendency for word-final devoicing before CFs, both corpora include small, but non-negligible, numbers of tokens in which CFs trigger (V), or at least do not block (N), voicing of word-final obstruents. These tokens suggest that CFs' participation in voicing assimilation cannot be completely refuted. Furthermore, following the discussion in Section 1.3, the account for these tokens require either treating CFs as 'regular' words or treating voicing assimilation as a purely phonetic process. Interestingly, further analysis of these tokens in *SK-Games* reveals that 21 out of 23 N-type tokens include prepositions like *od* and *z*, both meaning 'from'. Short Slovak prepositions (consonantal or mono-syllabic ones) tend to restructure with the following word and form a single prosodic word. The intervening surface filler and its tendency to function as a prosodic disjuncture might thus clash with the intended single prosodic word unit.

Regarding the duration of fillers, our data showed tendencies for the longer CFs to be preceded by voiceless obstruents on the surface (more so for the K type in which underlyingly voiceless consonant remains voiceless in the environment of the following CF). This observation is consistent with the analysis that CFs function as prosodic boundary markers since the longer the CF, the stronger is presumably the prosodic break signaled by this CF, and thus greater is the tendency for devoicing of word-final obstruents. However, this should be treated as a preliminary finding since CF durations were not normalized and possible silent pauses between the CF and the following words were not considered.

In sum, despite a strong bias for word-final devoicing preceding CFs, and thus their functioning as prosodic boundary markers on par with silent pauses, the hypothesis that CFs participate in voicing assimilation cannot be rejected.

5. Acknowledgements

We thank E. Cyran for inspiring discussion on Krakow Polish voicing and fillers. This research was supported by the ERDF's Research & Development Operational Programme, grant ITMS 26240220064 (RPKOM).

6. References

[1] Shriberg, E., “To "Errrr" is human: Ecology and acoustics of speech disfluencies,“ Journal of the International Phonetic Association 31(1): 153-169, 2001.

[2] Beňuš, Š, Enos, F., Hirschberg, J., Shriberg, E., “Pauses and deceptive speech,” 3rd International Conference on Speech Prosody, Dresden, 2006.

[3] Christenfeld, N., “Does it hurt to say um?” Journal of Nonverbal Behavior, 19: 171 – 186, 1999.

[4] Stenström, A., “Pauses in monologue and dialogue,“ In J. Svartvik (ed.) *London-Lund Corpus of Spoken English: Description and Research*, Lund: Lund University Press, 1990.

[5] Brennan, S., Williams, M., “The feeling of another’s knowing: prosody and conversational fillers as cues to listeners about the metacognitive states of speakers,” Journal of Memory and Language, 34: 383–398, 1995.

[6] Swerts, M., “Conversational fillers as markers of discourse structure, “ Journal of Pragmatics 30: 485–496, 1998.

[7] Bortfeld, H., Leon, S., Bloom, J., Schober, M., Brennan, S., “Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender,“ Language and Speech 44(2): 123-147, 2001.

[8] Taboada, M., “Spontaneous and non-spontaneous turn-taking,” Journal of Pragmatics 16(2-3): 329-360, 2006.

[9] Stewart, O., Corley, M., “Hesitation disfluencies in spontaneous speech: the meaning of um,“ Language and Linguistics Compass 4: 589–602, 2008.

[10] Shriberg, E., Lickley, R., “Intonation of clause-internal filled pauses,“ Phonetica 50: 172-179, 1993.

[11] Beňuš, Š., Mády, K., “Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels,” Proceedings of 5th Speech Prosody Conference, 2010.

[12] Cyran, E., “Polish voicing,” Lublin: KUL, 2013.

[13] Port, R., Crawford, P., “Incomplete neutralization and pragmatics in German,” Journal of Phonetics, 17: 257-282, 1989.

[14] Slowiaczek, L., Dinnsen, D. “On the neutralizing status of Polish word-final devoicing,” Journal of Phonetics, 13: 325-341, 1985.

[15] Dinnsen, D., Charles-Luce, J., “Phonological neutralization, phonetic implementation and individual differences,” Journal of Phonetics, 12: 49-60, 1984.

[16] Bárkányi, Z., Kiss, Z., “Phonological categoricality vs. phonetic gradience: The laryngeal properties of Slovak three-consonant clusters”, paper presented at the 11th Old-World Conference in Phonology, Leiden, the Nederlands, 2014.

[17] Beňuš, Š., “Cognitive aspects of communicating information with conversational fillers in Slovak”, Proceedings of 4the IEEE Conference of Cognitive Infocommunication, 2013.

[18] Rubach, J., “Nonsyllabic analysis of voice assimilation in Polish.“ Linguistic Inquiry, 27:69–110, 1996.

[19] Lombardi, L., “Positional faithfulness and voicing assimilation in Optimality Theory,“ Natural Language and Linguistic Theory, 17:267–302, 1999.

[20] Gafos, A., Beňuš, Š., “Dynamics of phonological cognition,” Cognitive Science, 30: 905-943, 2006.

[21] Gravano, A., Hirschberg, J. and Beňuš, Š., ”Affirmative cue words in task-oriented dialogue,“ Computational Linguistics, 38(1): 1-39, 2012.

[22] Darjaa, S., Cerňak, M., Trnka, M., Rusko, M., Sabo, R., “Effective triphone mapping for acoustic modeling in speech recognition,” INTERSPEECH, pp. 1717–1720, 2011.

[23] Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework," S.-A. Jun, ed., Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, pp. 9–54, 2004.

[24] Baayen, R. H., “Analyzing Linguistic Data. A Practical Introduction to Statistics Using R,“ Cambridge: Cambridge University Press, 241–302, 2008.

6 Tuesday 2

Labeling expressive speech in L2 Italian: the role of prosody in auto-and external annotation

Marta Maffia, Elisa Pellegrino, Massimo Pettorino

Department of Literary, Linguistic and Comparative studies,
University of Naples "L'Orientale", Italy
{mmaffia, epellegrino, mpettorino}@unior.it

Abstract

The present study is intended to compare two approaches of labeling expressive corpora: auto-annotation and annotation by external lay listeners. These two methods have been applied to the semi-spontaneous emotional speech produced by Chinese learners of L2 Italian, involved in the CardTask, a mood-induction procedure that allows us to control the context of interaction, preserving the spontaneity of reactions.

The emotional responses to the stimuli presented in the task were the object of an auto-annotation session. The same samples were then administered only in the auditory mode to 20 Italian and 20 Chinese lay listeners. The results of perceptual tests have underlined some similarities and differences between both auto- and external annotation, and between the ratings given by external Italian and Chinese listeners. The labels chosen by native Italians were similar to those selected in the auto-annotation session, particularly in the case of anxiety, fear and disgust. The correspondence between the results of the two annotation methods may be ascribed to the different prosodic patterns characterizing the emotional states. The results of the annotation made by Chinese listeners show that they found it hard to give a specific emotional label to utterances produced in a second language relying solely on prosodic patterns.

Index Terms: L2 emotional speech, prosodic cues, emotion annotation methods

1. Introduction

The study of emotional speech poses many problems, both methodological and analytical. When dealing with the collection of expressive spoken corpora, the first issue to be addressed is the style of speech to analyze. According to Scherer taxonomy [1], three kinds of productions can be considered, each with some points of strength and weakness: recited, induced or spontaneous emotional speech. Other methodological problems concern the nature of the speakers involved (actors or naïf), the kind of stimulus for emotion elicitation (pictures, videos, dyadic interactions) and the linguistic materials analyzed (nonsense words, syllables, interjections, utterances) (see [2] for a review of the most common elicitation techniques).

In addition to these methodological difficulties, another issue to consider is the labeling of emotions, particularly in the case of authentic expressive speech. The most common approach is the annotation by human experts, trained to deduce labeling paradigm from theoretical hypotheses on the nature of emotions. The limits of this annotation technique, though presumed to be the most objective, have been already underlined [3]. Alternative methods used to label emotional speech are the administration of perceptual tests to naïve listeners and the auto-annotation technique, in which the speakers themselves are asked to judge their own emotional state. There is no doubt that the annotation method chosen has an impact on the labels obtained and that, in order to test the validity of emotional labels, various methods should always be combined [4].

Despite the complexities of this kind of research, the identification of acoustic correlates of emotions has been the object of many studies, both on a production and perception level (see [1], [5] for a review). A correlation between the activation dimension and the most frequently measured acoustic parameters has been demonstrated [6], [7]. High activation emotions (such as fear, joy, surprise and anger) are generally characterized by shorter pauses, a wider tonal range, higher values of F0 and intensity, and faster speech rate. Low activation emotions (for example sadness and disgust), by contrast, are vehiculated by longer pauses, a narrower tonal range, lower values of F0 and intensity, and a slower speech rate.

Moreover, cross-linguistic studies on emotional speech encoding and decoding have emphasized the role of prosodic features in the identification of different emotional categories and have indicated that specific emotional states are vehiculated by universal prosodic patterns [8], [9].

Although studies on vocal emotions have been conducted for a wide range of languages, L2 emotional speech has not yet been extensively analyzed, either from the acoustic or from the perceptual point of view. Previous researches on expressive interlanguage, carried out on learners of L2 English with different mother tongues, have focused mainly on the emotional force of swearwords, taboo words or love words, when pronounced or listened in a second language [10], [11], [12], [13]. The impact of emotional expressions in L1 and L2 has also been the object of psycho-physiological studies. Harris et al. [14], for example, monitored the skin conductance of Turkish–English bilinguals via fingertip electrodes while they were rating for pleasantness a variety of stimuli in Turkish (L1) and English (L2). The results of this study demonstrated a difference between L1 and L2 emotional forces, being more noticeable in late bilinguals [15], [16]. Harris' hypothesis, therefore, is that L1 is the language of emotional expressiveness, while L2 is that of emotional distance.

2. The study

The present study has a twofold objective: firstly it is intended to identify the acoustic features of expressive speech in L2 Italian on the basis of an auto-annotation labeling; secondly, in order to verify the effectiveness of the auto-annotation, the auto-labels are compared to those given by external Italian and Chinese listeners. To achieve this, 10 Italians and 10 Chinese learners of L2 Italian (C1 Level – CEFR [17]) were involved. They were all female, university students, aged between 18 and 23.

The high level of linguistic competence has allowed the non-native students' active participation in the task and, as a consequence, the collection of a large corpus of emotional speech in L2 Italian. The decision to recrute Chinese learners depended on the results of previous studies, according to which the expressive speech of Chinese speakers is characterized by a more moderate and restrained style than that of Italians [18].

In order to collect emotional speech, Italian and Chinese participants were involved in the CardTask, a mood-induction procedure that allows us to control the context of interaction preserving the spontaneity of reactions.

2.1. The CardTask

The CardTask is a speaking activity where a Giver and a Follower work in pairs. They sit at the same table but they cannot see each other because of the presence of a dividing panel. The Giver receives five cards; the Follower is given a deck of 25 cards. The Giver has to describe her five cards (fig. 1) and the Follower has two minutes to find each card in her deck. The Follower's task is rather difficult because the deck consists of very similar cards only differing from each other in small details (fig. 2).

Figure 1: *Giver's cards.*

Figure 2: *Examples of some cards in the Follower's deck.*

In the room there is also an experimenter, whose function is basically to ensure that the task is performed as planned and to control some disturbing unexpected events which occurred during the game [19]. The following events were designed to elicit emotional linguistic reactions in the players and to arouse five different emotions (anger, anxiety, disgust, fear and surprise):

1. at the beginning of the game, while the Follower is looking for the first card in the deck, the chronometer rings after only 20 seconds instead of the expected two minutes;
2. the experimenter pretends to find a big beetle (obviously fake) in the room;
3. the experimenter tries to leave the room but the exit door seems to be temporary locked;
4. the fifth card is not in the deck.

In this study, the CardTask game was organized in two sessions: one only involving Italian speakers, the other only Chinese participants. The interactions took place in the silent chamber of University of Naples "L'Orientale" and were videotaped.

2.2. Method

2.2.1 Acoustic analysis

In order to identify the emotional responses to the stimuli presented by the experimenter during the task, we extracted the utterances occurring in the time interval between the end of each stimulus and the resolution of the unexpected events. A total of 132 utterances of expressive speech was collected and, by means of Praat, the whole corpus was segmented and annotated in two tiers: syllables and speech runs.

For each speech portion, we measured the duration and number of syllables, the length of burst phenomena, silent and filled pauses, and the lowest and highest F0 values. We consider bursts as "very brief, discrete, non verbal expressions of affect in both face and voice triggered by clearly identifiable events" [20]. On the basis of these measures, we calculated the following indexes: articulation rate (AR) (syll/s), speech time composition (percentage of silence, disfluence, syllables and burst), and tonal range (st). Additionally, in order to highlight the F0 variations connected to the considered emotions, for every utterance we related the F0 min and max values to the lowest F0 value - the F0 floor (st) - reached by the speaker in the whole corpus.

In order to preserve the spontaneity of the interaction, we did not control the movements of the participants in the room and thus the production of background noises. For this reason we decided not to consider the intensity and voice quality features.

2.2.2 Labeling emotions

The extracted emotional speech samples, both in L1 and L2 Italian, were object of an auto-annotation session. Moreover, in order to obtain more reliable labels and to assess the perceptual effect of emotional speech in a second language, the auto-labels were compared to those given by external Italian and Chinese lay listeners (henceforth external annotation).

During the auto-annotation session, the Italian and Chinese participating in the CardTask were instructed to watch their video recordings and to annotate them with one of the six labels, the five expected emotions, plus a generic option "other". In the case of auto-annotation, the video was supposed to help the players contextualize the utterances and recall what they were feeling during the task.

As for the external annotation, since our attention was focused exclusively on the acoustic correlates of emotions in L2 and on their communicative effectiveness, the emotional utterances of Chinese speakers were administered only in the auditory form to 20 Italian and 20 Chinese listeners. They labeled the utterances following the same protocol adopted for the auto-annotation session.

In order to prevent Chinese participants from misunderstanding the relationship between the labels of emotions and referents of these labels, a native Chinese speaker, specialized in Italian language and linguistics, translated the emotion categories to their L1.

It is important to underline that the acoustic data presented in the following paragraph are organized on the basis of the auto-annotation labels.

3. Results of spectro-acoustic analysis

Figure 3 illustrates the mean values of tonal range and F0 height (register) for each emotion in L1 and L2 Italian.

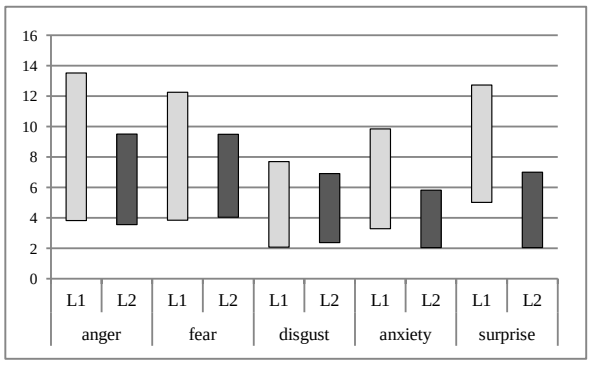


Figure 3: *Tonal Range and F0 height in L1 and L2 emotions (st).*

As the chart shows, the values regarding L1 Italian confirm the patterns for the high and low activation emotions, mentioned in the relevant literature. Of course anger, fear and surprise correlate with the highest F0 values and the widest tonal range. Anxiety, a high activation emotion as well, presents pitch height and tonal range values that are slightly lower than surprise, but higher than disgust, a low activation emotion [21].

Shifting our attention to the acoustic correlates of the emotions expressed by the Chinese subjects, it is possible to underline that the F0 values hardly match those attained by native speakers. As a matter of fact, F0 height and tonal range are quite steady in the whole corpus. The only exception is represented by anger and fear that are expressed with slightly higher values. These data seems to suggest that Chinese learners do not vary their pitch contour to distinguish different emotional states as in the case of native Italian speakers. After all, smaller pitch excursions are not only unique to L2 emotional speech, but they also represent one of the main acoustic correlates of Chinese accented Italian [22], [23].

Another parameter under study was articulation rate. Table 1 shows mean values for each emotional state in L1 and L2 Italian.

Table 1. *Articulation rate (syll/s) in L1 and L2 Italian.*

	Anger	Fear	Disgust	Anxiety	Surprise
L1 Italian	6.2	5.6	5.5	6.1	6.4
L2 Italian	5.2	5.6	4.3	4.8	4.8

Before considering AR variations in expressive speech, it is worth underlining that AR is a quite stable parameter, communicative setting being equal. With the exception of fear where L1 and L2 speakers reach the same values, in the other cases non-native utterances are produced with a slowing down of 1 or 1.5 syl/s, thus confirming the data available in the literature on foreign accented Italian [24]. The lower values of AR in L2 are essentially due to the learners' greater accuracy in uttering the single vowels and consonants. In order to reach the articulatory targets, the lengthening of syllable duration is needed, with a consequent slowing down of the speech.

Nevertheless, variations in AR do not seem to correlate with the five target emotions either in L1 or in L2. The only exception is represented by disgust in L2 Italian, whose values are particularly low (4.3 syll/s.) with respect to the whole corpus.

Figure 4 shows the composition of the utterance for each emotion in the two groups of participants.

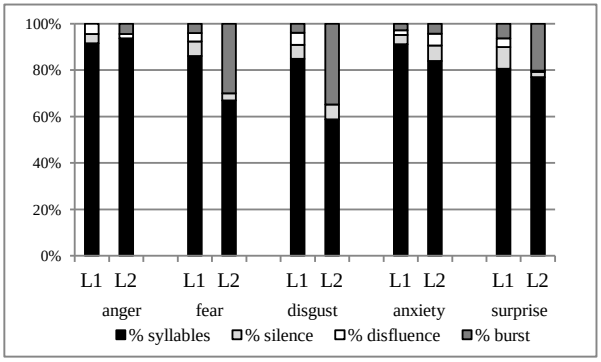


Figure 4: *Speech time composition in L1 and L2 Italian.*

In L1 Italian the percentage of syllables ranges from 85% to 95%, while in the second language it changes in the range of 60% and 93%. In the emotions that mostly differentiate L1 and L2 (fear, disgust and surprise), the bursts increase at the expense of syllable percentage. This means that when non-native speakers are not able to verbalize their emotions, bursts substitute the textual component. However, non-verbal expressions are not equally distributed between the five emotions. For example, in anger they are nearly absent, on the contrary, in disgust they represent one-third of the total time of the utterance. This datum is in line with Schröeder's findings [25]. Accordingly there are some emotions like disgust that are typically expressed by bursts, while others such as anger are not. A further observation is that in highly emotionally charged situations, L2 expressive speech is characterized by a very low percentage of disfluencies. L2 speakers rely on spontaneous and relatively universal vocal expressions to vehiculate their emotional states, instead of editing their performance with self-repairs, repetitions and substitutions. On the other hand, in L1, the percentage of disfluencies, above all in the form of lengthenings and vocalizations, is rather constant in the whole corpus. Such kind of filled pauses are typical of spontaneous speech and signal that the speaker is planning his/her speech.

4. Comparing auto- and external annotations

In order to verify the validity of labels given in the auto-annotation and to evaluate the communicative effectiveness of the selected expressive utterances in Chinese-accented Italian, we proceeded to compare the results of auto-annotation with the perceptual judgments given by external Italian and Chinese listeners. The results of the comparison underline some similarities and differences between auto- and external annotation and between the ratings given by native and non-native listeners. Tables 2 and 3 show the confusion matrices of auto- and external annotation by the two groups of listeners.

Table 2. *Confusion matrix of auto- and external annotation by Italian listeners.*

		External annotation				
		anger	fear	disgust	anxiety	surprise
Auto-annotation	anger	**29.2**	9.2	3.8	23.1	33.1
	fear	10.3	**48.3**	11.5	16.7	11.1
	disgust	5.8	25.0	**44.2**	7.7	16.3
	anxiety	6.1	14.2	2.4	**61.1**	13.4
	surprise	9.0	10.7	0.6	48.5	**28.6**

Table 3. *Confusion matrix of auto- and external annotation by Chinese listeners.*

		External annotation				
		anger	fear	disgust	anxiety	surprise
Auto-annotation	anger	**26.0**	8.0	4.0	28.0	34.0
	fear	5.6	**42.2**	8.9	8.9	34.4
	disgust	10.0	13.8	**30.0**	13.8	32.5
	anxiety	4.8	14.4	7.4	**66.7**	5.9
	surprise	9.4	9.4	6.1	47.2	**27.8**

As we can infer from the matrix of table 2, the labels chosen by the native Italians are pretty similar to those selected in the auto-annotation session by Chinese speakers, particularly in the case of anxiety, fear and disgust. The correspondence between the results of the two annotation methods may be ascribed to the specific prosodic patterns characterizing these emotional states. Anxiety is expressed by the lowest register and the narrowest tonal range. Fear and disgust are vehiculated by the same intonational patterns as in L1 Italian, though with not such marked differences. Moreover, disgust presents the lowest value of articulation rate and the widest portion of bursts. The similarities between values of F0 and articulation rate for surprise and anxiety induce listeners to confuse these two emotions. The emotional state that scores the highest percentage of mismatching between auto- and external annotations is anger, because of the considerable distance of its prosodic pattern from the model produced in L1 Italian. This result can be also explained by considering social conventions and different cultural standards in the expression of emotional states. As it has been noted in recent literature, Chinese speakers tend to inhibit the expression of emotions, which could threaten relational harmony as in the case of anger [8], [18].

The results of the annotation made by Chinese listeners show that they found it hard to give a specific emotional label to utterances produced in a second language. With the exception of anxiety, recognized by more than 60% of non-native listeners, the ratings given by Chinese subjects to the other emotions are definitely more uncertain than those expressed by Italians and more subject to random variations.

5. Conclusions

The present study had a twofold objective: firstly we intended to highlight the acoustic differences in the expression of the same emotions in L1 and L2 Italian; secondly we aimed to compare two approaches of labeling expressive corpora (auto- and external annotations) and verify the perceptual effects played by the prosodic patterns of L2 speech on native and non-native listeners.

As regards the acoustic correlates of emotional speech, the comparison between L1 and L2 confirmed Harris' hypothesis, according to which the second language is the language of emotional distance. Indeed the expressive speech of Chinese learners is characterized by a lower degree of variability in terms of F0 register and tonal range, a slowing down of articulation rate, and a different speech time composition. The reduced competence in the second language determines a lower percentage of syllable time and an increase of the burst component. During the CardTask, Chinese participants were not allowed to use their L1, the language of emotional expressiveness, but at the same time they were not able to use Italian to express their emotional states. Consequently, they tended to overcome this difficulty by relying on bursts.

As for the perceptual effectiveness of prosodic cues in the expression of emotions in L2 Italian, the data show that there is a correlation between the labels assigned by the speakers themselves in the auto-annotation and those chosen by native Italian listeners. This is particularly evident in the case of emotions characterized by similar prosodic patterns in L1 and L2 Italian (anxiety, fear and disgust). The lack of shared prosodic models, on the contrary, provokes in the listeners a high degree of uncertainty when labeling emotional states.

The random judgments given by Chinese external listeners reveal the objective difficulty of L2 learners to identify a specific emotional state in utterances produced in a second language.

Although intra- and cross-linguistic studies have already emphasized the difficulty of find coherent labels to the different emotions, this task seems to be much more demanding when analyzing emotional speech in a second language. The reason can be ascribed to the status of the L2 as the language of emotional distance, both on acoustic and perceptual levels.

In a further step of the research, we intend to extend the CardTask to learners with different mother tongues in order to verify whether the acoustic patterns characterizing the expression of emotions in L2 Italian are imputable to L1 transfer or are constrained by interlanguage development. The administration of perceptual tests to native and non-native speakers of Italian with different L1s will enable us to evaluate the extent to which the prosodic patterns and the cultural standards of the first language influence the perception of L2 emotional speech.

Further analysis of the corpus presented in this study will include a more detailed description of burst phenomena both in L1 and L2 Italian and the evaluation of the possible effect of the Chinese tonal system on the intonational features of the second language.

6. References

[1] Scherer, K. R., "Vocal Communication of Emotion: a Review of Research Paradigm", Speech Communication, 40:227-256, 2003.

[2] Coan, J. A., Allen, J. J. B. [Ed], The Handbook of Emotion Elicitation and Assessment, Oxford University Press, 2007.

[3] Aubergé, V., Audibert, N. and Rilliard, A., "Auto-annotation: an alternative method to label expressive corpora", in Proceedings of LREC, 2006.

[4] Truong, K. P., van Leeuwen D. A., Neerincx M. A. and de Jong F. M. G., "Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion", in Proceeding of: INTERSPEECH 2009, 10th Annual Conference of the

International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009.
[5] Johnstone, T. and Scherer, K. R., "Vocal communication of emotion", in M. Lewis and J. Haviland [Eds], Handbook of emotion (2nd ed.), 220-235, The Guilford Press, 2000.
[6] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijik, M. and Gielen, S., "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis", in Proceedings of the EUROSPEECH 2001, Aalborg, Denmark, 3-7 September, 1:87-90, 2001.
[7] Schröder, M., Speech and Emotion research. An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis, PhD Dissertation, 2003.
[8] Yang, L., Campbell, N., "Linking form to meaning: the expression and recognition of emotions through prosody", in Proceedings of the 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland, 2001.
[9] Pfitzinger, H. R., Amir, N., Mixdorff, H., Bösel, J., "Cross-language perception of Hebrew and German authentic emotional speech", in Proceedings of ICPhS2011, Hong Kong, 1586-1589, 2011.
[10] Kaneko, T., "How non-native speakers express anger, surprise, anxiety and grief: a corpus-based comparative study" in M. Archer et al. [Eds], 384-393, 2003.
[11] Dawaele, J. M. and Pavlenko, A., "Emotion Vocabulary in Interlanguage", Language Learning, 52(2):263–322, 2002.
[12] Dawaele, J. M., "The Emotional Force of Swearwords and Taboo Words in the Speech of Multilinguals", Journal of Multilingual and Multicultural Development, 25(2-3): 204-222, 2003.
[13] Dawaele, J.M., "The emotional weight of I love you in multilinguals' languages", Journal of Pragmatics, 40(10): 1753-1780, 2008.
[14] Harris, C. L., Aycicegi, A. and Gleason, J. B., "Taboo words and reprimands elicit greater autonomic reactivity in a first language than in a second language", Applied Psycholinguistics, 24: 561-578, 2003.
[15] Harris, C. L., "Bilingual Speakers in the Lab: Psychophysiological Measures of Emotional Reactivity", Journal of Multilingual and Multicultural Development, 2:223-247, 2004.
[16] Harris, C. L., Gleason, J. B. and Aycicegi, A., "When is a First Language More Emotional? Psychophysiological Evidence from Bilingual Speakers" in A. Pavlenko [Ed], Bilingual minds: Emotional experience, expression, and representation, 257-282, Clevedon, Mulilingual Matters, 2006.
[17] Council of Europe, The Common European Framework of Reference for Languages: Learning, teaching assessment, Cambridge University Press, 2001.
[18] Anolli, L., Wang, L., Mantovani, F. and De Toni A., "The Voice of Emotion in Chinese and Italian Young Adults", Journal of Cross-Cultural Psychology, 39:565-598, 2008.
[19] Maffia, M., Pellegrino, E., Vitale, M., De Meo, A., Pettorino, M.,"Expressive (Inter)language: A new method to elicit emotional speech in L1 and L2 Italian", paper presented at WASSS (Workshop on Affective Social Speech Signals), Grenoble 22-23 August 2013.
[20] Scherer, K. R., "Affect Bursts", in S. H. M. van Goozen, N. E. van de Poll and J. A. Sergeant [Eds], Emotions, 161-193, NJ: Lawrence Erlbaum, 1994.
[21] Jones, M., Anagnostou, F. and Verhoeven, J., "The vocal expression of emotion: an acoustic analysis of anxiety", in W.S. Lee, [Ed], Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 17-21 August 2011, 982-985, 2011.
[22] De Meo, A. and Pettorino, M., "Prosodia e Italiano L2: cinesi, giapponesi e vietnamiti a confronto", in R. Bozzone Costa, L. Fumagalli and A. Valentini [Eds], Apprendere l'Italiano da lingue lontane: prospettiva linguistica, pragmatica, educativa, Guerra Edizioni, 59-72, 2011.
[23] De Meo, A. and Pettorino, M., "L'acquisizione della competenza prosodica in Italiano L2 da parte di studenti sinofoni", in E. Bonvino and S. Rastelli [Eds], La didattica dell'Italiano a studenti cinesi e il progetto Marco Polo, Pavia University Press, 67-78, 2011.
[24] Pellegrino, E., "The perception of foreign accent and speech. Segmental and suprasegmental features affecting degree of foreign accent in Italian L2", in H. Mello, M. Pettorino e T. Raso [Eds], Proceeding of the VIIth GSCP International Conference – Speech and Corpora, Firenze University Press, 261-267, 2012.
[25] Schröder, M., "Experimental study of affect bursts", Speech Communication, Special Issue following the ISCA Workshop on Speech and Emotion, 40:1-2, 99-116, 2003.

Socio-affective interactions between a companion robot and elderly in a Smart Home context: prosody as the main vector of the "socio-affective glue"

Yuko Sasa[1], *Véronique Aubergé*[1]

[1] LIG-lab CNRS, Grenoble University, France

Yuko.Sasa@imag.fr, Veronique.Auberge@imag.fr

Abstract

The aim of this preliminary study of feasibility is to give a glance at interactions in a Smart Home prototype between the elderly and a companion robot that is having some socio-affective language primitives as the only vector of communication. Through a Wizard of Oz platform (EmOz), a robot is introduced as an intermediary between the technological environment and some elderly who have to give vocal commands to the robot to control the Smart Home. The robot vocal productions increases progressively by adding prosodic levels: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. The elderly subjects' speech behaviors confirm the hypothesis that the socio-affective "glue" effect increase towards the prosodic levels, especially for socio-isolated people.

Index Terms: socio-affective "glue", human-robot interaction, affective prosody, elderly, Smart Home.

1. Introduction

Prosody carries emotional, socio-affective and interactional information where each language has its own values [10]. This communicative information appears in different prosodic levels as in non-lexical sounds. Those can be non-phonetic sounds like grunts, affect bursts or mouth noises [20,24], phonological as fillers, mind markers or interjections [25], or onomatopoeia, widely studied in Japanese [26]. These sounds that we can consider as pure prosodic tools, were studied for specific and supposed emotional [5] and pragmatic [1] functions, as well as moods, emotions, intentions, attitudes, cognitive processes and mental states also known as "Feeling of Thinking"[14]. Moreover lexicons, sentences and paraphrases prosodic form also support various socio-affective values [29]. These cues can be extended from simple sounds to sentences produced in a same context, which have been tested in synthesis [12,15,17]. Lately, the prosody carrying this communicative information was introduced as a way to develop "socio-affective glue", terms introduced for the first time in [2]. In face-to-face interaction, the communication channel used by the speakers depends of the context and their social role, giving clues on how people have to talk to each other. Moreover, the "glue" refers to the fact that the interlocutors build dynamically their relation and adjust constantly the way they converge or not, globally changing the basic communicative channel firstly introduced by the context and their role. However, this kind of dynamic process can be more difficult to handle for some isolated-person, typically the elderly [3] but also younger person, like the Japanese hikikomori [11] or autistic people [6]. Furthermore, imitation has been studied as a basic process to create the same kind of "glue" in children language acquisition [8] or as a primitive of robots learning [23]. By the way, since the 90's, Affecting Computing and multidisciplinary communities have been focusing their work on the face-to-face interactions, especially on facial, gestural and vocal expressions using virtual agents and robots as in various studies in social computing [4,16]. It is interesting to see that when a robot is not explicitly humanoid, human creates by himself a socio-affective relationship with this device toward its « pet » stance [7]. Because all these different prosodic levels have not been studied together particularly to see their functions in the "socio-affective glue" building, our work will test them progressively thanks to a robot interacting with elderly towards gradual vocal productions: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. This will be tested with a Wizard of Oz protocol in a Smart Home prototype that has been used to study speech in a natural context environment with vocal commands, including elderly speech recognition [28].

2. Methodology

First of all, this study is testing the feasibility of our protocol to observe the gradual prosody productions effect on the "socio-affective glue" develop or not between elderly who are giving Smart Home's vocal commands and a robot "controlling" this Smart Home. The spontaneous corpus briefly analyzed in this study has been collected thanks to the EMOX robot (www.awabot.com/) interacting with the elderly in a Smart Home prototype named DOMUS. A Wizard of Oz platform called EmOz [2] was developed to control both the robot and the environmental system. The Wizards followed an accurate script describing which robot vocal productions they have to use and when they need to be produced. The Wizards also controlled the Smart Home each time the subjects gave a vocal command addressed to EMOX as we told them it was the robot that controlled DOMUS. This protocol is associated to a scenario with a specific recruiting pretext we were able to play thanks to elderly's caregivers (www.bienalamaison.com). The caregivers who were accompanying the subjects during the experiment were aware of the Wizard of Oz trick and our scientific goal being the researchers accomplices.

2.1. Experimental tools

2.1.1. DOMUS, a voice commanded Smart Home

As concept of living labs [19], DOMUS is designed like a 40m^2 flat with a kitchen, a bedroom and a living room equipped with two cameras and two microphones in each room, and a shower room with a microphone. Here, we selected a few possible actions in DOMUS to propose 31 voice commands that we can simulate from a control room. E.g.*; Mettre/Eteindre la lumière (fr) – To turn on/off the light*

(en); Monter/descendre les stores (fr) - To up/down blinds; Moins/plus fort la télé (fr) - Lower/louder TV (en).

2.1.2. EMOX robot with gradual vocal productions

The EMOX robot used in this study doesn't look like a human or an animal which morphology can induce the way we picture this tool and would create artifacts that we cannot control, and so the only way to find some anthropomorphism for this robot is its speech. In primarily study we attempted to change the Fundamental Frequency (F0) of spontaneous vocal micro-expressions, without modifying the prosodic contours to find a coherent voice for the robot [22]. From this database, we selected some of non-lexical sounds as part of the robot speech (e.g. *euh, and laughs*). We also recorded extra-sounds *(see table1a*), lexicons/interjections (*see table1b*) and commands imitations (*see table1c*) all with specific prosodies that we suppose to have an effect on the "socio-affective glue". In order to have a homogeneous voice for all our stimuli, we used Voxal (www.nchsoftware.com/voicechanger), voice conversion software. For parameterization, we increased the pitch by 1.52 from the original female speaker's voice, who recorded the sounds and so the robot has got a pitched "cartoon-like voice". So the voice esthetics has a reduced anthropomorphism, as we only want to observe information on the communicative functions carried by the chosen prosodic forms. Table1 shows the 30 vocal stimuli used by EMOX. 16 other sentences (also with supposed "glue prosody") were also pre-recorded, used only if it is necessary reacting to keep the subject motivation [*e.g.: Bonjour, je suis Emox(fr)= Hello, I am Emox(en); Oui, je suis là, j'écoute(fr) =Yes, I am here, I am listening (en); Oh pardon(fr)=Oh sorry(en)*].

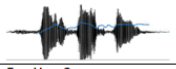
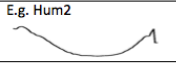
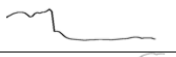
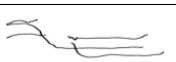
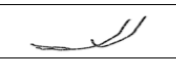
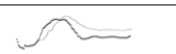
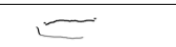
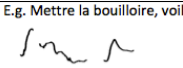
Sounds levels	Sounds nature	Examples of F0 contours
(a) Mouth noises	3 types of laughs	
	Various prosody : euh - hum1 - hum2	E.g. Hum2 
	prosody 1: humhum	
	2 onomatopoeia « woup » (associated to up/down blinds movements)	
(b) Interjections	prosody 1 : d'accord1 - ok1 - oui1	
	prosody 2: d'accord2 - ok2 - oui2	
	ça va - comme ça - ça y est	
	voila1, voila2	
	ouais, oops	
(c) Commands imitations	Commands in infinitive form + (b) interjections or (a) onomatopoeia	E.g. Mettre la bouilloire, voilà 

Table 1. *Emox robot stimuli types and Praat F0 contours*

2.1.3. EmOz – the Wizard of Oz Platform

In order to operate EMOX and DOMUS from the control room, we developed a Java interface associated to the robot Urbi system and the Smart Home protocols [2]. This interface is filled with buttons and each one of them controls one speech act of EMOX or one DOMUS command *(see Figure1)*. To move the robot we use a game controller so the robot can move around to follow the subjects, turn on itself, move its head (top/down, right/left) and go forward/backward. We also have the possibility to record our voice (changed through Voxal) and live play it if needed, to manage a coherent answer when the subjects' reactions were unpredictable.

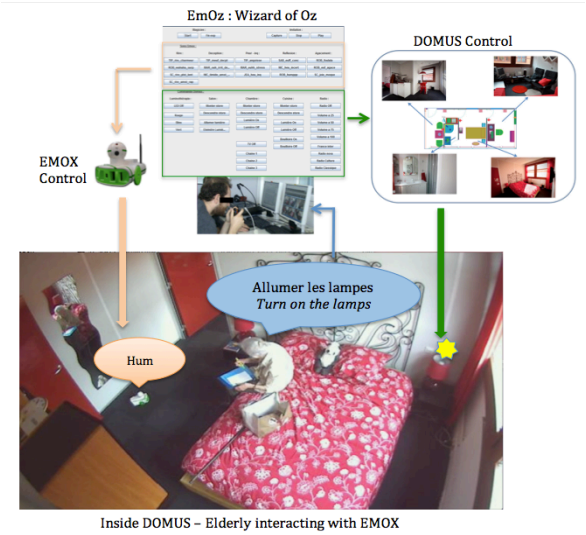


Figure1. *Scheme of EmOz wizard of Oz platform*

2.2. Experimental scenario

2.2.1. Subjects and caregivers

This study hired 4 French elderly subjects from 68 to 89 years old (one man and three women) accompanied by a caregiver/home helper they know well. The need of a helper was a criterion to select a "fragile but not sick subject", as it is difficult to find this kind of people and their degree of frailty corresponds to the scale 6 and 5 of the GIR grid [21], which is a French reference to evaluate an elderly ability. Moreover, these helpers were our accomplices, knowing the real goal of our study with the robot. Their role allowed us to reinforce the safety of the "fragile subjects". They also helped us to observe to observe the relation that the elderly had been creating with the robot. Furthermore, the helpers gave us an excuse to leave the subjects alone with the robot during one experiment step.

2.2.2. Experiment context and consent

To bring the subjects in DOMUS, we used a pretext to motivate the elderly to come, giving them a false task to do during the experiment. We did not mention the presence of the robot at first so they can interact spontaneously with it. Every subject was told to be the first participant of the experiment in order to keep a coherent scenario. In term of consent, the subjects had to sign three documents. Before the experiment we gave to the elderly a "pre-experimental consent" based on the pretext. Once they finished, they signed a "post-experimental consent" and an "image rights document" revealing the purpose of the Wizard of Oz. The helpers had their own single consent as the experimenters' accomplices.

2.2.3. Pretext of recruitment

We proposed a fake object of study to recruit the elderly. The subjects were asked to come in our Smart Home prototype because we were developing new technologies for elderly to allow them to live as long as possible in their own house, thinking of their safety and their welfare. Nevertheless we are currently not able to equip directly their home yet, so we need the subjects to come in our flat prototype. However, the problem we observed in "previous study" was that when elderly change their living environment (e.g. move into a retirement/nursing home or a hospital) they mostly have difficulties to accustom to this new place, especially when it is based on technologies. One of our "so-called hypotheses" was that if people bring some personal items (e.g. books, trinkets, decorative objects…etc.), which can be benchmarks as in Alzheimer disease, and they arrange their living environment with these items, people get used to the place more easily. Finally we asked the subjects to bring around ten items they choose and care about to come and spend some time (an hour or two) to evaluate our flat equipment, at the same time they personalize it with their own items. For their security, we asked them to be accompanied by one of their caregivers, who were also the experimenters' accomplices.

2.2.4. Experimenters roles

To play our experiment's scenario, we need two experimenter-actors interacting with the subjects:

- A "student-recruiter" who explains the pretext to the subjects and organizes the arrangement with the helper. He pretends not knowing what kind of technologies the flat is equipped with because it is his advisor who communicated with the engineering staff.
- An "engineer" who does not know the recruiter and was only asked by the student's advisor to explain how the Smart Home works. He is not aware of the study's aim.

Technically, we also needed at least two other experimenters to manipulate the EmOz platform from the control room: the first one used the java interface *(see Figure 1)* producing the EMOX vocal reactions *(in orange)* and the DOMUS commands *(in green)*. The second one was coordinating the robot movements. All these reactions were listed within a specific order regarding the voice commands that can be produced by the subjects, and these two Wizards followed this experimental script.

2.2.5. Experiment steps

The experiment scenario itself is divided into six steps:

- Step 1: the engineer welcomes the subject, the helper and the recruiter to present the Smart Home. He does not mention at this time the voice commands use and the robot is hidden. Very quickly, the helper receives a fake emergency call simulated from the control room, (before showing EMOX). The helper is asked to act like he has an urgent mission and needs to leave DOMUS for a short time, but she will come back very soon. The helper has not got any mean of transportation so the recruiter offers the ride, while the engineer explains to the elderly how the Smart Home works, waiting for their return. The helper and the recruiter use this excuse to go to the control room. Before leaving, the recruiter asks the subject to start arranging the flat with the personal items.
- Step 2: the engineer introduces EMOX to control the Smart Home. He presents a 31 commands list that can possibly be "activated". To work on a better quality speech signals, the engineer places a lapel microphone on the subject, justifying it by the fact that the robot vocal recognition system is not good enough without it. For the robot to able to recognize the subject's voice, the elderly is asked to try once all the voice commands one by one. When this training period is done, the engineer leaves the subject pretending to have some work to do.
- Step 3: It is an improvisation stage. The subject is alone with the robot, setting down his things in DOMUS.
- Step 4: After about 10 to 20 minutes, we ask the helper to go back into DOMUS and let the subject explain how the Smart Home works. If asked, the caregiver says that the recruiter was caught to talk about administrative stuffs and will return soon. EMOX does not realize the helper's commands and it only obeys to the subject.
- Step 5: When the subject has showed most of the voice commands to the helper, it is the recruiter turn to ask how DOMUS is working. If the subject himself tells something about the robot, the recruiter listens and asks about it, just out of curiosity. Nevertheless, the recruiter real task is to remind the pretext, so he starts a "fake interview" to know how the subject got used to DOMUS thanks to the items he brought.
- Step 6: During this debriefing stage, the engineer comes back and asks what the subject has thought about the voice commands technologies, without mentioning the robot at first. Here we want to know if the subject will tell about the robot and how he will talk about it. Once EMOX is evoked, we ask about the robot's voice, morphology, use, ways of speaking, functions, ability, role and personality. After this interview, the experimenters tell the real purpose of the study, revealing the Wizard of Oz tricks.

3. Discussions

3.1. The Elderly Emox Expressions Corpus

This corpus is composed by 4 experiments lasting from 1.5 to 2 hours each. For each subject, we have six videos (two per rooms) and an audio file collected by the subjects' lapel microphone. We then selected 167 interactions between EMOX and the elderly (from 43 to 52 per subject), throughout the interactions held in step 2 to 5, while the subjects: (a) are learning the commands with the engineer, (b) are alone with Emox, (c) are explaining how DOMUS and Emox work to the helper and then to the recruiter. Each interaction lasts about 10 to 50 seconds, showing a sequence of exchanges around one voice command. In this corpus, we observed two types of subjects' vocal productions of the "socio-affective glue": the commands form produced by the subjects and some of their answers after EMOX feedback. As the robot was introduced like a simple vocal commands receptor, there is no reason for the subjects to talk or ask other things to the robots. Moreover, as they had a specific list of voice commands to control DOMUS, they had no reason to produce other forms of commands as well as react to simple automation. Then, if they change their way to address their commands to EMOX or try to talk to him, we suppose that something in the robot's speech

prosody lead them to change their way to interact with it. The purpose of this for preliminaries analyses

3.2. Results

3.2.1. Debriefing global impressions

While debriefing, the main question we ask is "what opinion do you have concerning the system and the way you can control the Smart Home by voice?" (without mentioning at all the robot). For all the subjects, the first answer they gave us was: "the robot is a good company", so the technical support of the vocal commands is not mentioned first, meanwhile we introduced it in this way. Moreover, when we ask the subjects if they prefer the same commands but address directly to the Smart Home, they assert again that the robot is useful but not only: "we don't feel lonely, he talks to us". So we can say that some "socio-affective glue" was established between the robot and the elderly. We then want to know how by observing: (a) the commands form, particularly the prosody associated to them; and (b) the reactions of the subjects after EMOX feedbacks with gradual prosody levels.

3.2.2. Subjects commands tendency

During the interactions held in step 2 to 5, there are of course some variations concerning the subject's behaviors during the experiment first steps, but some common main characteristics emerged as features of the "glue" building toward different steps resumed in Figure 2: (a) declarative commands without paraphrasing; (b) the same original form commands but with a positive attitude prosody (in particular fundamental frequency arise which systematically appears at the end of the sentences, with a breathy voice); (c) commands paraphrased variations (used in synergy with a "we", illustrating the same kind of joint attention that we can observe in a mother-child dyad in an early language development context [27]) with a globally high fundamental frequency and a great arise at the end of the sentences; and finally (d) multiple prosodic focuses of support terms with a higher fundamental frequency. These phenomena are observed as well as a voice quality becoming more and more breathy. This elderly's voice quality breathiness seems to vary particularly while the robot produced a feedback based on pure prosodic vocal micro-sounds.

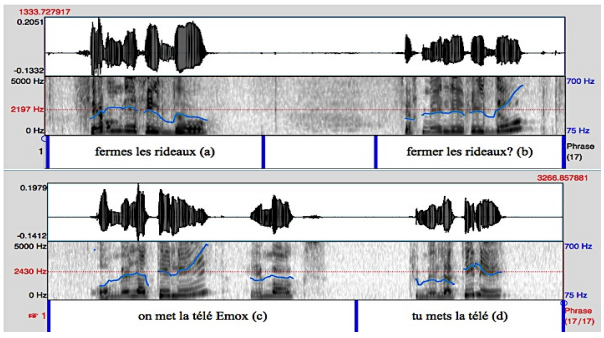


Figure 2. *Commands prosody: Praat spectrogram and F0 contours of some S1 subject's commands*

3.2.3. Subjects reactions to EMOX feedbacks

For every subject, nothing happened while there was no Emox speech reaction. The subjects just continue to read commands. But after the first appearance of pure prosodic mouth noises *(see Table1.a),* we always notice specific subjects reactions as showed in Table 2. Moreover, the voice breathiness seems to be amplified after mouth noises. Once the relation is established, the robots' lexicons and mouth noises become the base of the subjects' imitations and echolalia as answer to Emox feedbacks. This can be a clue on the degree of "glue" [13] to know when to introduce imitations that increase the "glue". The timing of EMOX's reactions and the silence duration are also an important factor in the "socio-affective glue" process that is also a major key for the interaction synchrony [9], but it also seems to be the moment that we can see if the "glue" is established. Indeed, even if EMOX (in fact the Wizards) makes mistakes or is slow, the subjects do not blame the robot but spend this response time justifying the commands, being politer or heartening it with compliments.

Types of subjects feedbacks	Quantity
Commands associated to politness and compliments	5
*Commands justifications***	13
Echolalia	10
Imitation	45
Kind repraoches	9
Mouth noises, interjections and laughs	49
*Politeness forms***	11
*Ponctual compliments***	16
Spatial proximity	28

**During silence or timelag after a vocal command

Table 2. *Subjects' reactions distribution as answer to Emox feedbacks*

4. Conclusion

The elderly's speech behaviors confirm that the effect of socio-affective "glue" increases towards the prosodic levels. This starts by the pure prosodic mouth noises that seem to be essential to initiate the relationship between the robot and a human. Once the link is created, the supposed prosodic "glue" tends to reinforce the bond through lexicons echolalia, imitations supported by a breathy voice quality. To have a cleaner corpora collection, the way to introduce the robot becomes essential as it induces the basic communication channel, where the glue is adding its effects after. To allow a precise control of the robot reactions timing and order, we need an efficient interface so the cognitive effort of the Wizard of Oz experimenter is the same as the effort the robot "seems to produce" to execute the commands. This will be one of these study perspectives, in addition to expanding this experiment with more various stimuli and subjects, on a revised and longer scenario to confirm those tendencies.

5. Acknowledgements

This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the Interabot project (Investissements d'Avenir, DGCIS) collaboration with *Awabot* (robotics) company. We thank *Bien à la Maison* company (elderly personal services) for their active participation to held the corpus. Best thanks to Tim Robert, Nicolas Bonnefond and Brigitte Meillon, for their precious technical contribution and their organization support for the experiments. Finally thanks to Leandra Batista for her advice on the results analyses and to Gilles De Biasi for his reviewing.

6. References

[1] Ameka, F (1992) "*Interjections: The universal yet neglected part of speech*", Journal of Pragmatics, 18, 101-118, 1992.

[2] Aubergé V., Sasa Y., Robert T., Bonnefond N., Meillon B. (2013) "*Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot*". In proceedings of WASSS 2013, Grenoble, France.

[3] Bayles K.A. and Kaszniak A. (1987) "*Communication and Cognition: Normal Aging and Dementia*", Little, Brown, Boston.

[4] Breazeal C. and Aryananda L. (2002) "*Recognition of affective communicative intent in Robot-Directed speech*", Autonomous Robots 12, pp 83-104.

[5] Campbell N. (2004) "*Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language*", Languages Resources and Evaluation, 39, 109-118.

[6] Chaby L., Chetouani M., Plaza M., Cohen D. (2012) "*Exploring multimodal social-emotional behaviors in autism spectrum disorders*". In Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust; IEEE Computer Society: Washington, DC, USA, 2012, pp. 950–954.

[7] Darling K. (2012). "*Extending Legal Rights to Social Robots*". We Robot Conference, University of Miami, April 2012.

[8] Decety J. (2007). "*A social cognitive neuroscience model of human empathy*". In E. Harmon-Jones & P. Winkielman (Eds.), Social Neuroscience: Integrating Biological and Psychological Explanations of Social Behavior (pp. 246-270). New York: Guilford Publications.

[9] Delaherche E., Chetouani M., Mahdhaoui A., Saint-Georges C., Viaux S., and Cohen D. (2012) "*Interpersonal Synchrony: A Survey of Evaluation Methods across Disciplines*". IEEE Transactions on Affective Computing 3(3), pp. 349-365.

[10] Fonagy P., & Target M. (1997). "*Attachment and reflective function: Their role in self-organization*", Development and Psychopathology, 9, pp. 679-700.

[11] Furlong A. (2008) "*The Japanese hikikomori phenomenon: acute social withdrawal among young people*", The Sociological Review, 56(2), pp. 309-325.

[12] Greenberg Y., Tsuzaki M., Kato H. and Sagisaka Y. (2006) "*A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech*". In Proceedings of Speech Prosody 2006, pp. 37-40.

[13] Ladd D.R. and Cuttler A. (1983) "*Models and measurements in the study of prosody*". In A. Cutler & D. R. Ladd (Eds.), Prosody: Models and measurements, Heidelberg: Springer-Verlag, pp. 1-10.

[14] Loyau F., Aubergé V. (2006). "*Expressions outside the talk turn: ethograms of the Feeling of Thinking*", 5th LREC, pp.47-50.

[15] Mac D., Castelli E, Aubergé V. (2012). "*Modeling the prosody of Vietnamese attitudes for expressive speech synthesis*". Workshop of Spoken Languages Technologies for Under-resourced Languages (SLTU 2012), Cape Town, South Africa.

[16] Mairesse F., Walker A., Mehl Matthias M. and Moore R.K. (2007) "*Using linguistic cues for the automatic recognition of personality in conversation and text*". In Journal of Artificial Intelligence Research 30, pp. 457-500.

[17] Morlec Y., Bailly G. and Aubergé V. (2001). "*Generating prosodic attitudes in French: data, model and evaluation*". Speech Communication, 33(4), pp.357-371.

[18] Morenc L.P (2010) "*Modeling human communication dynamics*", Social Sciences. Signal Processing Magazine, IEEE, 27(5), pp. 112 -116.

[19] Niitamo V.-P., Kulkki S., Eriksson M., Hribernik K. A. (2006) "*State-of-the-art and good practice in the field of living labs*". In Proceedings of the 12th International Conference on Concurrent Enterprising: Innovative Products and Services through Collaborative Networks, Milan, Italy, pp.349-357.

[20] Poggi I. (2008) "*The language of interjections*". In COST 2012 School, pp.170–186.

[21] Renault S. (2004). "*Du concept de fragilité et de l'efficacité de la grille AGGIR*". In Gérontologie et société. 2004, n° 109, pp.83-107.

[22] Sasa Y., Aubergé V., Franck P., Guillaume L., Moujtahid S. (2012). "*Des micro-expressions au service de la macro-communication pour le robot compagnon EMOX*", WACAI 2012, Grenoble, pp.54-59.

[23] Schaal S. (1999) "*Is imitation learning the route to humanoid robots?*" Trends Cognit. Sci.3, 233-242.

[24] Scherer K.R., "*Affect bursts*" (1994). In S.H.M. van Goozen, N. E. van de Poll & J.A. Sergeant (Eds.), Emotions, Hillsdale (NJ, USA), Lawrence Erlbaum, 161-193

[25] Schröder M., Heylen D., Poggi I. (2006) "*Perception of non-verbal emotional listener feedback*". In Speech Prosody, R. Hoffmann & H. Mixdorff, Eds., 2006, pp. 1–4..

[26] Shibatani M. (1990) "*The languages of Japan*". Cambridge: Cambridge University Press.

[27] Tomasello M., Carpenter M., Call J., Behne T. & Moll H. (2005) "*Understanding and sharing intentions: The origins of cultural cognition*". Behavioral and Brain Sciences 28(5), pp. 675–91.

[28] Vacher M., Fleury A., Portet F., Serignat J.-F., Noury N. (2010) "*Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living*", New Developments in Biomedical Engineering, Intech Book, pp. 645-673.

[29] Wichmann A. (2000) "*The attitudinal effects of prosody, and how they relate to emotion*". In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Towards Automatic Recognition of Attitudes: Prosodic Analysis of Video Blogs

Noor Alhusna Madzlan[1 3], *JingGuang Han*[1], *Francesca Bonin*[1 2], *Nick Campbell*[1]

[1] CLCS, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin
[2] SCSS, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] ELLD, Faculty of Languages and Communication, UPSI, Malaysia

madzlann@tcd.ie, hanf@tcd.ie, boninf@tcd.ie, nick@tcd.ie

Abstract

Understanding of speakers' attitude is essential for establishing successful human interaction. In this paper we analyse attitude manifestations in video blogs. We describe the main features of this novel communication medium and focus our attention on its possible exploitation as a rich source of information for human-human and human-machine communication. We describe the manual annotation of attitudes and the prosodic analyses. Finally we present a preliminary automatic attitude annotation system that attains 65% accuracy.

Index Terms: video blog, prosody, attitude, automatic classification, statistical modeling, SVM.

1. Introduction

Social media is becoming a major form of interaction and of personal expression. While the popularity of content based platforms, such as web blogs, Twitter and Facebook, confirms that written text is still the major form of online interaction, new forms of expression are evolving. Conversational video blogs (in short vlogs) are becoming a widespread phenomenon of online social media, which create a huge amount of user generated content. Video blogs can be defined as personal diaries made available to the larger public in the form of self-recorded videos, where the users express themselves, their personality and share life events. They combine the best qualities of pre-recorded broadcasted speech with the naturalness of spontaneous conversations. However, at the same time, the speaker expresses himself in a de-contextualized situation, addressing an imagined audience without being influenced by the listener's reaction, as would happen in a face to face context.

Video blogs have piqued the interest of scholars in recent years. The main stream of research on vlogs involved the study of personality recognition [1, 2] and they have been widely studied with respect to non-verbal behavior and social attention. However, to our knowledge, the linguistic pragmatic aspect of this new form of expression has been given little attention. Video blogs are a unidirectional form of communication where the user intends to convey a message, an emotion or a personal opinion. In this work we are not interested in investigating personality as reflected in video blogger behavior in front of a camera, but rather the impression of what he wants to transmit at a purely pragmatic level. For this reason, we explore five attitudinal classes that appear to be representative of the videos in our corpus and analyse their prosodic characteristics. In this work we define attitude as social affective states that the video bloggers intend to transmit and we rely on the classification of attitudes in [3]. We are not interested in the inner emotion of the video blogger but in what he intends to express. Finally, we explore the predictive value of extracted prosodic cues for automatic recognition of attitudes in video blogs.
Main contributions of this work are:

- We focus on a new social media, describing the potentialities of this source of material and a qualitative analysis of its characteristics.
- We describe a novel corpus of vlogs and its manual annotation with respect to attitudes.
- We analyse prosodic features of attitude impressions in video blogs.
- We address the task of automatically predicting video bloggers' attitude impressions using multimodal nonverbal cues and machine learning techniques.

The paper is structured as follows: Section 2 describes some of the previous literature on video blogs. Section 3 outlines the characteristics of the dataset. Section 4 explains the annotation process and adaptation of the schema. Section 5 presents selection and analysis of prosodic features. Findings and results are presented in Section 6. Section 7 explains and discusses the analysis in detail. We conclude the study in Section 8.

Figure 1: Example of Video Blogs

2. Related work

Numerous studies with relevance to various fields of research have been conducted on video blog analysis. Biel et.al [1] conducted a study on recognition of the Big 5 Personality Traits

[4] of video bloggers represented through non-verbal signals. They were able to predict personality traits including Extraversion, Openness, Agreeableness, Conscientiousness and Emotional Stability from analysis of non-verbal characteristics in video blogs. Prosodic and visual feature extraction was conducted to identify labels of personality annotated by five people using Amazon Mechanical Turk. Findings suggest that speakers who demonstrate higher pitch range and higher motion activity possess Extroversion and Openness traits.

Other related work involves multimodal sentiment analysis of opinion videos among Spanish speakers on YouTube [5]. Automatic feature extraction and prediction using audio, visual and textual features were conducted to identify sentiments of the speakers. Videos were labelled according to three positive, neutral and negative sentiments. Results showed that the smile offers the best feature prediction, while number of pauses and voice intensity followed suit. Positive sentiments are indicated by increased number of smiles and pauses, whereas negative sentiments are represented by higher voice intensity.

Treating attitudes as expressions of opinions made by the speaker towards related issues during interaction with their interlocutor, Mac et.al [6] investigated audio-visual prosodic attitudes involving cross-cultural relations. Their study examines perceptions of Vietnamese and French participants in identifying attitudinal expressions in the Hanoi standard dialect. A perception test was conducted and results showed that native listeners recognised attitudes better than foreign listeners. For Admiration, however, foreign listeners were able to successfully recognise the state better than native listeners.

Analysis on prosodic characteristics in speech not only facilitates but further enhances understanding of speakers' attitudes towards a particular topic or notion [3]. Henrichsen and Allwood [3] analysed the NOMCO speech corpus containing eight dialogues by automatically extracting multimodal features to identify attitudes. A standard set of ten attitudes called the A10-based annotation was developed. Multimodal features were extracted for automatic prediction of attitude categories and results showed that attitude labels could be predicted by the trained model.

3. Video Blog Dataset

This section introduces the dataset of videos that we use throughout the paper. We first describe the data collection process and then provide a high level analysis of the contents and of the type of speech typical of video blogs.

A total of 100 video blogs were selected from the YouTube channels of four different speakers [1]. The speakers have been selected according to the following characteristics:

- Native English speaker
- American English
- Male speaker
- Aged between 18-25 years old

Among the videos of these speakers, we selected the ones with the larger number of visualizations. On average each speaker is represented by about 25 videos. The videos were downloaded using a free add-on tool for Mozilla Firefox browser [2]. The audio tracks were extracted from each video using the same tool. Annotation and labeling was conducted manually using WaveSurfer [7], as in Fig.3.1. Sample rate for all videos is set at 44100bit rate with mono sound. Total duration of the corpus is 286 minutes, and the mean duration of each video is 2.88 minutes (sd=1.09).

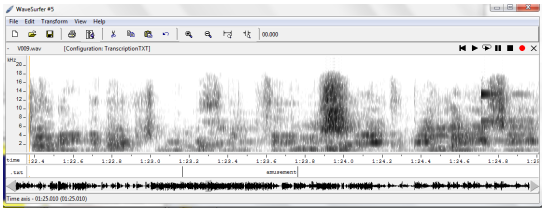


Figure 2: Annotation with Wavesurfer

3.1. Qualitative overview of the dataset

The video blogs in our dataset are typical of the video blog genre. They are pre-recorded monologues, recorded with the speaker facing the camera. They represent an expression of asynchronous communication with a delayed feedback provided by the comments of the audience, and they are usually stored in reverse chronological order. The speech is semi spontaneous. While in broadcast recordings, speech is typically prepared and scripted with a rigid format that the speaker needs to adhere to, in video blogs, the speech is more flexible. Video blogs may be characterised as prepared speech, because a substantial amount of time is given for preparation and planning prior to recording, but video bloggers do not strictly conform to the pre-written texts. As a result of this, utterances in video blogs resemble unprepared or spontaneous speech, particularly with regards to disfluencies, such as ungrammaticality, filled pauses, repetitions, repairs and false starts [8].

These two elements (the preparation of the script and the spontaneity) create a very interesting linguistic genre that preserves features of broadcast speech as well as features of natural spontaneous speech. The following is an example of video blog speech:

```
Hey guys!
If there's one thing I can't stand,
it's people who judge others.
Whether it's based on looks,
or what you heard about them, it
doesn't matter. If you don't know
the person personally, you have
no right to judge them. High
schools are the worst.
```

4. Annotation process

In the present work we are interested in analysing video blogger attitudes during the recording. In the unfolding of video blogs, the user is "playing" different roles and showing different attitudes to the audience. In order to conduct our experiments, we first labelled the corpus with respect to 5 main attitudes shown in Table 1. We base our annotation schema on the A10-set annotation identified in Henrichsen and Allwood [3].

[1] http://www.youtube.com/user/nigahiga
http://www.youtube.com/user/kevjumba
http://www.youtube.com/user/JustinJamesHughes
http://www.youtube.com/user/uncuthashbrown

[2] https://addons.mozilla.org/en-US/firefox/addon/easy-youtube-video-download/

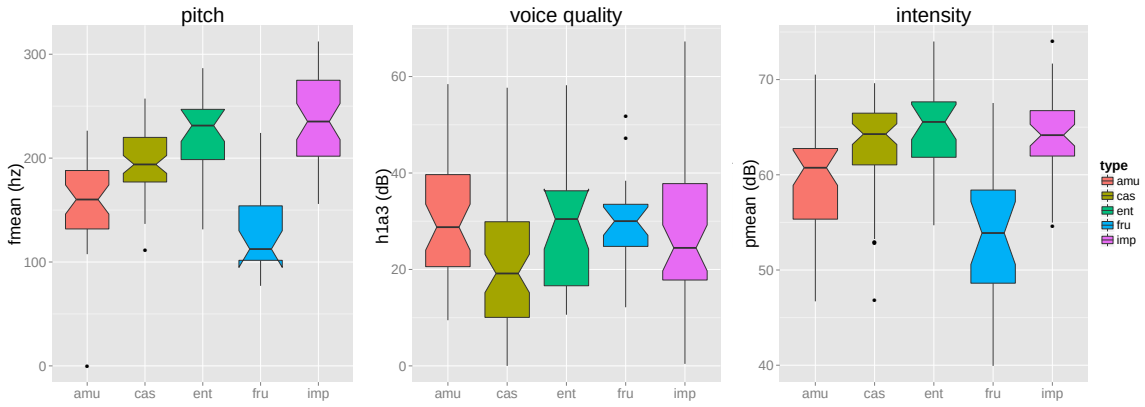


Figure 3: Pitch, voice quality and intensity distributions over the five attitude states (**amu**sement, **cas**ual, **ent**husiasm, **fru**stration, **imp**atience). The figure shows good separation in both Pitch and Intensity with clear differences in Voice Quality for Casual utterances.

The manual annotation took six weeks to complete. We annotated through several stages. A preliminary observation was devoted to understanding attitude expressions of video bloggers. After this qualitative analysis of the corpus, we selected from the A-10 the most representative attitudes for our dataset (Amusement, Impatience, Casual, Enthusiasm, Frustration). In a preliminary phase we listened to the audio files and marked segments of speaker higher activity. Subsequently, two raters were asked to label the segments using the annotation scheme presented in Table 1. Inter annotator agreement was calculated using Cohen's Kappa, and resulted in k=0.75 [9].

After labeling, the chunks' start and end times were extracted using a TCL/TK script.

Attitude	Description
Amusement	speaker laughs, chuckles
Impatience	speaker shouts, appears annoyed, harsh
Casual	speaker informally addresses the audience
Enthusiasm	speaker appears excited
Frustration	speaker appears 'defeated'

Table 1: Attitude Annotation schema

5. Feature extraction

The prosodic analysis was performed after extracting the following acoustic parameters from the video blogs audio channels, using a TCL/TK script:

- Fundamental Frequency (f0) - pitch level (high/low) [max,mean,min,median]
- Pitch target (fpct) - peak position of pitch (rising/falling)
- Voicing (fvcd) - percentage of voicing (vocal fold vibration) within each utterance
- Power/Intensity (dB) - loudness of the voice [max,mean,min,median]
- Power/Intensity movement (ppct) - peak position of power (rising/falling)
- Voice Quality (H1/a3) - tenseness of the voice (creaky/breathy) [h1h2,h1,a3] [10]
- Duration (dn) - length of the utterance (short/long)

In addition to the traditional prosodic parameters; pitch, intensity, and duration of the attitude segments, voice quality is included as a relevant acoustic parameter for communicative speech analysis. This parameter has been shown to have significant correlates with the interlocutor, speaking style and speech act [11].

5.1. Prosodic features analysis

In Table 2, we report the average and standard deviation of the prosodic values for each category. We notice that the attitudinal category Impatience shows the highest pitch and Frustration the lowest. Figure 3 (left and middle) shows the distributions of pitch and voice quality respectively. We observe that Frustration and Amusement differ significantly in terms of pitch from the other categories (lower pitch, $p<0.005$), while Casual significantly differs from the others with respect to voice quality ($p<0.005$). Impatience and Enthusiasm show similar distributions in terms of pitch, with a similar high average pitch. Figure 3 (right) shows the distributions of intensity. Again, Frustration is represented by speech with low intensity (significantly differing from the others, $p<0.005$), while, as expected, Impatience is the attitude characterised by a higher intensity. Significance was tested with a one tailed T-Test (alternative less).

6. Experiments and Results

In this section, we address the task of automatically predicting video bloggers' attitude impressions. Specifically, we were interested in assessing the prediction performances using the selected features for the annotated attitude classes. We will describe the experimental settings and present the results obtained. The study is based on the ground truth data collected (see Section 3) to evaluate classification performance; a user study focusing on the subjective assessment of the quality of the automatically extracted annotations is planned as future work.

We conducted our experiments using the data collection described in Section 4 in a 16 dimensional feature matrix. With the two different features sets as in Table 2, we trained a Support Vector Machine (SVM) with radial basis function (Gaussian) kernel as the classifier for the 5 attitude categories. We used a 10-fold cross validation approach to evaluate the trained model. We evaluated different feature sets:

Feature set ALL: All 16 features described in Table 2.

Type	fmean	fmed	fmax	fmin	fpct	fvcd	pmean	pmed	pmax	pmin	ppct	h1h2	h1a3	h1	a3	dn
Amused	162.97 *(36.09)*	172.24 *(50.94)*	228.04 *(59.84)*	103.01 *(30.43)*	0.43 *(0.25)*	0.55 *(0.19)*	59.86 *(5.75)*	60.51 *(5.88)*	75.37 *(5.38)*	36.71 *(10.43)*	0.49 *(0.27)*	6.49 *(4.74)*	30.00 *(12.17)*	-25.28 *(9.53)*	-55.28 *(5.81)*	1.07 *(0.31)*
Impatient	234.39 *(39.72)*	242.39 *(49.92)*	311.76 *(47.04)*	139.05 *(44.12)*	0.43 *(0.27)*	0.61 *(0.16)*	64.09 *(4.13)*	66.05 *(3.99)*	77.90 *(2.67)*	36.48 *(13.09)*	0.38 *(0.25)*	6.00 *(6.34)*	27.96 *(14.88)*	-26.06 *(11.18)*	-54.03 *(7.41)*	1.26 *(0.50)*
Casual	196.14 *(32.2)*	193.97 *(30.02)*	248.40 *(43.85)*	147.33 *(40.80)*	0.26 *(0.16)*	0.73 *(0.13)*	63.23 *(4.69)*	66.75 *(3.28)*	76.60 *(2.97)*	32.08 *(15.61)*	0.32 *(0.20)*	4.34 *(4.67)*	20.65 *(15.28)*	-34.72 *(17.54)*	-55.37 *(5.53)*	0.55 *(0.16)*
Enthusiastic	220.1 *(38.65)*	242.62 *(40.23)*	310.56 *(47.55)*	115.04 *(50.65)*	0.39 *(0.28)*	0.72 *(0.18)*	64.70 *(4.7)*	66.50 *(4.51)*	79.28 *(3.08)*	35.66 *(14.28)*	0.35 *(0.24)*	6.08 *(6.74)*	29.54 *(13.54)*	-26.72 *(11.87)*	-56.27 *(6.28)*	1.20 *(0.52)*
Frustrated	124.78 *(35.15)*	136.84 *(40.16)*	175.89 *(55.96)*	86.67 *(33.09)*	0.40 *(0.30)*	0.49 *(0.23)*	53.65 *(7.32)*	54.49 *(7.24)*	70.35 *(5.64)*	30.62 *(14.80)*	0.57 *(0.26)*	3.59 *(5.02)*	29.22 *(9.81)*	-27.43 *(9.72)*	-56.64 *(4.51)*	1.27 *(0.57)*

Table 2: Mean values for each attitude category with standard deviation in brackets

Feature set SEL$_1$: Select only the features that are not highly correlated (with Pearson's correlation coefficient $r < 0.7$): fmean, fmin, fpct, fvcd, pmean, ppct, h1h2, h1a3, h1, a3, dn. A correlation study of the feature set revealed that some of the features are highly correlated (correlation coefficient $r > 0.7$ with $p < 0.01$ in T-test). Only one of the features in the highly correlated feature pairs was selected and a new 11 dimensional feature set: SEL$_1$ was generated.

Table 3 shows the results of 10-fold cross validation SVMs with the different feature sets.

Feature Set	Accuracy
ALL	61.85
SEL$_1$	**65.46**

Table 3: Results for the different feature sets.

Results show that the feature set selected after removing the highly correlated features attained the best prediction accuracy. This result has also been compared with other feature set experiments by reducing the threshold of r to 0.65 or increasing to 0.75. SEL$_1$ resulted to be the feature set showing better prediction performance with a 65.46% accuracy rate.

7. Discussion

Understanding of speakers' attitude is not only essential for establishing successful human-human interaction, but could significantly contribute to the development of a robust system for human-robot interaction. Video blogs, spontaneous and intimate conversations, represent a rich source of natural attitude expressions.

We have reported the prosodic analyses of a collection of video blogs, for a better understanding of the acoustic dynamics behind five different attitudes including Amusement, Enthusiasm, Casual, Impatience and Frustration. Pitch and intensity show a strong discriminative value, and voice quality emerges as a feature characterising Casual friendly talk. We also investigate the predictive performance of prosodic cues and results from the automatic classification experiments show a prediction accuracy of 65.46%.

Although preliminary, results of this work are in line with [3]. In building conversational agents, attitude management (recognition and synthesis) is a key aspect. Given the sensitivity of conversational partners to an inadequate attitude response, attitude recognition is required to be reliable and robust. We believe that the combination of three main prosodic components such as pitch, intensity and voice quality can provide a robust attitude classification. In order to create a natural and spontaneous interaction, a conversational agent should also be able to synthesise the prosodic dynamics representative of the different attitudes. Our study goes in the direction of a better understanding of these dynamics.

8. Conclusion

In this paper we have explored attitude manifestations in video blogs. We have described the main features of this novel communication medium and focused attention on its possible exploitation as a rich source of information in human communication. We have presented a novel corpus of video blogs, its collection and annotation according to five attitudes, and analysed the main prosodic features characterising these classes. Finally we have presented a machine learning approach for the automatic detection of attitudes in video blogs and reported its preliminary results. Future work will be dedicated to extending the corpus and the annotations, and to exploring other feature selection approaches such as Principal Component Analysis.

9. Acknowledgements

This work is supported by the English Language and Literature Department, UPSI, Ministry of Education Malaysia, the Innovation Bursary of Trinity College Dublin, the Speech Communication Lab at TCD, and by the SFI FastNet project 09/IN.1/1263.

10. References

[1] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube." in *ICWSM*, 2011.

[2] J.-I. Biel and D. Gatica-Perez, "The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers," *Ethnicity*, vol. 16, no. 4.8, pp. 0–7, 2012.

[3] P. J. Henrichsen and J. Allwood, "Predicting the attitude flow in dialogue based on multi-modal speech cues," *NEALT PROCEEDINGS SERIES*, 2012.

[4] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[5] V. Rosas, R. Mihalcea, and L. Morency, "Multimodal sentiment analysis of spanish online videos," 2013.

[6] D.-K. Mac, V. Aubergé, A. Rilliard, and E. Castelli, "Cross-cultural perception of vietnamese audio-visual prosodic attitudes," in *Speech Prosody*, 2010.

[7] K. Sjlander and J. Beskow, "Wavesurfer - an open source speech tool," 2000.

[8] R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès, "Spontaneous speech characterization and detection in large audio database," *SPECOM, St. Petersburg*, 2009.

[9] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, p. 466, 1997.

[11] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, 2003, pp. 2417–2420.

Processing emotional prosody in Mandarin Chinese: A cross-language comparison

Pan Liu, Marc D. Pell

School of Communication Sciences & Disorders, McGill University
Centre for Research on Brain, Language and Music (CRBLM)
Montréal, Canada

pan.liu@mail.mcgill.ca, marc.pell@mcgill.ca

Abstract

To understand how emotional prosody is processed in Mandarin Chinese and whether it differentiates from that of other languages, we conducted a perceptual-acoustic study on a set of Chinese vocal emotional stimuli and examined how they were perceived and acoustically characterized, in comparison with four other languages, English, Arabic, German, and Hindi, reported by Pell et al. [1]. Chinese pseudo-utterances spoken in seven emotions (anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality) were first identified by a group of native Mandarin speakers in a seven forced choice task, and then subjected to acoustic analyses. Results revealed that among the seven emotions, neutrality, anger, sadness, and fear tended to be recognized most accurately. Acoustic analysis demonstrated the importance of three acoustic parameters (f0 mean, f0 range, and speech rate) in characterizing vocal emotions in Mandarin. Both the perceptual and acoustic characteristics are highly similar, although not identical, to that observed by Pell et al. [1] in English, Arabic, German, and Hindi, indicating a set of universal principles in vocal emotion communication across languages.

Index Terms: Mandarin Chinese, emotional prosody, perceptual-acoustic, cross-language

1. Introduction

Humans can efficiently understand each other's emotions from speech cues without any visual information available, in cases such as telephone conversations. The vocal features of speech, including variations in pitch, loudness, speech rate, etc., which are considered as 'emotional prosody', are a universal means to communicate emotions used by speakers from different language backgrounds [2], [3].

Existing cross-language studies of emotional prosody have reported that listeners can successfully recognize the meaning of vocal emotions in a foreign language, but always with an "out-group" disadvantage (i.e., lower accuracy than that of their native language), suggesting that there is a set of shared properties encoding emotions across languages but on the other hand, discrepancies across languages exist and exposure to vocal emotions in a specific language plays a role (e.g., [4]-[6]). Acoustic studies have also demonstrated a number of consistent tendencies in the acoustic features of vocal emotions across languages (e.g., [1], [2], [7]). Thus, it will be important and interesting to further clarify the extent to which there exists a set of universal perceptual-acoustic characteristics shared across languages in vocal emotion communication, based on direct comparisons of data from different language groups.

However, the previous literature have mostly focused on Indo-European languages, while little is known about how vocal emotions are communicated in other major languages such as Mandarin Chinese. Mandarin is a Sino-Tibetan language which is spoken by more than a billion people around the world and diverges in fundamental ways from Indo-European languages (e.g., [8]-[12]). So far, little work has been done to systematically explore the perceptual-acoustic features of emotional prosody in Mandarin, with direct comparisons with other languages. By employing a well-established database of vocal emotional stimuli in Mandarin [13], this study aims to provide evidence on the perceptual-acoustic features of Chinese vocal emotions; by qualitatively comparing these data directly with four languages from different linguistic families (English, Arabic, German, and Hindi) that were examined in one of our previous studies [1], this study will shed light on to what extent there are central tendencies in how discrete emotions are communicated vocally across different languages and cultures.

To ensure the comparability of our data with that of the previous study, identical procedures of testing and analyses were adopted as those of Pell et al. [1]. In particular, emotionally-inflected "pseudo-sentences" in Mandarin (e.g., *她在一个门文上走亮*) were recognized by native Mandarin speakers in a seven-option forced-choice identification task and then subjected to acoustic analysis. Pseudo-sentences are composed of pseudo content words conjoined by real function words, rendering them meaningless but resembling the phonetic-segmental and supra-segmental properties of the target language; they have been used effectively in the literature investigating the perception of vocal tones independent of the linguistic-semantic content (e.g., [1], [5], [14]). Based on the evidence that many languages share perceptual-acoustic properties of discrete emotions, it is hypothesized that Chinese vocal emotions will show similar perceptual-acoustic patterns as those found in the four languages in comparison [1], although unique variations may also be observed.

2. Method

2.1. Participants

Twenty-four native Mandarin speakers (12 female, 12 male) with a mean age of 25.5 ± 3.3 years were recruited for the perception study. They were all students from China who learned Mandarin from birth, lived in China until at least 18 years of age, had been away from China for less than two years, and spoke English as a second language. Each participant gave written consent for the testing was compensated $10 CAD per hour for their participation.

2.2. Materials and procedure

Eight-hundred and seventy four Chinese pseudo-utterances spoken by 4 native Mandarin speakers (2 male, 2 female) in 7 emotion categories (anger, disgust, fear, sadness, happiness,

pleasant surprise, and neutrality), were adopted from a validated database of Chinese vocal emotional stimuli [13]. In the perceptual study, the 874 utterances were randomly combined and divided into four blocks which were presented by Superlab presentation software (Cedrus, USA) in two testing sessions, two blocks during each session. During the testing, each utterance was played once over headphones at consistent comfortable listening level, for which the participants identified which emotion was being expressed from a list of the seven categories presented on the computer screen by clicking the mouse. All participants received practice trials prior to the first block and frequent breaks during each session. All instructions were conducted in Mandarin.

2.3. Analyses

2.3.1. Perceptual-acoustic analyses

Identical perceptual-acoustic analyses were conducted as those of the study of Pell et al. [1]. For the perceptual data, the recognition rates of each of the 7 emotion categories were calculated. Acoustic analysis was performed on the 874 items and focused on three acoustic parameters that are widely employed in the literature: mean fundamental frequency (f0Mean, in Hertz), fundamental frequency range (f0Range, in Hertz), and speech rate (SpRate, in syllables per second). The values of mean f0, maximum f0, minimum f0, and utterance duration for each item were obtained in Praat [15], based on which the three parameters were calculated. Following Pell et al. [1], in order to correct for differences in a speaker's mean voice pitch and expressive range, all f0 measures (mean f0, maximum f0, and minimum f0) were normalized in relation to the individual *resting frequency* of each speaker (i.e., the average minimum f0 value of all neutral utterances produced by that speaker, see [1] for details). Therefore, for the normalized values of f0Mean and f0Range, a value of 1 for an utterance represents a 100% increase in the speaker's resting frequency, which could be compared across speakers as a proportional value. Measures of speech rate (SpRate) were calculated by dividing the number of syllables of each utterance by the duration of that utterance, in syllables per second.

2.3.2. Statistical analysis

To evaluate whether the seven emotion categories could be differentiated both perceptually and acoustically, univariate and multivariate analysis of variance (ANOVA/MANOVA) were conducted on recognition rates and acoustic measures (f0Mean, f0Range, and SpRate), respectively. In addition, a step-wise discriminant analysis was performed on the acoustic data to explore whether the seven emotion categories could be successfully classified based on the three acoustic measures.

3. Results

3.1. Perceptual data

Emotion recognition rates were calculated for each emotion category as the target emotion hit rate (% correct). A one-way ANOVA performed on recognition rates as a function of emotion category showed a significant effect of emotion category, $F(6, 154) = 26.87$, $p < .01$. Post hoc (Tukey's HSD) comparisons revealed that neutrality (86%) was recognized most accurately, followed by anger (82%), sadness (81%), fear (80%), and happiness (70%); then disgust (67%), which is significantly lower than neutrality, anger, sadness, and fear (86%; *ps* < .01). Pleasant surprise (56%) was recognized less accurately than the other six categories (*ps* < .01). Qualitative comparisons were conducted on recognition rates between Chinese and the four languages studied by Pell et al. [1]; see Figure 1 for an illustration.

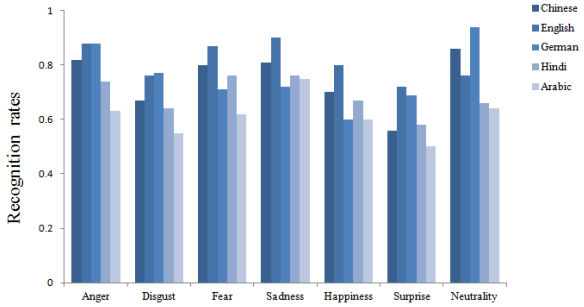


Figure 1: *Recognition rates for each emotion category of each language.*

3.2. Acoustic data

The three acoustic measures (normalized f0Mean, normalized f0Range, and SpRate) were obtained of the 874 items. To explore how the seven emotions differed in these parameters in Mandarin, a one-way MANOVA was performed on the acoustic data as a function of emotion category, with the three acoustic parameters as the dependent variables. The MANOVA indicated that the effect of emotion category on the three acoustic parameters was significant, Wilk's $\Lambda = 0.29$, $F(18, 2447) = 73.54$, $p < 0.01$. Following univariate analyses showed that the effect of emotion category was significant for f0Mean, $F(6, 867) = 106.22$, $p < 0.01$, f0Range, $F(6, 867) = 62.78$, $p < 0.01$, and SpRate, $F(6, 867) = 108.27$, $p < 0.01$. Post hoc comparisons were carried out on each acoustic parameter separately.

For f0Mean, pleasant surprise was expressed with a significantly higher f0Mean compared to all other categories (*ps* < .01), and neutrality exhibited the lowest f0Mean (*ps* < .05). Following surprise, anger and happiness yielded significantly higher f0Mean than fear (*ps* < .01), while the three emotions (surprise, anger, and happiness) showed a higher f0Mean than disgust, sadness, and neutrality (*ps* < .01). For f0Range, fewer significant differences among emotions emerged: pleasant surprise, anger, happiness, and disgust were conveyed with a significantly greater f0Range than fear, neutrality, and sadness (*ps* < .01); in addition, surprise showed a greater f0Range than disgust ($p < .01$). Finally, for SpRate, anger was expressed significantly faster than all other emotions (*ps* < .01), while disgust was produced the slowest (*ps* < .01). After anger, fear and neutrality were spoken significantly faster than happiness and sadness (*ps* < .01), while surprise and happiness also revealed a faster SpRate than sadness (*ps* < .01). Qualitative comparisons were conducted on the three acoustic parameters between Chinese and the four languages [1]. See Figure 2 an illustration.

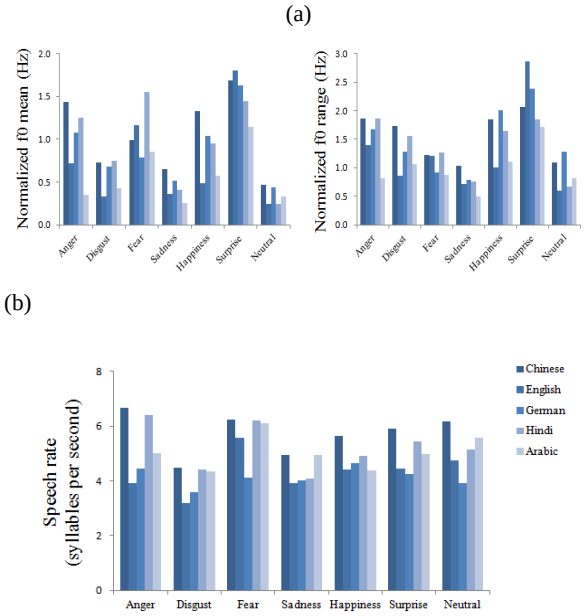


Figure 2: *(a) normalized f0 mean values for each emotion category of each language (left); normalized f0 range for each emotion category of each language (right); (b) speech rates for each emotion category of each language.*

3.3. Discriminant analysis

In order to examine how well the three acoustic parameters predicted the perceptual classification of the seven emotion categories, a discriminant analysis was performed which revealed three significant canonical functions: Function 1, $F(18, 2447) = 73.54$, $p < 0.01$; Function 2, $F(12, 1732) = 99.18$, $p < 0.01$; Function 3, $F(6, 867) = 108.27$, $p < 0.01$. Function 1 explained 55.3% of the variance and correlated significantly with f0Mean ($r = 0.80^{*}$) and SpRate ($r = 0.76^{*}$). Function 2 accounted for 38.4% of the remaining variance. Function 3 accounted for 6.3% of the remaining variance and correlated with f0Range ($r = 0.67^{*}$). This model successfully predicted the classification of the seven emotion categories at an overall rate of 49.9% (436/874), which is similar to those reported by Pell et al. [1] for English (58%), German (49%), Hindi (56%), and Arabic (53%) when the same analysis was performed.

4. Discussion

By conducting a perception study and acoustic analyses of emotional prosody in Mandarin which were identical to the methods of Pell et al. [1], our study allows direct comparison of the perceptual-acoustic features of emotional prosody among different languages; this allows certain inferences to be made about the extent to which vocal emotions display universal tendencies across languages.

The analysis of recognition rates showed expected variations among emotion categories, indicating that certain vocal emotions, e.g., neutrality, anger, sadness, and fear, were recognized more accurately than others in Mandarin, which is compatible with previous data in the four languages in comparison [1] and other languages such as Portuguese, Spanish, and Swedish [2], [7], [16]-[18]. A general advantage to recognize *negative* emotions from vocal speech cues, independent of language, is compatible with evolutionary views of emotion communication that vocal signals associated with threats, such as fear and anger, must be highly salient over long distances and across language systems to ensure human survival [19], [20]. While sadness is often considered a signal of the need for support from conspecifics and is therefore instrumental to maintain cohesion of social groups [21], [22], [23], [24], it is less clear from an evolutionary standpoint why sadness is often recognized most accurately in the vocal channel. Quite possibly, the recognition advantage of vocal sadness is due to their acoustic distinctiveness as speech unfolds, at least for expressions of 'depressed' sadness which lack acoustic variation and energy, as argued recently by Pell and Kotz [29].

Based on qualitative inspection of the perceptual data in Figure 1, it should be noted that fear demonstrated comparatively lower recognition rates in Arabic and German than in the other languages. One potential explanation is that the intensity of fear expressions across studies might be inconsistent; e.g., expressions of fear may be immediate and intense (i.e., 'panic' fear) or more sustained ('dread' fear) and these different forms appear to have distinct vocal cues [2] influencing recognition of this emotion. It seems that in Mandarin of this study and English and Hindi of the previous study, fear is more salient as intense portrayals of 'panic' fear with higher speech rate and f0 mean value.

Disgust and happiness were recognized relatively poorly in Mandarin compared to the other emotions, compatible with the four languages and previous findings (e.g., [2], [5], [7], [16]). For the case of happiness, it has been argued that instead of a unitary category of "happiness", there exist several positive emotions that share the facial expression of smile but possess distinct vocal cues (e.g., achievement/triumph, amusement, contentment, sensual pleasure, and relief), each of which could be reliably recognized through non-verbal vocalizations [25]. Given the fact that the unitary term 'happiness' has been used in the vocal emotion literature and the current study, it is possible that the distinct vocal cues were confounded with each other which led to difficulty in recognizing 'happy' vocal expressions.

In the case of disgust, it is more likely that in natural communication, this emotion is expressed predominantly by facial cues or by non-verbal vocalizations rather than through vocal inflections of the whole utterance (e.g., [7], [14], [26]). Note also that while disgust is considered one of the basic emotions in most studies, there is ongoing debate about whether disgust is either a sensory/interoceptive affect, or a socially constructed moral emotion; this raises the question of whether disgust would be associated with differentiated communicative properties like other basic emotions [27]. Pending further data, our findings do strongly suggest that vocal cues signifying disgust are highly distinct from other emotions in Mandarin and other languages, although the ability to recognize disgust based on prosodic cues tends to be problematic for many listeners [28], [29].

Finally, qualitative inspection showed that pleasant surprise yielded the lowest recognition rates among the seven emotion categories in Mandarin, replicating the observations of the four languages in the study of Pell et al. [1]. Interestingly, surprise was found to be recognized well in several other languages, e.g., Portuguese [7], Swedish [16],

Spanish [30], and Standard Basque [31]. A significant distinction is that Pell et al. [1] and the present study elicited surprise with a positive valence ("*pleasant surprise*"), whereas all other studies elicited vocal cues conveying "surprise" without a specific valence. Compared to surprise, pleasant surprise may be more easily confounded with "happiness" and more difficult to recognize. However, as all these conclusions are based on a small group of speakers and simulated speaking contexts, it is also possible that individual biases in producing certain emotions is contributing to the variation in listeners' performance on emotion identification tasks [32].

Acoustic correlates of specific emotions

Based on qualitative inspection of the acoustic data in Figure 2, the acoustic patterns of several emotions in Mandarin demonstrated a number of consistencies with those in the four languages studied by Pell et al. [1]. Specifically, sadness was conveyed with a low f0Mean, a low f0Range (i.e., reduced variation in f0), and a slow speech rate; disgust exhibited a low f0Mean, moderate f0Range, and the slowest speech rate, whereas pleasant surprise exhibited the highest f0Mean, the largest f0Range, and a moderate speech rate. In addition, neutral speech displayed a relatively low f0Mean, narrow f0Range, and a moderate speech rate (see also [5], [7], [17], [33]-[37]). The fact that many vocal emotions are encoded acoustically in similar ways, irrespective of the language, is consistent with the idea that emotional communication is constrained to a large extent by biological factors and share a set of universal properties across languages [38].

Qualitative comparison across languages also showed greater acoustic variability in certain emotions. For example, exemplars of fear in Mandarin exhibited a moderate f0Mean whereas fear exhibited a much higher f0Mean than other emotions in English, Hindi, and Arabic in the study of Pell et al. [1]. Anger exhibited a high f0Mean whereas f0 values of anger were comparatively lower in Arabic, English, German [1]. As noted earlier, these different acoustic patterns observed of fear and anger across languages are likely to reflect two types of the target emotion. Specifically, for the case of fear, it was suggested that higher f0Mean values characterized "panic fear" while low f0Mean values characterized "sustained fear" (e.g., [2], [5]). Similarly, for anger, higher f0Mean values tend to depict hot anger (i.e., rage or intense frustration), while moderate or low f0 values tend to depict cold anger (i.e., threat; [2], [5], [39]). The possibility that speakers from different cultural-linguistic backgrounds arrive at different interpretations of the labels "fear" and "anger" that influence how they display these emotions vocally could partly explain the acoustic discrepancies observed in f0 values of these two emotions across languages.

Happiness in Mandarin was spoken with a moderate f0Mean, a moderate f0Range, and a moderate speech rate in comparison with the other emotions, whereas this emotion showed a faster speech rate than all other emotions in German [1], and much higher f0 values in Portuguese [7] and Spanish [30]. As discussed earlier, the inconsistent acoustic patterns may be due to the existence of different types of positive emotions rather than a unitary category of happiness, which were associated with different constellations of acoustic features [40]. Therefore, the "happy" expressions in the current study and the literature may actually be a mixture of these different patterns of acoustic cues, which result in discrepant results across languages. Another explanation for the acoustic variability of happiness involves different functions of negative and positive emotions. While the communication of negative emotions such as fear and anger may reflect biologically-driven responses to threat that are signaled to conspecifics in similar ways across language groups, the communication of positive emotions, which facilitate cohesion and affiliation within group, may be largely restricted to in-group members with whom the primary social connections are built and maintained [40]. Thus, as opposed to negative emotions, expressions of positive emotion are influenced by cultural rules and language variables to a larger extent (see [1], [40], [41]), and as such, display greater variability in how they are encoded and decoded across cultures and languages (see [42], [25]). In future studies, it is necessary to disentangle the unitary term "happiness" that has been used in the literature to elaborate the perceptual-acoustic features of distinct types of positive emotions. In addition, cross-cultural/linguistic studies are in need to further clarify the extent to which different types of negative and positive emotions are shaped by culture and language.

Despite focusing on only three acoustic parameters that are critical in emotional communication, a discriminant analysis of the perceptually validated items in Mandarin showed that combined changes in f0Mean and speech rate accounted for approximately 55% of the variance in the acoustic data across emotions. These results are compatible with the established view that a speaker's voice register and articulation rate are essential cues in communicating vocal emotions across languages (e.g., [1], [7], [18]). In total, the three parameters successfully predicted the classification of 49% of items into their perceived emotion categories, which is comparable to that found in other languages (English = 58%, German = 49%, Hindi = 56%, Arabic = 53%, Pell et al., 2009). Nonetheless, as approximately half of the items could not be classified by this small set of acoustic variables, future work will need to include measures of intensity/amplitude, voice quality, and other parameters to fully capture how listeners use acoustic cues to recognize emotion from speech prosody.

5. Conclusion

By conducting a perceptual-acoustic study on vocal emotion expressions in Mandarin Chinese and comparing the data directly with four other languages from a previous study (Pell et al., 2009), this study supplies new evidence of the perceptual-acoustic features of emotional prosody in Mandarin which exhibit many similar, although not identical, patterns to those found in Indo-European and Semitic languages studied by Pell et al. [1] as well as others (e.g., [2], [6], [7], [17], [18]). The new evidence from Mandarin Chinese, which is a Sino-Tibetan ideographic language with little similarity with other languages in the literature, implies that linguistic structure does not impact the core acoustic-perceptual features of vocal emotion expressions in fundamental ways and further strengthens the notion that vocal emotion communication is governed by 'universal' principles across different languages and cultures. Future research using stimuli from spontaneous speech of multiple speakers will further clarify the perceptual-acoustic features of emotional prosody in more natural day-to-day settings. Another interesting topic for future studies is to investigate how the variation of lexical tones of Mandarin influences the acoustic characteristics of emotional prosody in this language.

6. References

[1] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A. and Kotz, S. A., "Factors in the recognition of vocally expressed emotions: A comparison of four languages", Journal of Phonetics, 37(4): 417-435, 2009.

[2] Banse, R. and Scherer, K. R., "Acoustic Profiles in Vocal Emotion Expression", Journal of Personality and Social Psychology, 70(3): 614-636, 1996.

[3] Bollinger, D., "Intonation: Selected readings", England: Penguin Books Ltd, 1972.

[4] Elfenbein, H. A. and Ambady, N., "On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis", Psychological Bulletin, 128(2): 203-235, 2002.

[5] Pell, M. D., Monetta, L., Paulmann, S. and Kotz, S. A., "Recognizing Emotions in a Foreign Language", Journal of Nonverbal Behavior, 33(2): 107-120, 2009.

[6] Scherer, K. R., Banse, R. and Wallbott, H. G., "Emotion inferences from vocal expression correlate across languages and cultures", Journal of Cross-Cultural Psychology, 32(1): 76-92, 2001.

[7] Castro, S. L. and Lima, C. F., "Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudosentences for research on emotional prosody", Behavior Research Methods, 42(1): 74-81, 2010.

[8] Li, K., "The information structure of Mandarin Chinese: Syntax and prosody", Dissertation Abstracts International: The Humanities and Social Sciences, 70(4): 1258, 2009.

[9] Yin, H., "The so-called Chinese VV compounds—A continuum between lexicon and syntax", Proceedings of the 2010 annual conference of the Canadian Linguistic Association, Montreal, Canada, 2010.

[10] Lin, T. H. J. and Liu, C. M. L., "'Again' and 'again': A grammatical analysis of you and zai in Mandarin Chinese", Linguistics, 47(5): 1183-1210, 2009.

[11] Gandour, J., Tong, Y., Talavage, T., Wong, D., Dzemidzic, M., Xu, Y., et al., "Neural basis of first and second language processing of sentence-level linguistic prosody", Human Brain Mapping, 28(2): 94-108, 2007.

[12] Lai, C., Sui, Y. and Yuan, J., "A Corpus Study of the Prosody of Polysyllabic Words in Mandarin Chinese", Proceedings of Speech Prosody, Chicago, USA, 2010.

[13] Liu, P. and Pell, M. D., "Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal stimuli", Behavior Research Methods, 44: 1042-1051, 2012.

[14] Paulmann, S. and Pell, M. D., "Is there an advantage for recognizing multi-modal emotional stimuli?", Motivation and Emotion, 35(2): 192-201, 2011.

[15] Boersma, P. and Weenink, D., "Praat, a system for doing phonetics by computer", Glot International, 5(9/10): 341-345, 2001.

[16] Abelin, Å., "Spanish and Swedish interpretations of Spanish and Swedish emotions – the influence of facial expressions", Proceedings of Fonetik, Stockholm, Sweden, 108–111, 2004.

[17] Abelin Å. and Allwood J., "Cross linguistic interpretation of emotional prosody", Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 110–113, 2000.

[18] Thompson, W. F. and Balkwill, L. L., "Decoding speech prosody in five languages", Semiotica, 158: 407-424, 2006.

[19] Tooby, J. and Cosmides, L., "The past explains the present: Emotional adaptations and the structure of ancestral environments", Ethology and Sociobiology, 11(4-5): 375-424, 1990.

[20] Ohman, A., Flykt, A. and Esteves, F, "Emotion drives attention: Detecting the snake in the grass", Journal of Experimental Psychology: General, 130(3): 466-478, 2001.

[21] Barbee, A. P., Rowatt, T. L. and Cunningham, M. R., "When a friend is in need: Feelings about seeking, giving, and receiving social support", In P.A. Andersen & L.K.Guerrero (Eds.), Handbook of communication and emotion: Research, theory, applications, and contexts, 281-301, San Diego, CA: Academic Press, 1998.

[22] Miceli, M. and Castelfranchi, C., "The Plausibility of Defensive Projection: A Cognitive Analysis", Journal for the Theory of Social Behaviour, 33(3): 279-301, 2003.

[23] Sadoff, R. L., "On the nature of crying and weeping", Psychiatric Quarterly, 40(3): 490-503, 1966.

[24] Sarbin, T. R., "Emotions as narrative employments", in Entering the Circle: Hermeneutic Investigation in Psychology, 185-201, Albany, NY: State University of New York Press, 1989.

[25] Sauter, D. and Scott, S. K., "More than one kind of happiness: Can we recognize vocal expressions of different positive states?", Motivation and Emotion, 31(3): 192-199, 2007.

[26] Scherer, K. R., Banse, R., Wallbott, H. G. and Goldbeck, T., "Vocal cues in emotion encoding and decoding", Motivation and Emotion, 15(2): 123-148, 1991.

[27] Panksepp, J., "Criteria for basic emotions: Is DISGUST a primary "emotion"?" Cognition and Emotion, 21(8): 1819-1828, 2007.

[28] Jaywant, A. and Pell, M. D., "Categorical processing of negative emotions from speech prosody", Speech Communication, 54: 1-10, 2012.

[29] Pell, M. D. and Kotz, S. A., "On the time course of vocal emotion recognition", PLoS ONE, 6(11): e27256, 2011.

[30] Montero, J.M., Gutiérrez-Arriola, J.M., Colás, J., Guarasa, J.M., Enríquez, E. and Pardo, J.M., "Development of an emotional speech synthesiser in Spanish", Proceedings of Eurospeech, Budapest, Hungary, 1999.

[31] Navas, E., Hernáez, I., Castelruiz, A. and Luengo, I., "Obtaining and evaluating an emotional database for prosody modeling in standard basque", Lecture Notes in Artificial Intelligence, 393-400, 2004.

[32] Wallbott, H. G. and Scherer, K. R., "Cues and Channels in Emotion Recognition", Journal of Personality and Social Psychology, 51(4): 690-699, 1986.

[33] Iriondo, I., Guaus, R., Rodriguez, A., Lazaro, P., Montoya, N.,Blanco, J. M. et al., "Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques", Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland, 161-166, 2000.

[34] Pell, M. D., "Influence of emotion and focus location on prosody in matched statements and questions", Journal of the Acoustical Society of America, 109(4): 1668-1680, 2001.

[35] Juslin, P. N. and Laukka, P., "Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?", Psychological Bulletin, 129(5): 770-814, 2003.

[36] Liscombe, J., Venditti, J. and Hirschberg, J., "Classifying Subjective Ratings of Emotional Speech", Proceedings of Eurospeech, Geneva, Switzerland, 2003.

[37] Sobin, C. and Alpert, M., "Emotion in speech: The acoustic attributes of fear, anger, sadness, and joy", Journal of Psycholinguistic Research, 28(4): 347-365, 1999.

[38] Scherer, K. R., "Vocal Affect Expression: A Review and a Model for Future Research", Psychological Bulletin, 99(2): 143-165, 1986.

[39] Scherer, K. R., London, H. and Wolf, J. J., "The voice of confidence: Paralinguistic cues and audience evaluation", Journal of Research in Personality, 7(1): 31-44, 1973.

[40] Sauter, D. A., Eisner, F., Ekman, P. and Scott, S. K., "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations", Proceedings of the National Academy of Sciences of the United States of America, 107(6): 2408-2412, 2010.

[41] Rigoulot, S., Wassiliwizky, E. and Pell, M. D., "Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition", Frontiers in Psychology, 4: 1-14, 2013.

[42] Cornew, L., Carver, L. and Love, T., "There's more to emotion than meets the eye: A processing bias for neutral content in the domain of emotional prosody", Cognition and Emotion, 24(7): 1133-1152, 2010.

Acoustic-prosodic and paralinguistic analyses of "uun" and "unun"

Carlos T. Ishi [1], *Hiroaki Hatano*[1], *Miyako Kiso*[1]

[1] Intelligent Robotics and Communication Labs., ATR, Kyoto, Japan

carlos@atr.jp, hatano.hiroaki@atr.jp, miyakokiso@atr.jp

Abstract

The speaking style of an interjection contains discriminative features on its expressed intention, attitude or emotion. In the present work, we analyzed acoustic-prosodic features and the paralinguistic functions of two variations of the interjection "un", a lengthened pattern "uun" and a repeated pattern "unun", which are often found in Japanese conversational speech. Analysis results indicate that there are differences in the paralinguistic function expressed by "uun" and "unun", as well as different trends on F0 contour types according to the conveyed paralinguistic information.

Index Terms: interjections, acoustic-prosodic features, paralinguistic information, spontaneous conversational speech.

1. Introduction

In spontaneous dialogue speech, repeated "un" utterances ("unun…") or lengthened "un" utterances ("uu…n") are often used as variations of the (short/single) backchannel "un". These two patterns may cause different impressions to the interlocutor depending on the situations they are used. Therefore, in order to achieve a smooth communication between humans and machines, these two patterns should be discriminated.

In our past works, we have focused on interjections appearing in Japanese natural conversational speech, and analyzed the relationship between speaking style and the paralinguistic information (intentions, attitudes and emotions) conveyed by the interjections [1-4]. So far, several monosyllabic interjections (such as "un", "ee", "oo", "ha", "he", "ya") have been analyzed. However, utterances where interjections are repeated in sequence have not been focused.

So far, several works have been conducted regarding the interjection "un" in Japanese (including its variations in speaking styles) [1-8]. However, there are only few studies devoting attention to the functions of "unun" comparing to "un" in discourse. For example, the relations between acoustic features and functions of interjection "un" have been investigated using speech data extracted from TV drama [5]. It is mentioned that "uun" indicates "unknown information is being stored", "embarrassment" or "hesitation". In [6], it has been described that "unun" functions as a marker that the interlocutor's last utterance evoked the speaker's attention for the current conversation topic. In [7], the distribution of backchannel expressions appearing in natural conversations has been analyzed. It has been concluded that "un" is commonly used by both information providers and followers, while "unun" is only used by information followers.

Regarding F0 pattern analysis, the relationship between speaker's attitudes and F0 patterns of "un" have been analyzed in [8]. They found that mean F0 is higher when attitudes of "activation, acceptance, confidence" are expressed, while durational change structures are related with "affirmation/negation" expression. In our past works [1-4], we have shown that the tone type (rising, falling, flat tones) is useful for discriminating between functional speech acts (such as positive reactions, asking for repetition, and thinking), while voice quality features are useful for discriminating emotion expression (such as surprise, admiration, and disgust). However, the speaking styles and paralinguistic functions of "unun" utterances have not been clarified so far.

Regarding backchannel analysis in other languages, it is reported that occurrence rates and environmental location of Japanese backchannels differ from Mandarin and English [9]. It is reported that, in general, appropriate forms and timings of backchannels ("reactive tokens" in their term) show the interest for the speakers and encourage the speaker to keep talking. However, a relatively heavy usage of backchannels in Japanese provides emotional support for speakers [9]. In [10], a large variety of non-lexical tokens, mainly backchannels, uttered in English conversations was examined and the relationship between prosodic features and its meanings was reported. It was stated that duration lengthening (equivalent to "uun" in Japanese) means "amount of thought", while syllabification (equivalent to "unun") means "lack of desire to talk".

Another motivation for the present work is the generation of head motion of robots synchronized with speech utterances [11,12]. Our analysis of head motion during speech utterances has revealed that backchannels are often accompanied by nods, and a sequence of repeated backchannels, such as in "unun" are usually accompanied by multiple nods, approximately one nod per "un" repetition [11]. On the other hand, it was also found that in "uun" utterances where the speaker is thinking, a head tilting is often accompanied. Thus, in the synchronization of head motion with speech, the discrimination of these two types becomes important.

Another issue is that "unun" and "uun" utterances may have similar spectral features, so that acoustic features commonly used in speech recognition, such as MFCC (Mel-Frequency Cepstral Coefficients), would not be enough for distinguish these two patterns. Thus, acoustic analyses are conducted on spectral features and F0 contours.

2. Analysis of "unun" and "uun" utterances

2.1. Speech data

The dataset for analysis was extracted from the ATR multi-modal natural dialogue database. The database contains 65 dialogue sessions of 10 ~ 15 minutes, including 11 male speakers and 14 female speakers with ages from 10s to 60s. The conversation topics are free, including past experiences, future plans, topics about a common known person.

Firstly, a text search was conducted on the transcriptions in the database for collecting "unun" (including two or more repetitions of "un") and "uun" (including two or more repetitions of the vowel lengthening) utterances. In Japanese a vowel lengthening symbol is used for long vowels. As the transcriptions in the database were conducted by multiple

annotators, the transcription criteria might not be unified. Thus, we asked three native speakers (research assistants) to check the consistency of the transcriptions. As a result, 6% of the utterances were corrected, resulting in a total of 342 “unun”-type utterances, and 926 “uun”-type utterances.

The left panel in Fig. 1 shows the distribution of the utterance duration in “unun” and “uun” utterances. Due to a big overlap in the distributions, segmental duration cannot be used to distinguish these two utterance types. In the right panel of Fig. 1, the distribution of the number of “un” repetitions are shown for the “unun” utterances. It can be noted that the most frequent was the pattern with two repetitions (45%), followed by the pattern with three repetitions, i.e. “ununun” (32%), four repetitions (16%) and five or more repetitions (7%).

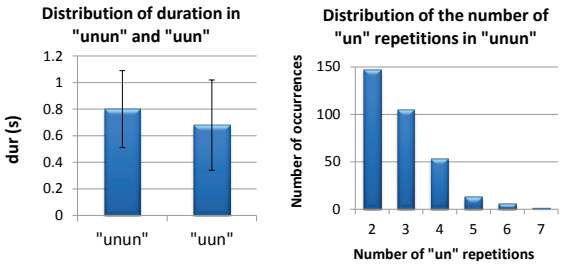


Figure 1: *Distribution of duration in “unun” and “uun” utterances (left panel). Distribution of the number of “un” repetitions in “unun” utterances (right panel).*

2.2. F0 contour type analysis

For each utterance, tone labels were annotated by the first author (which is experienced in prosody annotation), based on F0 curve displays and auditory impression. As tone categories, the following labels were used for monosyllabic utterances:

- “Fa”: falling tones
- “Ft”: flat tones
- “Rs”: rising tones
- “Rt”: pitch reset
- “FtFa”: pitch remains flat and ends with a falling tone.
- “?”: F0 cannot be observed due to vocal fry or low power.

For utterances with two or more syllables, such as in “unun”, the following labels were used:

- “FaFa”: sequence of falling tones
- “Fa_Fa”: sequence of falling tones separated by a short pause

Fig. 2 shows the distributions of different tone types in “unun” and “uun” utterances.

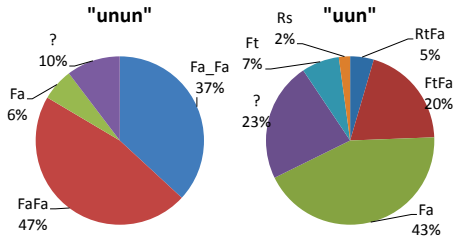


Figure 2: *Distribution of tone types in “unun” and “uun” utterances.*

Flat tones were observed in 7% and rising tones in 2% of the “uun” utterances, while such tone types were not observed in “unun” utterances.

Figs. 3 and 4 show examples of F0 contours and MFCC-smoothed spectrograms of “unun” and “uun” utterances found in our dataset.

Among the “unun” utterances, short pauses smaller than 100 ms between successive “un” syllables were found in 37% of the utterances. In this type, the F0 breaks between successive “un” syllables, being observed as a sequence of falling tones, as the example shown in Fig. 3a.

Most of the “unun” utterances appeared with the pattern without pauses between the “un” syllables (47%). Among them, we observed patterns where the nasal “n” portion is identifiable in the spectrogram and patterns where the (smoothed) spectral features do not change clearly, as the example shown in Fig. 3b. This last pattern, in particular, has little difference to the “uun” utterance spectrogram, so that the only use of spectral features (such as MFCC) would not be enough for their discrimination.

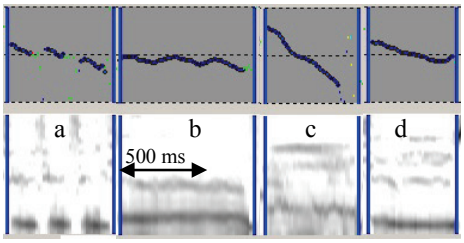


Figure 3: *Examples of F0 contours and MFCC smoothed spectrograms for “unun”. F0 is in log scale from 110Hz ~ 440Hz (center dashed line at 220 Hz); spectrogram is in mel-scale 8kHz band. a) Fa_Fa_Fa ; b) FaFaFa ; c) FaFa? (pitch reset in the second “un” is unclear) ; d) Fa (F0 up-down motion is unclear)*

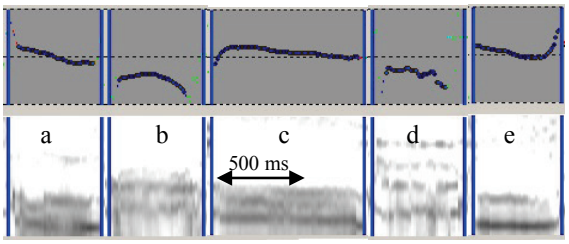


Figure 4: *Examples of F0 contours and MFCC smoothed spectrograms for “uun”. a) Fa; b) FtFa; c) RtFa; d) FtFa with laughing; e) Fa+final rising.*

The phonetic transcriptions of “unun” utterances are usually represented as /uNuN/ (where /N/ is the syllabic nasal). However, in practice, the vocal tract shape does not change much its shape, and only pitch changes upward and downward. This change in pitch is thought to be used in the perceptual distinction of “uun” and “unun” by native speakers.

Another often observed pattern was the one shown in Fig. 3c, where the pitch reset (upward F0 motion) between successive “un” syllables is not clear. Nonetheless, visually one can still observe that the gradient of the F0 contour changes between the “un” syllables.

However, utterances where the F0 resets were almost non-identifiable were also observed in 6% of the "unun" utterances, as the example shown in Fig. 3d. In such cases, the tone label "Fa" was attributed. Although no clear F0 changes can be observed, spectral changes in the /N/ portion can be observed (note that the second formant is broken in Fig. 3d). This means that "unun" utterances are not necessarily accompanied by strong upward-downward F0 changes.

On the other hand, falling tones were observed in most of the "uun" utterances (43% for Fa, and 20% for FtFa). Examples of these patterns are shown in Fig. 4a ~ c. Fig 4d shows an example of "uun" utterance accompanied by laughing, where F0 considerably fluctuates upward and downward.

2.3. Paralinguistic information analysis

Three native speakers annotated the paralinguistic information conveyed by "uun" and "unun" utterances. The paralinguistic items were attributed according to the list below, prepared based on past research on paralinguistic information annotation of interjections. The original terms in Japanese are shown in brackets.

- backchannel ("aiduchi"): "I'm listening."
- affirmation ("koutei"): "Yes, that's right."
- agreement ("dooi"): "Yes, I agree."
- denial ("hitei"): "No, that's wrong"; "No, I disagree."
- negative reaction: dissatisfaction, blame, suspicion ("hiteiteki: fuman, hinan, utagai"): "I'm not satisfied"; "I'm suspicious about"; "I can't accept immediately."
- disgust ("ken-o"): "That's disgusting."
- understanding ("rikai"): "I see", "Yes, I understand."
- admiration ("kanshin"): "I'm admired"; "I'm impressed."
- embarrassment, hesitation ("tomadoi", "chuucho", "konwaku"): "I'm embarrassed/hesitated (on how to react to your utterance)."
- thinking ("kangaechuu"): "I'm preparing my next utterance."
- sympathy, compassion, pity ("kyoukan, doujou, zannen"): "I feel the same"; "It's a pity".

The annotations were conducted by taking contextual information into account, by listening to utterance intervals of both dialogue partner voices, including five seconds before and five seconds after the target interjection part. Annotators were also allowed to include a new item, if they find the items in the list do not fit to the paralinguistic information conveyed. As a result, new labels were included:

- self-affirmation ("jiko-koutei"): affirmation-like interjection right after and directed to the speaker's own utterance.
- modest affirmation ("shoukyokuteki koutei"): affirmation-like interjection, but the speaker does not express it a straightforward manner.

The inter-annotator agreement rates (in terms of kappa values) were 0.59, 0.65 and 0.80 for each pair of annotators.

Fig. 5 shows the distributions of the paralinguistic information items for "unun" and "uun" utterances. A paralinguistic item was attributed to an utterance if two or more annotators agreed. Utterances where the number of utterances for a specific paralinguistic item was smaller than 10, or where agreement was not achieved among the annotators are included in the "others" category. Agreement was achieved in more than 90% of the "unun" utterances, and in more than 75% of the "uun" utterances.

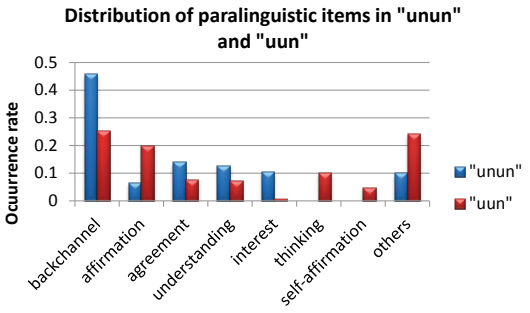


Figure 5: *Distributions of the paralinguistic information items in "unun" and "uun" utterances.*

It can be observed from Fig. 5 that both "unun" and "uun" utterances are used to express "backchannel", "affirmation", "agreement" and "understanding". However, the paralinguistic items "thinking" and "self-affirmation" were observed only in "uun" utterances, while the item "interest" was observed mainly in "unun" utterances.

Analyses were then conducted on the relationship between the tone types and the conveyed paralinguistic information.

Fig. 6 shows the distributions of the paralinguistic information items in "uun" utterances, for different tone types. The occurrence rates are normalized by the total number of utterances in each tone type, which are shown within brackets.

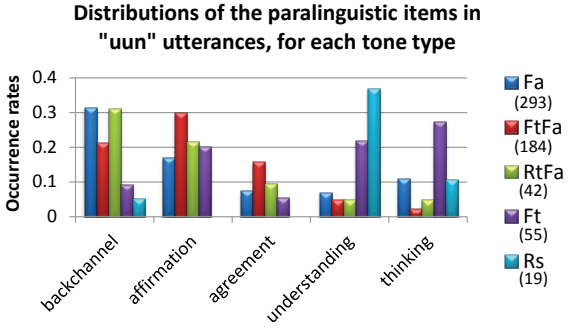


Figure 6: *Distributions of the paralinguistic information items in "uun" utterances, for different tone categories. The occurrence rates are normalized by the total number of utterances in each tone category.*

It can be observed from Fig. 6 that falling tones ("Fa" + "FtFa" + "RtFa") appear with high occurrence rates in backchannel, affirmation and agreement, flat tones ("Ft") are predominant in thinking, and rising tones ("Rs") are predominant in understanding. This is in agreement with the trends reported in past works for monosyllabic "un" utterances.

Regarding the polysyllabic "unun" utterances, no clear differences could be found between tone type and paralinguistic information, since almost all "unun" utterances have a sequence of falling tones as was shown in Section 2.2. Instead, we observed that the number of "un" repetitions could cause different impressions in the dialogue flow.

Fig. 7 shows the distributions of the paralinguistic information items in "unun" utterances, for different "un" repetition numbers (i.e., "2" for "unun", "3" for "ununun", and so on). The occurrence rates are normalized by the total number of utterances in each "un" repetition number category, which are indicated within brackets.

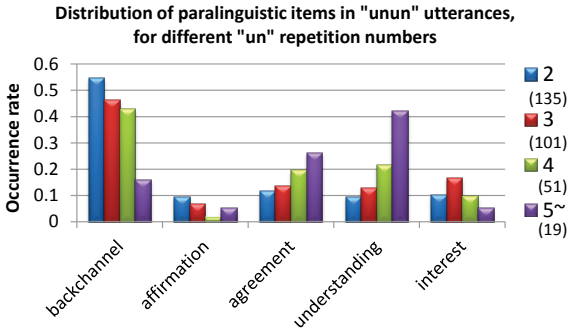


Figure 7: *Distributions of the paralinguistic information items in "unun" utterances, for different "un" repetition numbers (i.e.,"2" for "unun", "3" for "ununun", and so on). The occurrence rates are normalized by the total number of utterances in each "un" repetition number category.*

Among the "unun" utterances, it can be observed from Fig. 7 that utterances with two repetitions ("2") occur with high rates in "backchannels" (more than 50%). The utterances with more than five repetitions ("5~") can be observed with predominant occurrence rates when expressing "understanding" or "agreement". When expressing interest, number of "un" repetitions around three ("3") is found to be predominant.

Finally, from the comments from the annotators, in general, "unun" gives impression of actively pull out the interlocutor utterances by expressing interest, while "uun" gives impression of sharing the feelings of the interlocutor.

3. Discussion: Issues on the discrimination of "uun" and "unun"

For the utterances where pauses are present between successive "un" syllables or where clear spectral changes can be observed in the /N/ portion, spectral features could be used for identification of "unun" utterances. However, for other types, F0 patterns would be useful for discrimination. In this section the problems found in the discrimination of "unun" and "uun" utterances based on F0 patterns are discussed.

The upward-downward F0 motions would be one strong cue for identifying "unun" utterances. When F0 does not change, or when it moves upward and downward only once, the utterance is likely to be perceived as "uun".

However, it was shown in Fig. 4d that F0 can also move upward and downward in "uun" utterances accompanied by laughing. It would be difficult to discriminate such utterances from "unun" utterances by only using F0 range information. Further, discrimination of "unun" utterances accompanied by laughing would also be more difficult. Thus the dynamic features of F0 contours should be modeled for discriminating these features.

F0 rising in the end portion of the utterances was observed in 10% of the utterances, in 7 speakers (6 female and 1 male). Fig. 4e showed an example of "uun" utterance where the F0 rises in the end portion. This F0 rising was also observed (with less frequency) in the end portion of "unun" utterances (as in Fig. 3d), and is thought to be unconsciously produced when the vocal folds stop vibrating. As this F0 rising is not well perceived, it should be removed for tone analysis.

The "?" label was annotated in 23% of the "uun" utterances and in 10% of the "unun" utterances. Most of them were either due to low power in bad recording conditions or due to presence of vocal fry, so that F0 contours could not be obtained.

Regarding the recording conditions, in part of the database headset microphones are available, while in the other part only directional microphones on the table are available. For the table microphones, the distance to the mouth is 30 ~ 40 cm on average, so that both background air conditioner noise and the dialogue partner interference sound are strongly observed.

The problem is that "un" utterances have lower power even within the utterances of the same speaker, and their backchannel functions in dialogue make them highly probable of being overlapped with the interlocutor's utterances. Thus, the SNRs tend to become very low in "un" utterances if the microphone is not positioned close to the speaker's mouth, affecting both speech recognition and F0 extraction.

The other problem is the presence of vocal fry (or creaky phonation), where the vocal fold vibrations become irregular, so that the measured F0 would not correspond to perceived pitch contours. Vocal fry was observed mainly in 4 male subjects, where robust F0 contours could not be obtained. In such cases, the presence of short pauses between "un" utterances become more important for identification of "unun" utterances, rather than the F0 information.

4. Conclusions

We conducted analyses on acoustic-prosodic features and paralinguistic functions of "unun" and "uun" utterances, which are repeated and lengthened variations of the interjection "un", commonly appearing in conversational speech.

Analysis results indicated that both "unun" and "uun" appear in the expression of backchannels, affirmation, agreement and understanding, while "unun" appears more frequently when expressing interest, but does not appear for expressing thinking or self-affirmation, in contrast with "uun". Further, in "unun" utterances, the number of repetitions of "un" tends to be higher when expressing agreement or understanding.

Regarding the patterns of F0 contours in "unun" and "uun" utterances, it was shown that part could be discriminated by the up-down F0 movements or by spectral changes in the nasal part. However, it was also shown that utterances accompanied by laugh, final rising and vocal fry can be problematic. Future work includes evaluation of discrimination of "unun" and "uun" based on acoustic features.

5. Acknowledgements

This work was partly supported by the Ministry of Education, Culture, Sports and (MEXT Kakenhi) and Japan Science and Technology Corporation (JST). We thank Mika Morita and Kyoko Nakanishi for helping in the data analysis.

6. References

[1] Ishi, C.T., Ishiguro, H., Hagita, N., "Automatic extraction of paralinguistic information using prosodic features related to F0," duration and voice quality. Speech Communication 50(6), 531-543, June 2008.

[2] Ishi, C.T., Ishiguro, H., and Hagita, N, "The meanings of interjections in spontaneous speech," Proc. Interspeech' 2008, 1208-1211, 2008.

[3] Ishi, C.T., Ishiguro, H., and Hagita, N., "Analysis of acoustic-prosodic features related to paralinguistic information carried by interjections in dialogue speech," Proc. Interspeech' 2011, 3133-3136, 2011.

[4] Ishi, C.T., Hatano, H., Hagita, N., "Extraction of paralinguistic information carried by mono-syllabic interjections in Japanese," Proceedings of The 6th International Conference on Speech Prosody (Speech Prosody 2012), 681-684, 2012.

[5] Sudo, J., "The Japanese interjection un: From its meanings and functions to an analysis of its phonetic features", Journal of the Phonetic Society of Japan, Vol.11 No.3, 94-106, 2007 (in Japanese)

[6] Togashi, J., "Aizuchi hyougen keishiki-ni miru shinnai-no jouhou syori-ni tsuite", Working papers for special project of Tsukuba university "touzai gengo bunka-no ruikeiron", 27-42, 2002 (in Japanese).

[7] Yoshida, E., "Detecting patterns of sequences by coding scheme and transcribed utterance information: An analysis of Japanese reactive tokens as non-primary speaker's role", Proceedings of The 3rd workshop of Japanese corpus, 435-440, 2013 (in Japanese).

[8] Kokenawa, Y., Tsuzaki, M., Kato, H. and Sagisaka, Y., "An analysis of speaking attitude manifesting as fundamental frequency characteristics", Technical report of IPSJ SIG, 87-92, 2004 (in Japanese).

[9] Clancy, P. M., Thompson, S. A., Suzuki, R. and Tao, H., "The conversational use of reactive tokens in English, Japanese, and Mandarin", Journal of Pragmatics, 26, 355-387, 1996.

[10] Ward, N., "Non-lexical conversational sounds in American English", Pragmatics and Cognition, 14, 113-184, 2006.

[11] Ishi, C.T., Liu, C., Ishiguro, H., and Hagita, N. (2010). "Head motion during dialogue speech and nod timing control in humanoid robots," Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), 293-300.

[12] Liu, C., Ishi, C., Ishiguro, H., Hagita, N. (2013). Generation of nodding, head tilting and gazing for human-robot speech interaction", International Journal of Humanoid Robotics (IJHR), vol. 10, no. 1, January, 2013.

Speaking style prosodic variation: an 8-hour 9-style corpus study

Jean-Philippe Goldman[1], Tea Pršir[1,2], George Christodoulides[2], Antoine Auchlin[1]

[1]Département de Linguistique, Université de Genève
[2]Institut Langage & Communication, Université de Louvain

jean-philippe.goldman@unige.ch, tea.prsir@unige.ch,
george@mycontent.gr, antoine.auchlin@unige.ch

Abstract

This paper presents the results of a prosodic and phonostylistic analysis based on C-PhonoGenre, an 8-hour-long spoken French corpus, consisting of 9 speaking situations and (on average) 10 speakers per situation. The corpus was automatically segmented at the phonetic, syllabic and word levels (EasyAlign), and in larger pause-separated units. Part-of-speech annotation (DisMo) and prominent syllable detection (ProsoProm) was added automatically. The corpus was also manually annotated at the syllabic level for stylistic variants, such as post-tonic schwas, liaisons, elisions, disfluencies, audible breaths and noises. Acoustic analyses (ProsoReport, DurationAnalyser) provide more than 100 micro- and macro-prosodic measures, which we correlate with the phonostylistic features and the linguistic annotation. This analysis results in a contrastive, fine-grained *prosometric* description of phono-stylistic and situational variation, over 4 situational, gradual dimensions: audience, media, preparation, and interactivity. Further statistical analysis was carried out to explore the discriminative and explanatory power of combinations of prosodic measures.

Index Terms: situational variation, prosody, classification of speaking styles

1. Introduction

General knowledge of language includes that of its variants, as demonstrated by the study of *genres* [1, 2, 3], and phono-stylistic variation is such an area [4, 5, 6]. Recent research in prosody focuses on phonostylistic situational variation in large corpora, departing from previous binary oppositions (*e.g.* read vs. spontaneous speech), or one-dimensional characterisations of style (formal vs. informal).

This paper presents a selection of global and contrasted results of an on-going research project on situation-dependent speaking styles, or *phonogenres*. It applies the semi-automated methodology of corpus description introduced in previous work [7, 8]. It analyses speaking situations by features [9, 10, 11], reduced to four main dimensions: audience, media, preparation, and interactivity; each dimension has 3 different states (see Table 1). For example, ***audience** = 1* indicates that the speaker is physically present before an audience, while ***media** = 1* indicates speech directed to an individual or a small group, yet in front of a microphone or camera (*indirect* audience). ***Preparation** = 1* indicates semi-prepared speech, situated between spontaneous and read speech. In the case of parliamentary debates, a *question* is prepared, while the *answer* is semi-prepared. ***Interactivity** = 1* indicates that the main speaker may be interrupted. For example we distinguish between dialogue interaction and broadcast sports commentaries. The corpus under study, C-PhonoGenre, is approximately 8 hours-long, and covers 9 different genres, four of which are further subdivided into *sub-genres* (based on their situational specificities).

Results show that phonogenres and sub-genres can be distinguished and characterised by the relation between situational and prosodic dimensions. Various "hidden" or unpredicted influences of situational properties on prosody also emerge. An indirect, but equally important result of this study is a corpus processing methodology, based on the co-ordinated application of several semi-automatic tools.

2. Data

C-PhonoGenre contains data from 8 speaking styles: instructional speech [DIDA]; spontaneous narration [NARR]; speeches during "Question Time" at the French parliament [PARL]; sermons [RELG]; radio press reviews [RPRW]; three kinds of sports commentary [SPOR]: rugby, basketball and football; presidential New Year's wishes [WISH] and weather forecasts [WFOR]. The average sample duration per speaker is 5:30 min.

Table 1. *Situational features by PhonoGenre*

PhonoGenre		Audience	Media	Preparation	Interaction
DIDA	Radio	1	2	2	2
	TV	0	2	2	0
	Lecture	2	0	1	0
NARR	Narration	1	0	0	2
PARL	Question	2	1	2	1
	Answer	2	1	1	1
READ	Reading	0	0	2	0
RELG	Internet mass	0	1	2	0
	Sermon on TV	2	1	2	0
RPRW	Radio press review	0	2	2	0
SPOR	Basket	0	2	0	0
	Rugby/football	1	2	0	2
WFOR	Weather forecast	0	2	2	0
WISH	Pres. New Year	0	1	2	0

For this study, we compiled a corpus including the eight speaking styles of C-PhonoGenre and the "reading" style [READ] from C-PROM-PFC [12]. Table 2 shows number of samples, syllables, words and total speech time per genre. Although [SPOR] and [RELG] contain less than 10 speakers per genre, the total amount of data renders these genres comparable with the others. On the other hand, while 10 different speakers are included under the "weather forecasts" genre [WFOR], the total speech time is 9 minutes, i.e. less than a minute per speaker.

The corpus contains recording of both female and male speakers, originating from 3 different French-speaking areas: Metropolitan France, Belgium and Switzerland [13]. We do not present findings regarding regional variation here, but the information is present in the corpus metadata and can be used for further study. Regional variation may partly explain the observed intra-genre, inter-speaker dispersion.

Table 2. *Number of recordings, syllables and words, and duration by phonogenre*

PhonoGenre	Num. samples	Duration (min)	Num. syllables	Num. words
DIDA	17	100	26 304	18 717
NARR	10	44	11 396	9 546
PARL	10	20	5 710	3 613
READ	16	36	9 932	6 648
RELG	7	54	8 726	6 141
RPRW	15	95	26 359	17 531
SPOR	5	35	7 601	5 305
WFOR	10	9	2 861	1 947
WISH	15	98	18 614	12 578
TOTAL	**105**	**491**	**117 503**	**82 026**

3. Methodology

3.1. Data processing

After manual orthographic transcription in Praat [14], we obtained a phonetic transcription as well as a segmentation of words, syllables, phones and pauses, automatically using EasyAlign [15]. Manual corrections were made to reach a high quality alignment between the segments and the speech signal. A unique annotator manually added a <*delivery*> tier in order to enhance downstream data processing. It contains four types of annotation: i) disfluencies, articulation and phonological phenomena: schwa; vowel lengthening (whether associated to hesitation or not); creaky voice; liaison and elision; ii) symbols to distinguish between complete silence, audible and less audible breaths, and mouth noises; iii) indices of paralinguistic phenomena (laugh, cough) and external sounds; iv) overlapping segments and syntactic plan interruptions. Part-of-speech tagging and multi-word unit detection was obtained automatically using DisMo [16]; this annotation is used to study the interface between speaking styles and grammar. A five-level degree of prominence for each syllable was calculated using ProsoProm [17]. Three additional tiers were automatically added: <*lex*> distinguishing between lexical and functional words; <*if*> localising initial vs. final lexical words' syllables; and <*ap*> which is an automatically generated segmentation into accentual phrases (in Mertens' sense [18], i.e. phonological words), including an annotation of the initial and final syllables of each accentual phrase. Pitch was corrected manually for the entire corpus, since the accuracy of several acoustic measures and prominence detection depend on it.

3.2. Acoustic measures

Acoustic and prosodic features were extracted for the entire corpus. Initially, ProsoGram's [18] two-step algorithm for pitch stylisation was applied: for each syllable, vocalic nuclei are detected based on intensity and voicing, and then the F0 curve on the nucleus is stylised into a static or dynamic tone, based on a perceptual glissando approach. ProsoReport [8] summarises this information, taking into account information contained in other tiers (such as <*delivery*> and <*lex*>) to produce a detailed collection of descriptive statistical measures for each corpus sample. These measures can be grouped into four main families: temporal measures (e.g. articulation rate); pitch measures (e.g. pitch register and movement); syllabic prominence measures (e.g. percentage of prominent syllables in various positions); correlational measures (e.g. percentage of accentual-group-initial prominent syllables). Additionally, DurationAnalyser [19] produced a set of statistics based on segmental information (e.g. variance coefficients for vowels or consonants, nPVI etc.). All this information was compiled for further analysis; in total, 129 prosodic descriptors for each corpus sample were retained.

4. Results

Based on the aforementioned prosodic descriptors, many results appear to be significant. We show only some of them based on the *genre* classification, then on different prosodic domains (duration, intonation and accentuation of initial and final syllables, across genres and situational features).

4.1. Results by phonogenre

The articulation ratio at the genre level sets apart WISH and RELG, reflecting that both situations are solemn. SPOR also stands out, but for another reason: the commentator has to pause while the ball moves from one player to the next. Conversely, WFOR, and to a lesser extent RPRW, show the highest articulation ratio, because of the time pressure imposed by broadcasting media.

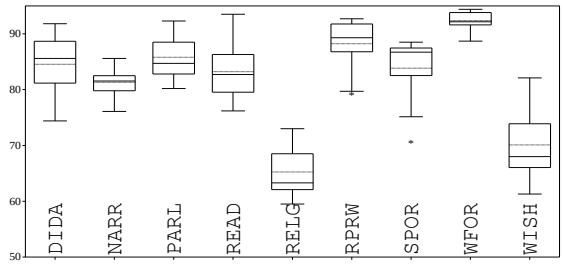


Figure 1: *Articulation ratio for the 9 phonogenres*

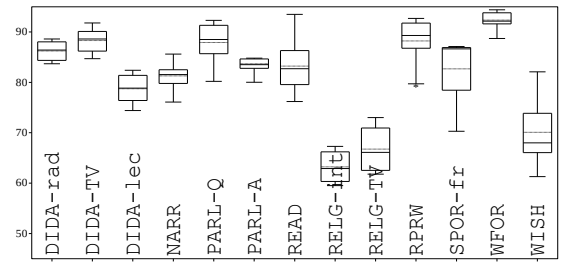


Figure 2: *Articulation ratio for the 13 sub-genres*

At the sub-genre level, articulation ratio also illustrates interesting contrasts. Parliamentary answers (PARL-A) clearly occupy more speech time than questions, because the listeners react during the answer and the speaker has to take into account this feedback. The three instructional sub-genres also show differences, mainly between the Radio and TV sub-genres. A possible explanation of this difference may be that radio samples are shorter than TV samples (DIDA-rad: approx. 3 min., DIDA-TV: 10-20 min.), as well as the fact that TV delivers images with speech, sharing time with visual flow [20]. The DIDA-lec subgenre (non-broadcast university lectures) is closer to DIDA-rad, suggesting that the media dimension is less important than the instructional one. RELG subgenres differ slightly on articulation ratio, though other prosodic measurements distinguish them much more clearly.

4.2. Segmental duration

Among the acoustic parameters based on segmental durations, the variance of vowel duration is the one exhibiting the best

discriminative power. Figure 3 is a box-plot of vowel duration variance for different sub-genres. The NARR genre detaches from others probably because of its spontaneous nature: at the syllable level, this results in frequent hesitation-related lengthening; at the discourse level, an irregular speech rate is entailed by the progressive construction of discourse. In contrast, the READ and WFOR genres have a lower variation of vowel duration, but for different reasons. READ readers are non-professionals and thus adopt a monotonous rhythm; whereas the WFOR genre has a very high speaking rate, causing a ceiling effect on vowel duration. Interesting differences occur again between in parliamentary sub-genres, showing a significantly lower variation for the answer [PARL-A] than for the question [PARL-Q], due to the increased interactivity (see 4.1).

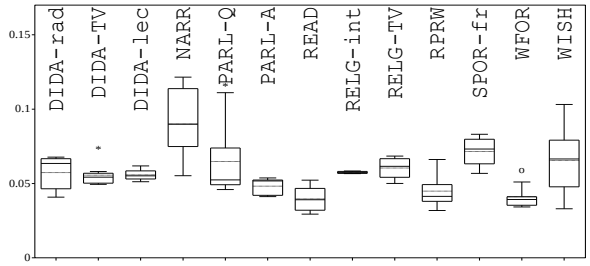


Figure 3: *Variation of vowel duration for the 13 sub-genres*

The *preparation* feature also shows a lower variation of vowel duration for prepared recordings (F(2,102)=50; p<0.001). This is explained by more lengthened hesitations in spontaneous speech. The *interactivity* feature shows similarly a greater variation (F(2,102)=31.4 p<0.001) for interactive recordings, usually the spontaneous ones (NARR, SPOR).

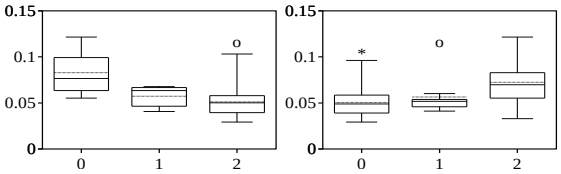


Figure 4: *Variation of vowel duration for 3 levels of preparation (left) and interactivity (right*)

4.3. Intonation

Intonational properties indicate a lower relative F0 variation for phonogenres with a larger audience (F(2,102)=10.5; p<0.001); this is surprising, as we hypothesised that public speaking would entail greater speaker involvement. However, this acoustic parameter varies according to our predictions across the media feature (F(2,102)=12.06; p<0.001).

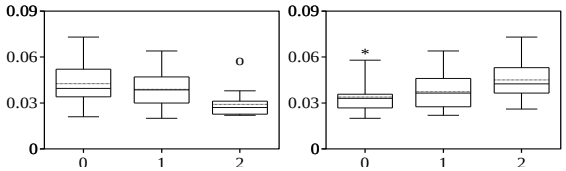


Figure 5: *Relative F0 variation for 3 degrees of situational features of audience (left) and media (right)*

4.4. Prominence in initial and final position

The study of initial and final positions of prominent syllables results in differentiating phonogenres based on their situational features.

4.4.1. Situational features

The percentage of prominent final syllables is decreasing as the phonogenre is getting more *interactive* (F(2,102)=8.88; p<0.001). This can be explained by a high score of hesitation in NARR and vowel lengthening, typical for sport commentaries SPOR (Figure 6, left).

The percentage of prominent initial syllables is getting higher if a phonogenre falls is *media*, where it is important to clearly distinguish discourse segments (Fig.6, right). The initial prominent syllables of AP shows similar results (F(2,102)=5.88; p<0.001).

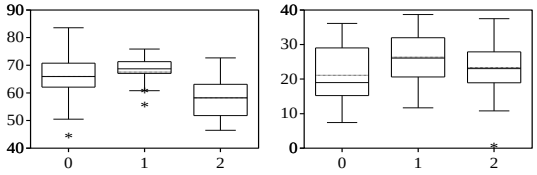


Figure 6: Percentage of prominent final syllables for interactivity (left) and percentage of prominent initial syllables for media (right) for each of the 3 degrees

The relative length of initial and final syllables of the AP varies in a significant manner across the *preparation* dimension (initial syllables F(2,102)=5.42 p<0.001; final F(2,102)=10.65 p<0.001). The variation is inverted: initial syllables of AP tend to be shorter in prepared discourse than in non-prepared, but final syllables become longer (Figure 7).

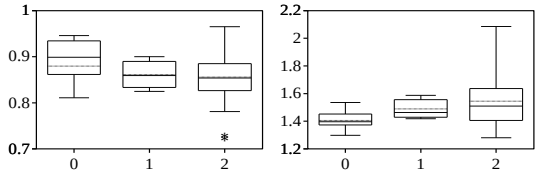


Figure 7: *Relative length of initial (left) and final (right) syllables of the AP for preparation*

The physical presence of *audience* implies lower percentage of initial prominent syllables per AP. This is the case for parliamentary speeches where one member of the house is talking directly to the government minister [PARL]; for university lectures, where teacher addresses the students [DIDA-lec], or sermons where the priest is addressing the faithful [RELG-TV] (F(2,102)=12.8; p<0.001).

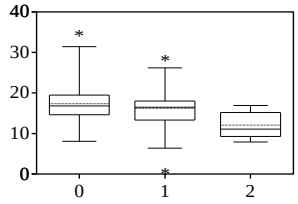


Figure 8: *Percentage of initial prominent syllables per AP for audience for each of the 3 degrees*

4.4.2. Sub-genres

Sub-genres level distribution of initial prominences (Fig. 9) partly reflects the values of situational features. Sub-genres in which an audience is present show the lowest level (PARL; DIDA-lec, RELG-TV, SPOR), whereas media ones (DIDA-rad, DIDA-TV, WFOR) show the highest level. RELG-int behaves like a media style, although it was graded as an intermediate style across the media dimension. Despite RPRW being a prototypical case of broadcast media style, it shows a low level of initial prominence. In fact, RPRW shows a high rate of prominent initial syllables of words, not of AP, which reflects the stylistic choice of marking initial syllables.

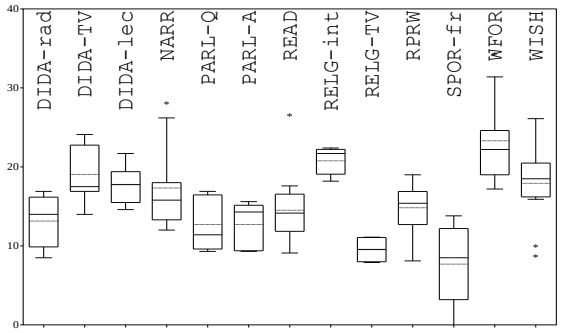


Figure 9: The distribution of the percentage of initial prominent syllables per AP for the 13 sub-genres

4.5. Principal Component Analysis

Since no unique prosodic/acoustic parameter is sufficient to clearly distinguish speaking styles, we explored a global statistical approach which finds the optimal linear combination of all parameters. A Principal Components Analysis (PCA) was thus performed, to model phonogenres and situational features with the parameters, knowing that some of them are high correlated. The first two principal components (PC) explain only 43% of the variance, while the first 8 explain 78.2%. A discriminating analysis with 8 PCs over 9 phonogenres showed that 93.3% of recordings were identified as belonging to the correct genre. Table 3 shows the percentage of correct classification decisions, over genre, sub-genre and the 4 situational features.

Table 3. *Percentage of correct classification with 8 principal components for Genre and Sub-Genre distinction as well the 4 situational features*

Genre	93.3 %
Sub-Genre	90.5 %
Audience	90 %
Media	84.8 %
Preparation	92.4 %
Interactivity	92.3 %

Figure 10 shows the distribution of the 105 speech recordings along the first two PCs. The first PC appears to be highly correlated with the articulation ratio, as the WFOR genre is clearly opposed to the genres WISH and RELG. The bottom of the figure shows the confidence ellipses for the 'media' situational feature according to the first 2 PCs with a clear distinction between semi-media (1) and media (2) condition, while the non-media (0) condition is over the first two conditions.

5. Conclusion

Our study led to two concrete outcomes: (a) a detailed methodology for future studies in *phonostylistics* and socio-phonetics, which is primarily based on automated tools and the study of prosody (an extension to other levels, e.g. the lexical level, is envisaged); and (b) a spoken corpus of high scientific value, due to its thorough multi-level annotation.

The results show that, while no single prosodic measure is sufficient to separate and classify speaking styles, a linear combination of several measures leads to a robust clustering of samples belonging to different genres. We next will explore feature selection techniques to determine which set of prosodic measures is the most relevant (instead of a PCA analysis that produces a global, but opaque, combination of measures).

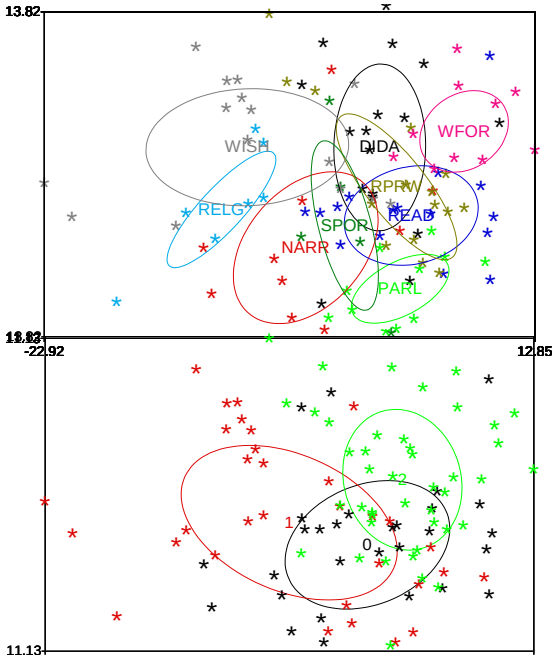


Figure 10: The 105 recordings according to the first two PC and confidence ellipses by phonogenres (top); the confidence ellipses for the media situational feature (bottom)

The study has also shown that following an iterative, adaptive procedure is necessary: while the initial, top-down approach was to select samples in order to create a balanced corpus (based on a predefined array of situational features), subsequent data analysis led to the observation that samples within a given genre could and should be further classified in *sub-genres*. Therefore, the interplay of prosodic measures and situational features gave rise to an *a posteriori* subdivision of genres (bottom-up approach) in order to ensure compact definitions and to reduce the excessive heterogeneity of some speaking situations. Results show (Fig. 1, 2, 3) that sub-genre groupings *transcend genre differences* (e.g. PARL-A and DIDA-rad, in Fig. 2), and that some of them are related to common, controlled situational features (cf. Fig. 9, DIDA-lec: non-media, public audience). They also present evidence for groupings due to unpredicted, or hidden, situational features, like "external time pressure" (cf. Fig. 1 PARL, WFOR, RPRW), "speech sequence duration", or "solemnity / ritual conventions" (RELG and WISH), that belong to the prototypical image of speaking style. The differences observed between questions and answers within the PARL genre suggest a situational feature [+ interactivity] at the sub-genre level. They reveal a prosodic reflection of a discursive (not situational) feature, namely the presence of other listeners that react to the exchange even though they do not participate in it directly. The results also indicate that several different speakers per genre must be included in the corpus for the genre-specific features to rise above individual variation.

Although some annotation steps remain manual, most of our methodology is automatic. This framework was built in a very generic way: we plan to provide additional prosodic measures and test corpora in other languages. Such research enables both verification of linguistic hypotheses and automatic genre identification. Finally, we should mention that the C-PhonoGenre corpus will be made available to the community for any other research purposes.

6. Acknowledgements

This research is funded by Swiss National Science Foundation – FNS Grant nr 100012_134818.

7. References

[1] Beacco, J.-C. "Trois perspectives linguistiques sur la notion de genre discursif", Langages 38(153): 109–119, 2004.

[2] Solin, A. "Genre", in J. Zienkowski, J.-O. Östman and J. Verschueren [Eds], Discursive Pragmatics, 119–134, John Benjamins, 2011.

[3] Bawarshi, A. and Reiff, M. J. Genre: An Introduction to History, Theory, Research, and Pedagogy, Indiana, Parlor Press, 2010.

[4] Fónagy, I. and Fónagy J. "Prosodie professionnelle et changements prosodiques", Le Français Moderne 44: 193–228, 1976.

[5] Fónagy, I. La vive voix. Précis de psycho-phonétique, Paris, Payot, 1983.

[6] Léon, P. Précis de phonostylistique, Parole et expressivité, Nathan Université, Paris, 1993.

[7] Simon, A.C., Auchlin, A., Avanzi, M. and Goldman, J.-Ph., "Les phonostyles: une description prosodique des styles de parole en français", in M. Abecassis and G. Ledegen [Eds], Les voix des Français. En parlant, en écrivant, 71–88, Peter Lang, Berne, 2010.

[8] Goldman, J.-Ph., Auchlin, A. and Simon A.C., "Discrimination de styles de parole par analyse prosodique semi-automatique", in H.-Y. Yoo and E. Delais-Roussarie [Eds] Actes d'IDP 2009, Septembre 2009, Paris, 2011.

[9] Lucci, V., "Étude phonétique du français contemporain à travers la variation situationnelle", Université des langues et lettres, Grenoble, 1983.

[10] Koch, P. and Oesterreicher, W., "Langage parlé et langage écrit", in G. Holtus, M. Metzeltin and Ch. Schmitt [Eds], Lexikon der Romanistischen Linguistik, I/2, 584-627, Niemeyer, Tübingen, 2001.

[11] Pršir, T., Goldman, J.-Ph. and Auchlin A., "Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français", in P. Mertens and A.C. Simon [Eds], Proceedings of the Prosody-Discourse Interface Conference 2013, 107-111, Leuven, September 2013.

[12] Avanzi, M., Christodoulides, G., Schwab S., Bardiaux A., Goldman J.-Ph., "La variation prosodique régionale et stylistique en français – Analyse de neuf points d'enquête PFC", Journées PFC, Paris, December 2013.

[13] Simon, A. C. [Ed], La variation prosodique régionale en français, DeBoeck, 2012.

[14] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", online at http://www.praat.org

[15] Goldman, J.-Ph. EasyAlign: an automatic phonetic alignment tool under Praat, Proceedings of InterSpeech, Florence, Italy, 2011.

[16] Christodoulides, G., Avanzi, M. and Goldman, J.-Ph., "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator: An Evaluation on a Corpus of French Spontaneous and Read Speech", Proceedings of the Language Resources and Evaluation Conference (LREC) conference, Reykjavik, Iceland, 26-31 May 2014.

[17] Goldman, J.-Ph., Avanzi, M., Simon, A.C. and Auchlin, A., "A continuous prominence score based on acoustic features", Proceedings of InterSpeech 2012, 9-13 September, 2012.

[18] Mertens, P., "The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model" in B. Bel and I. Marlien [Eds], Proceedings of Speech Prosody 2004, Nara, Japan, 23-26 March, 2004.

[19] Dellwo, V., Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence, PhD Dissertation, Universität Bonn, 2010.

[20] Kern, F., "Speaking dramatically. The prosody of life radio commentary of football matches", in Barth-Weingarten, D., Reber E., and Selting M. [Eds] Prosody in interaction, John Benjamins, 217-237, 2010.

Certainty and uncertainty in Brazilian Portuguese: methodology of spontaneous corpus collection and data analysis

Leandra Batista Antunes[1, 2], *Véronique Aubergé*[2], *Yuko Sasa*[2]

[1] University of Ouro Preto, Brazil
[2] LIG (Laboratory of Informatics of Grenoble), CNRS, Grenoble University, France

antunes.leandra@yahoo.com.br; {veronique.auberge; yuko.sasa}@imag.fr

Abstract

This work presents a methodology used to collect some spontaneous social affect corpus and preliminary prosodic analysis of certainty and uncertainty in Brazilian Portuguese. The corpus was collected using a Wizard of Oz method (through EmOz platform). The scenario to induce certainty and uncertainty is based on the situation of a job interview, where a companion robot (Emox) is used as a trainer. The subjects were convinced to benefit from a free teaching of this "revolutionary" method to train to job interview. In this scenario the linguistic expressions are partially controlled, in order to focus the certainty/uncertainty expression mainly on paraphrasing and prosody. Data were preliminarily analyzed for audiovisual prosody: videos analysis were made regarding eyebrows, eyes, mouth and face/head movements, while audio analysis were made about acoustic prosody parameters of fundamental frequency and duration. The first results show that using EmOz within such a scenario is an efficient way to induct spontaneous but comparable speech production. Prosodic results show that fundamental frequency and duration measurements, as well as eyebrows, eyes, mouth and face/head movements are differently used in certainty and in uncertainty production in Brazilian Portuguese.

Index Terms: spontaneous attitudinal corpus acquisition, audiovisual prosody of (un)certainty, Brazilian Portuguese.

1. Introduction & theoretical background

As early noted by [1], one fundamental prosody's function is to express the speaker's affective states, such as emotions, attitudes or moods. More generally, in face to face communication the socio-affective states (intentions, attitudes, mental states) are the main cues of dialog efficiency. Here, the socio-affective prosodic expressions are assumed to be voluntary controlled and a part of the linguistic system, according to [2, 3, 4], differing in this way from emotions. In this paper we focused on two attitudes particularly relevant in dialog: *certainty* and *uncertainty*, specifically in Brazilian Portuguese (henceforth BP).

Since 15 years, many prosodic studies of BP have been focusing on the role of the prosody to express the speaker's attitudes or intentions [5, 6, 7, 8, 9]. Among the investigated attitudes in these prosodic studies we can find certainty and uncertainty, which were already studied in other languages.

Many studies held on several languages showed some similarities and differences between languages (confidence vs. doubt for French, English, Japanese, Vietnamese and Chinese [10]). More elicitated expressions were produced in some Dutch language studies [11, 12, 13, 14]: the audiovisual prosody was investigated to indicate the uncertainty degree in answers of factual questions. In these studies, the methodology for data acquisition was based on the answer to factual questions, as done in [15, 16]. The researchers (who do not show themselves to the participants at this moment) asked 40 questions to the subjects (students). Participants were recorded in audio and video (with a camera taking the participant's face). After this, the participants were asked if they would be able to recognize the answer to the same questions if they were in a multiple choice test. Subsequently, the same subjects answered the same questions in a multiple choice test. The authors' intention was to describe, through audiovisual prosody, the degree of uncertainty in each answer, measuring the Feeling of Knowing (FOK), defined as "people's ability to assess and monitor their own knowledge" [13].

Audiovisual prosodic differences between certainty and uncertainty in these studies include: i) concerning the verbal cues: the final boundary tone, where the High tone is associated to uncertainty, while the Low one is associated to certainty; the delay or the time to start the answer, longer times being linked to uncertainty; the presence of pauses, mainly filled ones (filled by 'uh', 'uhm' or 'mm'), is more frequent in uncertainty; ii) concerning the visual cues: gaze not focused and funny (marked) face, which are linked to uncertainty; presence of eyebrows movements or smiles, generally related to uncertainty [11, 12, 13, 14].

In BP, the prosody of certainty and uncertainty attitudes, as well as the prosody of other attitudes that are said to be related to them (like doubt, incredulity or obviousness) was mainly studied in acted speech. Most of these works investigate these attitudes in acted sentences, included or not in contextual situations [17, 18, 19, 20, 21]. The prosodic differences described in these attitudes concern fundamental frequency - F_0 (higher at the end of the sentence in uncertainty and doubt attitudes), intensity (stronger in certainty), duration (longer in uncertainty), and the presence of pauses and fillers (mainly in uncertainty).

For emotional prosody (involuntary production processing), some perceptive discriminations between acted and spontaneous speech were shown [22, 23]. The next step of this present study will be to compare acted and spontaneous expressions of certainty/uncertainty.

Thereby, we come to a paradox: how can we control the induction of spontaneous attitudes in order to get comparable data for several subjects, together with acted speech for the same subjects? In other words, how can we study real expressions of certainty and uncertainty attitudes on (quite) the same linguistic material? Using real interaction corpora can provide spontaneous speech, but we would not be sure that lexical items will be the same, as discussed in [24, 25]. In BP, the works investigating attitudes' prosody with spontaneous corpora [e. g. 8] had another problem: how to annotate the real productions, since even if the induction process is supposed to carry on certainty/uncertainty, it must be verified without influencing the annotation by the expected labels? To try to

solve these problems and with the aim to study certainty and uncertainty in a spontaneous comparable corpus we decided to create a job interview scenario.

2. Material and methods

2.1. Corpus

(Un)certainty occurs in many contexts of verbal interactions. One context, perhaps easier to control, would be situations where answering questions is needed. For that reason it is the kind of protocol used the most to collect data to study (un)certainty. We can find many real situations where one has to answer questions: knowledge games, job interviews, giving information and so on. To have spontaneous answers in which (un)certainty can be expressed, we chose to create a job interview scenario (rather than doing a real job interview, due to ethical reasons), in which we could also use factual questions (which answers would be comparable due to lexicon similarities). This way we were able to use two kinds of question/answer interaction.

Many studies held by the Grenoble team are based on this methodological choice (a combination of a scenario conducted by Wizard of Oz method) to induct spontaneous affective speech [23, 24, 26, 27]. The advantages of this choice were discussed in [23], and among them we can highlight the possibility to induct attitudinal production in a controlled way.

To play this scenario, we chose to use EmOz, a Wizard of Oz platform, developed at LIG laboratory [28], which works with a companion robot Emox (developed by the Awabot Company <www.awabot.com>). This choice is based on three assumptions: the small robot, whose form is neither human nor animal, cannot perform humanoid gestures, that is the anthropomorphisation and interpretation of robot movements are reduced and then all the attention can be paid to the vocal productions; the recorded sentences said by the robot could be exactly repeated in each interaction with each different subject, in the same order, to have the same stimuli for each participant; the EmOz platform allows to control the robot remotely, while the subject thinks (s)he was interacting with a(n) (intelligent) machine.

About the EmOz platform, we use Excel files (scripts) associated to the robot Urbi system through a Java interface to program the robot's actions [28]. In these files it is possible to include the robot's actions like body or head movements, sounds or a combination of sounds and movements. The sentences the robot will say are prerecorded as wave files. The platform has a protocol and an improvisation mode. In the protocol mode we have a table with all the programmed robot's actions, in order, and we can follow the sequence by just clicking each table line. In the improvisation mode we have all the scripts and sounds as clickable buttons (and we can choose the scripts which will appear as buttons and drag and drop these buttons to organize them), and finally we have a recording button to record and play sounds on the spot (See figure 1).

The job interview scenario was organized as follow: first we recruited 9 Brazilian students at Grenoble University, two men and seven women, aged from 21 to 32 years, telling them they would do a job interview with a robot that could do and analyze job interviews. This way, the subjects thought they would evaluate the robot system while the robot would train them to do job interviews. Systems like this exist or are in development [29, 30, 31], which makes our scenario credible.

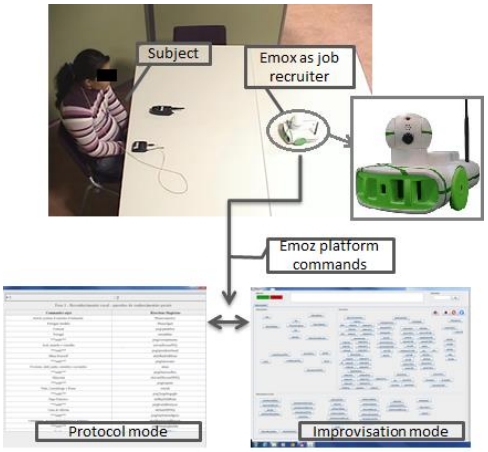


Figure 1: *Scheme to record the corpus*

When subjects arrive to do the job interview, we tell them there is a first interview phase in which the robot will make factual questions. This phase is justified by the need of voice recognition, i. e., the robot has to "learn" the subjects' voice, and the sequence of questions has this aim. In this context we placed a headset microphone on the subject, saying that this microphone would help with the voice recognition. The participants did not know they were being recorded. We also said this first phase was a pre-test used in some job interviews, in which the recruiter asks general knowledge questions. The factual questions were selected from a Brazilian factual questions collection, available in [32]. To ensure the production of answers with certainty and uncertainty we diversified easy and difficult questions. A group of 30 questions was evaluated by 10 Brazilian university students as easy, medium and difficult questions. After this evaluation, we saved 4 questions for each difficulty level (4 easy, 4 medium and 4 difficult), the four best within the judges agreement.

Along these questions we elaborated 12 mixed questions, which asked more than a thing at a time, and vary the level of difficulty among these asked items. For example, we asked which months of the year have less than 31 days. As February has 28 days, we expected subjects would easily remember February, but other months of the year that also have less than 31 days would take more time to remember. That's why in these questions, we expected to have productions in which subjects would deal with certainty and uncertainty at the same time, in the same answer. These 12 mixed questions were also evaluated by Brazilian students, and we saved the 8 in which judges agreed best in the classification. This way we saved 20 factual questions. These 20 factual questions were recorded by a Brazilian woman, with university education, who produced the questions with a standard Brazilian pronunciation. To start the experiment, the robot said it was ready to start. After each answer, there was a feedback from the robot saying it had heard the answer.

The job interview questions were taken from sites that list the main job interview questions in BP [e. g. 33 and 34]. We chose 12 of the most frequent questions in these sites. These questions were recorded by the same Brazilian woman who recorded the factual questions. To guarantee a real interaction with the subject in this phase of the experiment, the organization of the questions was made to make room to other questions, formulated at the moment of the interview, depending on the answers given by the subject. For example, one question asked was 'what are your strengths and your

weaknesses?'. If the subject answered that one of his(her) strengths was determination, we formulated a question like 'do you think your determination may help you in your work?'. These questions were included to make the robot look more intelligent, able to analyze a job interview, as we said, making our scenario more believable.

During the first (factual questions) and second (job interview) phases of our scenario we recorded the subjects, although they did not know, with two cameras: the first one directed to subjects' face and the second one taking the entire context. In Figure 1 we can see an image taken from the second camera, with the organization made for recording data.

In a debriefing phase, after the job interview, we asked the participants to evaluate the robot's system. After this, we told them that some participants evaluate as a strange situation to do a job interview with a little robot as a recruiter, so we told we would like to compare the manner to answer a robot to the manner to answer a person. With this pretext, we asked the subjects to answer again all the questions, this time asked by a human, thus we have one natural and one artificial production of the same linguistic material. Subsequently we told the subjects that we recorded all the experiment and we asked them if they could come back to watch their videos and to annotate how they felt during the experiment.

Three or four days after, subjects came back to watch their videos and freely annotate their feelings during the experiment. In this auto-annotation, the subjects could already remark they felt something related to (un)certainty. Nevertheless, to ensure this type of evaluation, we asked the participants to choose in a forced test if they felt certainty or uncertainty for each answer to factual questions.

At the end we revealed the real aim of the study: to find how we express (un)certainty in BP. According to this aim, we gave subjects their answers transcribed and we asked them to answer the questions twice again, firstly acting certainty and then acting uncertainty in each of their answers.

2.2. Audiovisual prosodic analysis

In this work, we will present only the analysis made on the answers of factual questions. With 20 answers from each of our 9 subjects, 180 answers were analyzed, and this number can vary due to the answers composed by more items.

The recorded videos were analyzed to verify the presence of eyebrows, eyes movements and funny face (considered as mouth movements), to watch gaze direction (if it was diverted or directed) and head movements (yes/no), while observing the attitude to which these factors were linked. Then we calculated the percentages of presence/absence of these factors.

Concerning audio, data were analyzed in Praat software [35], and fundamental frequency and duration measurements were made. Regarding fundamental frequency, initial, final, maximum and minimum F_0 values were taken for each sentence. After that, the final F_0 movement was measured, and its shape was observed. These values were taken in semitones, because different speakers produced the sentences and the relative semitones measurements can minimize interpersonal differences. However, results are presented separately for male and female subjects. Regarding duration, we measured the entire duration for the sentence and the number of pronounced syllables, and the ratio of these values generated the rate of articulation in syllables per second. We also took the duration of all pauses and noted if each pause was filled or not. The delay time before starting to answer was measured too.

3. Preliminary Results and Discussion

3.1. Video analysis

In table 1 we present the summary of the movements perceived in video analysis regarding answers with certainty and uncertainty. We added a third attitude, as described by subjects in their auto-annotations, when they were sure to not know the answer. In some questions (in which we asked more than a thing at the same time) they also annotated they were sure about some item(s), while they were not sure about other(s). Thus we analyzed 242 noted attitudes instead of 180.

Table 1. *Movements observed in the videos*

attitude	eyebrows movements	diverted gaze	Mouth movements	head movements
certainty (127)	7 (5.5%)	26 (20.5%)	2 (1.6%)	4 (3.1%)
cert. of not (56)	16 (28.6%)	29 (51.8%)	20 (35.7%)	17 (30.4%)
uncertainty (59)	21 (35.6%)	52 (88.1%)	25 (42.4%)	25 (42.4%)

As we can see in the first table, the percentages of movements are higher in attitudes that involve uncertainty or certainty of not knowing. In some certain answers we find movements (precisely in 25.2% of the answers classified by subjects as certain), that is mainly of the eyes. We believe these movements are not exactly the same they used to express uncertainty, but it can be related, in the case of the eyes, to a moment that the participant takes to remember a well-known answer. Regarding head movements, the times they appeared in certainty it was as a nodding movement, unlike the head movement in uncertainty, the head shake. The other movements found in certainty appearing at the end of the answer were related to the next answer item, which was not known by the subjects. Another interesting thing is that each subject has his own manner to express uncertainty by using gestures: while some of them use diverted gaze others prefer using eyebrows movements. In 97.8% of uncertain answers at least one of these gestures is used (often two or three gestures are used in the same answer). In the answers where the subjects are sure to not know the answer, 90.9% have at least one of these gestures.

In figure 2 we can see some examples of the movements found to express uncertainty, such as mouth, head and eyebrows movements or diverted gaze (on the right in each pair) in contrast with certainty (on the left in each pair) (it is important to observe that "neutral position" is to look down, with the face turned to the robot).

Figure 2: *Facial expressions in certainty (on the left in each pair) and in uncertainty (on the right in each pair).*

3.2. Audio analysis

The audio of the answers was analyzed regarding general characteristics such as the delay time, the filled pauses, the final F_0 movement (see table 2). We also analyzed the order

inversion or the presence of lexical items to say which attitudes were used (e.g. 'it can be...' or 'I think...').

Table 2- *General characteristics of the answers*

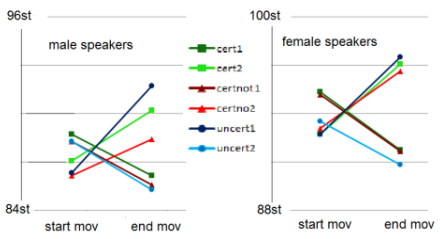
S	Delay (L-long or S-short)			Filled pauses (presence)			Final F_0 movement (A-ascending or D-descending)		
	C	CN	U	C	CN	U	C	CN	U
F	29L 84S	24L 14S	34L 13S	2	24	21	54A 59D	10A 28D	32A 15D
M	4L 10S	11L 7S	5L 7S	0	10	7	5A 9D	5A 13D	8A 4D
Tot (N)	**33L 94S (127)**	**35L 21S (56)**	**39L 20S (59)**	**2 (127)**	**34 (56)**	**28 (59)**	**59A 68D (127)**	**15A 41D (56)**	**40A 19D (59)**

Legend: F- Female subjects; M- Male subjects; C- certainty; CN- certainty of not knowing; U- uncertainty; N- number of occurrences

As can be seen in table 2, regarding the delay time, it was divided in short and long values, depending if they were longer or shorter than median value. The certainty attitude presents a tendency to have the smallest values of the delay time, while uncertainty and certainty of not knowing attitudes present the higher values. Regarding the presence of filled pauses, they were almost not found in certainty, but they could be commonly found in uncertainty and certainty of not knowing. Finally, regarding the final F_0 movement, despite some variation, we can state that mainly a descendent movement was used in certainty and certainty of not knowing attitudes, while an ascending one was used in uncertainty attitude. As we have different directions of the F_0 movements, if we compare the rising movement found in uncertainty with the falling one found in certainty and certainty of not knowing, the values found are significantly different ($p<0,05$, CI 95%).

The variations found in the F_0 movements were related to other characteristics of the attitudes' expression in our corpus: an ascending movement is expected in an uncertain answer, but if the speaker uses expressions like 'I think it is...' the uncertainty is already represented lexically, so the F_0 movement chosen was a descending one. We found ten uncertain answers with a descending movement that have these lexical expressions to express uncertainty in our data. In some answers we also found order inversion: the subject answered the most difficult item at the end of the sentence and, to indicate finality, the end of the sentence was descending. These reasons could explain why we found some uncertain answers with a falling F_0 movement. Concerning ascending movements, they were found in certainty and certainty of not knowing attitudes mainly in enumerations, in which the first items would be ascending to indicate continuousness; we have 8 questions which ask more than a thing at the same time, so we have 8 answers we expected to have an enumeration of some items.

Graphic 1 – *Final F_0 movement of the answers*



The two types of F_0 movement found in each attitude can be seen in detail in graphic 1. The darker colors represent the most recurrent movement for each attitude, while the lighter colors represent the other movement found. Comparing darker colors, we can see typical final F_0 movements: rising for uncertainty (dark blue) and falling for certainty and certainty of not knowing (dark red and dark green). In figure 3 we present a typical example of certainty and uncertainty answers.

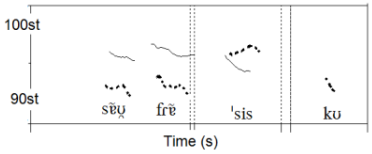


Figure 3: *F_0 contours of the answer "São Francisco" (*Saint Francisco*) with certainty (continuous line) and uncertainty (dotted ticked line), said by two women. Marked syllable is the tonic one, with a falling and a rising F_0 movement.*

Regarding duration measures, we compared results of the articulation rate (speech rate, in syllables per second) and duration of the last tonic syllable and its previous one (see table 3). The male speakers' speech rate varies, being used in a different way to express the attitudes: uncertainty is significantly different from the other attitudes, with a significantly higher value of speech rate. For female speakers no difference is statistically significant and the tendencies of higher values are different from men's tendencies. The duration of the final syllables, studied as an important cue to express uncertainty in BP previous works do not point out a clear tendency in our study; these values do not seem to play a role to express spontaneous (un)certainty attitudes in BP.

Table 3- *Speech rate and duration of the final syllables*

attitude	Male speakers			Female speakers		
	syl/s	tonic	pre-tonic	syl/s	tonic	pre-tonic
certainty	4.48	296.5	179.9	4.58	310.2	192.8
cert. of not	5.01	293.7	158.1	4.65	274.3	183.8
uncertainty	5.99	252.5	158.3	4.64	284.9	163.2

4. Conclusions

Collecting spontaneous speech data using a Wizard of Oz method within a scenario revealed to be an efficient way to induct spontaneous but comparable attitudinal speech utterances. Audiovisual prosodic results, such as F_0 and duration, eyebrows, eyes, mouth and head movements, showed personal behaviors but general tendencies used to express (un)certainty in BP. In further studies, we will compare spontaneous and acted utterances produced by the same subjects and we will also analyze other collected data with this method to study other (un)certainty characteristics in BP.

5. Acknowledgements

We would like to thank Brigitte Meillon, responsible for the Multicom Platform at LIG lab. Thanks to Nicolas, Gilles, Sylvie and Adrien for all the assistance. Thanks to CAPES Foundation, Brazil, for the postdoctoral scholarship granted to Leandra Antunes – BEX 18020-12-7. This work has been partially supported by the Major Program for the National Social Science Found of China (13&ZD189), the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and by The Interabot Project (Investissements d'Avenir, DGCIS).

6. References

[1] Fónagy, I. “Des fonctions de l’intonation: essay de sinthèse”, Flambeau, 29: 1-20, 2003.

[2] Couper-Kuhlen, E. An introduction to English Prosody. Tübingen, Niemeyer, 1986.

[3] Aubergé, V. “A Gestalt Morphology of Prosody Directed by Functions: the Example of a Step Model Developed at ICP”. Proceedings of the 1st Conference on Speech Prosody, 151-155. Aix-en-Provence, 2002.

[4] Wichmann, A. “The attitudinal effects of prosody, and how they relate to emotion”. In: Cowie, R., Douglas-Cowie, E. and Schröder, M. [eds], Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, 2000.

[5] Reis, C. “A entonação no ato de fala”. In: Mendes, E., Oliveira, P. and Benn-Ibler, V. [eds] O novo milênio: interfaces lingüísticas e literárias, 221-229, Belo Horizonte, UFMG/FALE, 2001.

[6] Rilliard, A., Moraes, J., Erickson, D. and Shochi, T. “Prosodic analysis of Brazilian Portuguese attitudes”. Speech Prosody, Chicago, 2010.

[7] Moraes, J., Rilliard, A. Erickson, D. & Shochi, T. Perception of attitudinal meaning in interrogative sentences of Brazilian Portuguese. Proceedings of 17th International Congress of Phonetic Sciences (ICPhS). Hong Kong, 2011.

[8] Antunes, L. O papel da prosódia na expressão de atitudes do locutor em questões. Tese de doutorado: FALE/UFMG, 2007.

[9] Antunes, L. B. “Propositional attitudes in interrogative sentences in Brazilian Portuguese”. In: Proceedings of WASSS- Workshop on Affective Social Speech Signals. Grenoble, France, 2013.

[10] Aubergé, V. “Attitude versus emotion: a question of voluntary vs. involuntary control”. In: GSCP, Invited talk. Belo Horizonte, 2012.

[11] Dijkstra, C., Krahmer, E. J., & Swerts, M. “Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence”. In Hoffmann, R. & Mixdorff, H. (eds). Proceedings of 3rd Speech Prosody, Dresden, 2006. Online: <http://arno.uvt.nl/show.cgi?fid=95687>. Accessed on 17 set. 2013.

[12] Krahmer, E.J., & Swerts, M. “How children and adults produce and perceive uncertainty in audiovisual speech”. In: Language and speech, vol. 48, n. 1, p. 29-54, 2005. Online: <http://las.sagepub.com/content/52/2-3/129.full.pdf+html>. Accessed on 17 set. 2013.

[13] Swerts et al. “Audiovisual cues to uncertainty”. In: Carlson, R. et al. In: Proceedings of the ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems. Chateau-D’Oex, 2003. p. 25-30. Online <http://wwwhome.cs.utwente.nl/~laar/CA2/Edwins%20Finest/Audiovisual%20cues%20to%20uncertainty.pdf> Accessed on 17 set. 2013.

[14] Swerts, M., & Krahmer, E. Audiovisual prosody and feeling of knowing. In: Journal of Memory and Language, v. 53, p. 81-94, 2005. Online: <http://arno.uvt.nl/show.cgi?fid=107744>. Accessed on 23 set. 2013

[15] Brennan, S. E. & Williams, M. “The feeling of another’s knowing: prosody and filled pauses as cues to observers about the metacognitive states of speakers”. In: Journal of Memory and Language, v. 34, p. 383–398, 1995. Online: <http://www.psychology.stonybrook.edu/sbrennan-/papers/brenwill.pdf>. Accessed on 23 set. 2013.

[16] Smith, V. L. and Clark, H. H. “On the course of answering questions”. In: Journal of Memory and Language, v. 32, p. 25-38, 1993. Online: <http://psych.stanford.edu/~herb/1990s/Smith.Clark.93.pdf> Accessed on 23 set. 2013.

[17] Moraes, J. “From a prosodic point of view: remarks on attitudinal meaning”. In: Mello, H., Panunzi, A., Raso, T (eds.) Pragmatics and Prosody: Illocution, modality, attitude, information patterning and speech annotation. Firenze: Firenze University Press, 2011.

[18] Rilliard, A. Moraes, J. A. de Ericson, D. & Shochi, T. “Prosodic analysis of Brazilian Portuguese attitudes”. In: Ma, Q., Ding, H. & Hirst, D. (orgs). Proceedings of the 6th Speech Prosody, v. 2. Shangai, may 2012. p. 677-680.

[19] Azevedo, L. Expressão da atitude através da prosódia em indíviduos com doença de Parkinson idiopática. Tese de doutorado: UFMG/FALE, 2007.

[20] Oliveira, B. A prosódia na expressão das atitudes de dúvida, incerteza e incredulidade no português brasileiro. Dissertação de Mestrado: FALE/UFMG, 2011.

[21] Silva, J. P. G. A prosódia na expressão da dúvida e da certeza no português brasileiro. Dissertação de Mestrado: FALE/UFMG, 2008.

[22] Audibert, N., Aubergé, V. and Rilliard, A. “Prosodic Correlates of Acted *vs.* Spontaneous Discrimination of Expressive Speech: A Pilot Study”. Proceedings of the 5th Speech Prosody, Chicago, 2010.

[23] Audibert, N. Prosodie de la parole expressive: dimensionnalité d’énonces méthodologiquement contrôlés authentiques et actés. Thèse de doctorat, Université de Grenoble, 2008.

[24] Aubergé, V., Audibert, N. & Rilliard, A. “E-Wiz: A trapper protocol for hunting the expressive speech corpora in lab”. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbonne, Portugal, p. 175-178.

[25] Aubergé, V., Audibert, N. & Rilliard, A. (2003). “Why and how to control the authentic emotional speech corpora?” Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH 2003), Genève, Suisse, p. 185-188.

[26] Sasa, Y. “EmOz : magicien d’Oz pour les interactions entre personnes âgées- un robot compagnon dans l’habitat intelligent Domus”. 123fls. Mémoire de master Sciences du langage. Grenoble: Université Stendhal, 2013.

[27] Sasa, Y., Aubergé, V., Franc, P., Guillaume, L., Moujtahid S. “Des microexpressions au service de la macro-communication pour le robot compagnon EMOX”. In: Actes du WACAI 2012, Grenoble, p. 54- 59,. 2012.

[28] Aubergé, V., Sasa, Y., Robert, T., Bonnefond, N., Meillon, B. and Guillaume, L. “EmOz: a wizard of Oz for emerging the socio-affective glue with a non-humanoid companion robot”. In: Proceedings of WASSS- Workshop on Affective Social Speech Signals. Grenoble, France, 2013.

[29] Anderson, K. et alii. “The TARDIS framework: intelligent virtual agents for social coaching in job interviews”. In Proceedings of Advances in Computer Entertainment. Springer-Verlag, p. 476-491, 2013. Online: <http://perso.limsi.fr/sabouret/ps/wacai2012-hazael.pdf>. Accessed on 17 out. 2013.

[30] Hoque, M. E, Courgeon, M., Martin, J. C., Mutlu, B. & Picard, R. W. “MACH - My Automated Conversation coacH”. 2013. Online: <http://web.media.mit.edu/~mehoque/Publications/13.Hoque-etal-MACH-UbiComp.pdf>. Accessed on 11 out. 2013.

[31] Jones, H., Sabouret, N. & Gondré, M. S. “Un modèle affectif pour un recruteur virtuel dans le contexte de simulation d’entretiens d’embauches”. In: Actes du WACAI – Workshop Affect, Compagnon Artificiel, Interaction. Grenoble, nov. p. 28-36, 2012. Online: <http://perso.limsi.fr/sabouret/ps/wacai2012-hazael.pdf>. Accessed on 17 out. 2013.

[32] PIG - Projeto de Informação Geral - 530 questões para trabalhar em sala de aula. Online: <http://www.coronelsarmento.xpg.com.br/pig.htm>. Accessed on 10 oct. 2013.

[33] Doze perguntas mais frequentes numa entrevista de emprego. Online: <http://www.curricular.com.br/artigos/entrevista-emprego/perguntas.aspx>. Accessed on 11 oct. 2013.

[34] Carreira. Testes. Revista EXAME. Online: <http://exame.abril.com.br/carreira/testes>. Accessed on 11 oct. 2013.

[35] Boersma, P. and Weenik, D. Praat: doing phonetics by computer. V. 5.24. Free software. Available from: www.praat.org.

Between Recognition and Resignation – The Prosodic Forms and Communicative Functions of the Czech Confirmation Tag "jasně"

Jan Volín,[1] *Lenka Weingartová*,[1] *Oliver Niebuhr*[2]

[1]Institute of Phonetics, Charles University in Prague, Czech Republic
[2]Department of General Linguistics, ISFAS, Christian-Albrecht-University of Kiel, Germany
{jan.volin|lenka.weingartova}@ff.cuni.cz, niebuhr@isfas.uni-kiel.de

Abstract

Like question tags, confirmation tags such as the Czech affirmative particle *jasně* can be used with various prosodic characteristics that augment, reverse or otherwise modify their relatively unspecific lexical meaning. We extracted 172 instances of *jasně* from several dialogues and assessed their discourse function. 36 prosodic correlates in temporal, amplitude and fundamental frequency domains were measured and used in three computational classifiers: linear discriminant analysis, classification trees and artificial neural networks. All three methods significantly reflected the functional assessments and additionally indicated the relative importance of individual predictors in a mutually consistent manner.

Index Terms: affirmative particle, confirmation tag, Czech, discourse, intonation, pragmatics.

1. Introduction

It is obvious and has been repeatedly shown for many languages that 'question tags' like *isn't it* in English, *nicht wahr* in German, or *verdad* in Spanish are very rich both phonetically and functionally, cf. [1,2,3,4,5,6,7]. They are used by speakers to keep the interaction going and/or promote the flow of information. At the same time, the lexical semantics of question tags is fairly unspecific. Taken together, this allows question tags to occur in very different communicative contexts and in combination with all kinds of prosodically expressed speaker attitudes and emphatic intensifications.

The same also applies to 'confirmation tags' like *of course* in English, *alles klar* in German, or *todo bien* in Spanish. Confirmation tags, which, if they are single words, can also be called affirmative particles, are moreover everything but rare in conversation. Their high frequency in combination with their flexible application and constant segmental basis make confirmation tags – just like the better investigated but probably rarer question tags – an ideal research subject for studying the prosodic forms of a language and their respective communicative functions. In this context, the present study deals with the disyllabic Czech confirmation tag *jasně* ['jas.ɲe] whose closest English equivalents would be *sure*, *agreed*, *of course*, or *fair enough*. Speakers ordinarily insert *jasně* at the beginning of their utterances, typically as a separate prosodic phrase, in order to react to a preceding turn of the interlocutor.

Our major aim is to determine, describe, and systematize the prosodic and functional variation that can occur on *jasně*, in this way also advancing our understanding the prosodic system of Czech in general. While particularly the phonological factors and prosodic correlates of lexical stress as well as the related phenomena of phrasing and rhythm have been intensively analyzed for Czech in the last decades (cf. [8,9,10,11,12,13]), only relatively little is known about the use of intonation and emphasis patterns in Czech. Unlike research on the prosody of emphatic expressions in Czech, which does virtually not exist, there are some studies on Czech intonation. However, the corresponding research has so far often been descriptively oriented in the sense that intonational forms and functions have been characterised and contrasted on the basis of exemplars, impressions and experience, or with the primary aim to develop annotation inventories and enhance speech technology applications, cf. [14,15,16]. Analyses that aimed at a detailed, empirically based understanding of intonational forms, functions, and their linkages have only just come up in the last few years, cf. [17]. Our paper follows this more recent, empirical line of research.

This paper will summarize the production part of our study. The production data come from a large corpus of enacted (i.e., text-based) dialogues conducted by 30 native speakers of Czech. The data were acoustically analyzed in terms of a number of different duration, F0, and intensity measures. The production part will soon be complemented by a perception part, serving to cross-validate the form-function links that emerge on Czech *jasně*. Both the communicative functions and the acoustic-prosodic parameters on which our analysis of *jasně* is based were inspired by German whose intonational and emphatic categories and structures have been thoroughly explored in the last decade, cf. [18,19,20,21,22]; and it is probably not exaggerated to state that intonation and emphasis structures in German are already fairly well understood.

Against this background, the specific questions that we address here are the following:

(1) Do we find systematic prosodic variation on Czech *jasně*?

(2) If the answer to (1) is positive, is this variation functionally motivated, i.e. meaningful? Or is the variation just contextually motivated and due to speaking rate, phrase structure, or speaker-specific effects?

(3) If the answer to (1) is positive, is this variation multiparametric or rather dominated by a single prosodic parameter?

The answers to these questions will later allow to put the issue into a cross-linguistic perspective. For example, the absolutely strict lexical stress position in Czech reduces the corresponding functional load of duration and/or intensity so that these parameters could even play a more important role in signalling emphasis categories than in German. If this is the case, will the respective prosodic patterns be still associated with the same communicative functions as in German?

Three classifiers will be used to gain a cross-evidenced view of the variables, of which some map very similar properties as the others differing only in conceptual detail (see below). This will provide a methodological advantage for further research.

2. Method

A total number of 180 *jasně* tokens from the Prague Phonetic Corpus [23] were used. The target word occurred in six different contexts in the corpus and each was uttered by 30 native Czech non-professional speakers, 24 female, 6 male, aged 20-25 years. Scripted texts were used to elicit short dialogues from pairs of speakers. The speakers were explicitly encouraged to familiarize themselves with the dialogues and then act them out as convincingly as possible. The participants got along with the task very well, taking various affective approaches. Nonetheless, two trained phoneticians, who controlled the recording process, asked for new trials when dysfluencies or unnatural renderings occurred.

The recordings were made digitally at a sampling rate of 32 kHz and with a 16-bit quantization in the sound-treated studio of the Institute of Phonetics in Prague, using an IMG ECM2000 microphone and a SB Audigy 2ZS soundcard.

2.1. Perceptual categories

First, all 180 target word tokens were surveyed on an auditory basis by three Czech trained phoneticians (two of whom were authors of this paper). This auditory survey in combination with the phoneticians' native-speaker intuitions led to setting up eight functional categories:

Type 1: neutral acceptance

Type 2: eager agreement

Type 3: impatience

Type 4: indifference, patronizing

Type 5: wonder, surprise

Type 6: recognition, realizing

Type 7: resignation

Type 8: reassurance, sympathy

Having set up these categories, each target word was listened to and assigned to one of the categories. The assignment procedure was conducted independently by the three phoneticians. In the case of disagreement the respective token was discussed and the majority vote was taken. In the end, eight tokens had to be discarded due to disagreement of all three listeners, so that 172 words were left for further analysis.

2.2. Acoustic measurements

Acoustic analyses of the target words were carried out in *Praat* [24], individual segment boundaries were manually labelled. The following parameters were measured:

Temporal:

- word duration (in ms)
- relative segment duration (in % of word duration, and in % of syllable duration)
- relative syllable duration (in % of word duration)
- difference between the duration of syllable nuclei (in ms, [a]-[e])
- difference between the duration of syllable onsets (in ms, [j]-[ɲ])
- durational profiles: the outcome of a cluster analysis (4 clusters, k-means) where the individual cases (renderings of the word) were clustered according to their segment durations (in % of word duration)

F0:

- first and second extreme of the F0 contour (i.e., maximum and minimum or vice versa) normalized to speaker range (in %) and speaker average (in ST)
- the difference of the first and second extreme in the F0 contour (in ST re 100 Hz)
- the difference between vowels, i.e., between the F0 mean values taken in the middle third of each vowel (in ST)

Speaker range and average values for normalization were taken from all six utterances in which the target word occurred, rather than just from the target word itself. Errors in F0 extraction were manually corrected, and portions of the signal with creaky voice were excluded, so that we obtained an estimate of the speaker's modal range. Values of the minima and maxima in the target word were measured manually, F0 micro-perturbation was disregarded. Creaky voice in the target words was subsequently set to be at 0 % of the speaker's range rather than at negative values, since this has improved the discriminatory power of the variable in preliminary analyses and most probably reflects the speaker's intention of hitting 'ultimate low' rather than a specific frequency target.

Energy:

- maximum SPL value in the target word, normalized by average utterance SPL (in dB; pauses and silences were excluded)
- location of the SPL maximum (in % of word duration, and in the corresponding segment)
- SPL in the middle of each segment, normalized by average utterance SPL (in dB)

Apart from these acoustic measurements, the intonation contour was also annotated by the third author with labels adapted from the Kiel Intonation Model [20,25]:

- prominence strength of each syllable in three levels (0 = no prominence, 1 = weak prominence, 2 = strong prominence)
- synchronization of the pitch-accent peak (early, medial, late) in weakly or strongly prominent syllables
- final boundary tone (0 = flat, 1 = moderately descending, 2 = falling to the lower end of the speaker's range)

Afterwards, the position of the (more) prominent syllable in the disyllabic target word and its and pitch-accent synchronization were merged into a single contour-descriptor label (e.g., 'FA' = early peak on the first syllable; 'MB' = middle peak on the second syllable).

All 36 parameters listed above were then used as variables in subsequent statistical analyses.

The discriminative strength of each variable was explored through one-way ANOVAs. For classifying the data into the eight perceptual categories, linear discriminant analysis (LDA), classification and regression trees (CART) and artificial neural nets (ANN) were used. The advantage of using CARTs and ANNs is the possibility to employ both continuous (e.g., temporal or F0 parameters) and categorical (e.g., duration profiles, intonation labels) variables as predictors. Moreover, CART can use one and the same variable repeatedly at different split decisions. All the classifiers used were from the STATISTICA software package [26].

3. Results

3.1. Counts of Tokens in Functional Categories

Out of the 172 investigated tokens, 43 cases were assigned to Type 2 (eager agreement) and 42 cases to Type 1 (neutral acceptance). Type 6 (recognition, realizing) was represented by 29 cases, Types 4 (patronizing) and 5 (wonder) both by 18 cases, while for Types 7 (resignation), 8 (reassurance), and 3 (impatience) only 9, 8 and 5 cases, respectively, were found.

3.2. Discriminant Analysis

Linear discriminant analysis was performed after searching for continuous variables that do not correlate too highly with each other and differentiate well among the individual functional categories. Unrestrained analysis (i.e., mapping the structure of the dataset with rather loose tolerance levels) resulted in a success rate of 57.6 % and identified the following variables as best reflecting the assumed functional categories: duration of the vowel /e/ in the second syllable of *jasně* relative to the word duration, vowel /e/ duration relative to syllable duration, consonant /s/ duration relative to syllable duration, duration of first syllable relative to word duration, durational difference between vowels, durational difference between syllable onsets, and F0 difference between the first and second extreme.

Several further analyses were performed with more stringent tolerance levels. Although the success rate for the best outcome dropped to 52.3 %, the results can be considered more generalizable. Only five variables were ultimately used: word duration, durational difference between vowels, durational difference between syllable onsets, F0 difference between vowels and normalized value of the first F0 extreme. Table 1 displays the ensuing confusion matrix. It is apparent that under-represented functional categories (Type 3, 7 and 8, i.e., impatience, resignation and reassurance respectively) were not recognized in this more rigorous setting of the LDA.

LDA		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	22	11	0	4	1	2	1	1
	2	15	28	2	2	0	7	2	1
	3	0	0	0	0	0	0	0	0
	4	2	1	1	9	5	0	0	4
	5	2	0	0	1	10	0	0	1
	6	0	3	2	2	0	20	6	0
	7	0	0	0	0	1	0	0	0
	8	1	0	0	0	1	0	0	1
Corr. %		**52**	**65**	**0**	**50**	**56**	**69**	**0**	**12**

Table 1. *Confusion matrix resulting from the most successful discriminant analysis. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

The success rate for other categories was 50 % and more. The highest numbers off the diagonal can be found for Types 1 and 2 (neutral acceptance and eager agreement). They seem to be highly confusable, although the correctly identified cases in these two abundant types still prevail. The most distinct functional category seems to be Type 6 (recognizing) with nearly 70 % of the cases correctly separated from other categories and with errors towards Types 1 and 2 again. As stated above, the lowest success was achieved for the smallest groups, of which Type 3 was represented by 5 instances only.

3.3. Classification and Regression Trees

The algorithm used in STATISTICA calculates automatically the usefulness of all the input variables and ranks them according to their effectiveness in the classification process. The most successful tree achieved the success rate of 65.7 %, which is by about 10 % more than in our earlier discriminant analyses. The best tree had 9 splits and was based on word duration (ranked as the most important predictor), F0 difference between vowels, durational difference between syllable onsets, relative duration of a syllable within the word, normalized F0 value of the first extreme and normalized intensity of the first vowel. Other intensity measures and categorical labels of intonation and temporal profile were found unimportant, while the word duration and F0 difference between vowels were used twice, i.e. for two different splitting decisions. The ensuing confusion matrix is shown in Table 2. Further splitting could still be ordered, but only at the expense of generalizability, hence we did not proceed with it. The number of confusions between Types 1 and 2 is lower than in previous analyses, but a considerable number of Type 2 cases were misclassified as Type 6 (see below, Table 2, the second numbered column).

CART		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	21	4	0	0	0	0	0	0
	2	12	28	2	0	0	0	1	0
	3	0	0	0	0	0	0	0	0
	4	1	0	0	11	0	0	0	1
	5	3	0	1	0	18	0	0	0
	6	2	11	2	4	0	29	8	1
	7	0	0	0	0	0	0	0	0
	8	3	0	0	3	0	0	0	6
Corr. %		**50**	**65**	**0**	**61**	**100**	**100**	**0**	**75**

Table 2. *Confusion matrix resulting from the most successful CART analysis. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

The best classification of categories was achieved for Types 5 and 6 (wonder and recognizing) whose all instances were correctly found. However, some other types were also mistakenly added to these categories. From this point of view, Type 5 seems to be better as only four improper cases were added to it. The rare Types 3 and 7 (impatience and resignation) were not recognized at all, but Type 8 (reassurance), which was represented by 8 cases in our dataset was classified relatively successfully.

3.4. Artificial Neural Nets

Thirty different architectures and settings were tried always with eight output neurons (corresponding to eight functional categories). The initial analyses used all the variables available as the input with the aim to evaluate of their individual usefulness. Sub-sequent analyses only utilized the most effective predictors. It turned out that Multilayer Perceptron Neural Networks outperformed other available types (RBF, LNN).

ANN		Observed Types							
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Predicted Types	*1*	30	11	1	2	1	3	1	1
	2	6	27	1	1	2	3	0	1
	3	1	0	2	1	1	1	0	0
	4	2	1	0	9	0	1	0	2
	5	2	1	0	3	14	0	2	2
	6	0	3	1	2	0	20	2	0
	7	0	0	0	0	0	0	4	1
	8	1	0	0	0	0	1	0	1
Corr. %		**71**	**63**	**40**	**50**	**78**	**69**	**44**	**12**

Table 3. *Confusion matrix resulting from the most successful ANN classification. Figures represent individual cases, except in the last line with percentages of correctly recognized cases within a category.*

As in the previous analyses, the most efficient variable was the duration of the word. Rather surprisingly, the second best was the durational profile of the word (a categorical variable), followed by the prominence strength on the first syllable (categorical), durational difference between syllable onsets, F0 difference between vowels, synchronization of the F0 peak (categorical), and normalized intensity of the first vowel. The automated neural networks that we used weigh some of the input variables by zero and make them ineffective. As a result, they do not suffer from dimensionality problems. The unrestrained model achieved a success rate of 66.3 %. When only the nine best variables were used in a three-layer perceptron architecture, the success rate dropped to 62.2 %, but probably with the advantage of better generalizability. Confusion matrix of the final analysis is displayed in Table 3.

Unlike in CART analyses, there are no 0 % or 100 % success rates in mirroring the functional categories. Type 5 (wondering) and Type 1 (neutral acceptance) were the best recognized with the success rates over 70 %. Some correct classification occurred even in the small groups of Type 3, 7 and 8 which were previously found difficult to capture (apart from Type 8 in CART analysis).

4. Discussion

Three classifiers performed their analyses with comparable levels of success. However, the lowest success rate in the case of the conventional discriminant analysis suggests that continuous linear relationships do not model prosodic dependencies best. As noticed in the past, acoustic correlates of prosodic features are used in various combinations, and the same features can be used for different communicative functions, in this way creating discrete 'islands' in a multidimensional space. If this is true, more advanced classifiers should be advantageous. More specifically, the best recognized categories, Types 5 and 6 (wondering and recognizing), were each found by CART at two different endpoints of the classification tree. This supports the idea that the same pragmatic or discourse effect can be achieved through different prosodic means. One way or another, our results allow for positive answers to the first two questions from the introduction: the prosodic variation in our data set appears to be systematic and functionally motivated.

The third question concerned the variables responsible for prosodic profiling of the individual functional categories. The word duration as an expression of the articulation rate was identified as a useful discriminatory element in all analyses performed. It seems that the rapidity (or slowness) with which the word *jasně* is pronounced is a reliable marker of the appended functions. Various other durational characteristics kept reoccurring as well, of which the most important one was the difference in duration of the consonantal onsets of the syllables. Interestingly, auditory inspections turned our attention to the duration of the word-initial consonant, which was markedly longer for some functional categories than for others, but the duration of this consonant relative to the duration of the word was computationally less useful than the same duration relative to the duration of the second syllable onset. Local durations of consonants thus might function in speech by being contrasted against each other rather than by being compared with the carrier unit as a whole.

As to the melodic correlates, the one repeatedly occurring as effective was the difference between F0 means measured in the middle thirds of the vowels (in ST). This variable could be sometimes replaced with the difference in F0 extremes within the word with a few percent shifts in the success rates. The relative pitch of the first syllable – expressed as either the F0 value within the speaker's range or as the annotated labels adapted from the Kiel Intonation Model – were also found relevant by the computational classifiers, although they were not eventually utilized in the most successful models.

The profiles of functional categories to be further examined in perceptual tests appeared to be as follows. Types 1 and 2 were spoken significantly faster than all the other types. It seems plausible to find neutral and eager stances brisk, whereas patronizing, wonder, realizing, resignation, and reassurance spoken more slowly. The major discriminator between Types 1 and 2 was then the difference in duration of the syllable onsets. Type 2 (eagerness) has significantly longer the word-initial consonant. A similar relationship is found between patronizing (short word-initial consonant) on the one hand, and wondering and realizing on the other hand (longer word-initial consonant). Melodically, Type 5 (wonder) was the only one with clearly rising F0 contour. Patronizing (Type 4) and reassurance (Type 8) were spoken with flat contour, while the rest of the types had falling melodies. As to energy, Types 1, 2 and 6 exhibited high SPL in first (i.e, stressed) vowel, whereas Types 4, 5 and 8 low SPL.

Similarly to English [27] or German [21], Czech functional categories seem to rely to a great extent on temporal and melodic cues, although intensity plays its role, too. An experiment is currently in progress, testing the perceptual response of German and Czech listeners to the exemplars from our current study.

The under-represented categories 3, 7 and 8 were difficult to classify. Although this can be due to the computational safeguards (not generalizing for small samples), our intuitive evaluation suggests that these categories are not just relatively rare, but also internally disparate. The pragmatic messages they signal (impatience, resignation, or reassurance, respectively) may be expressed by various means and, therefore, be less well-defined than their more frequent counterparts. Further research in this respect is needed, but prior verification of these categories by larger listener groups is crucial.

5. Acknowledgements

The 1st & 2nd author were supported by the Programme of Scientific Areas Development at Charles University in Prague, Subsect. 10, Linguistics: Social Group Variation. We also thank Hana Bartůňková for her help with the categorization.

6. References

[1] Cattell, R. “Negative transportation and tag questions”, Language 49(3): 612–39, 1973.

[2] Millar, M. and Brown, K., “Tag questions in Edinburgh speech”, Linguistische Berichte 60: 24–45, 1979.

[3] Cruz-Ferreira, M., “Tag Questions in Portuguese: Gramnnar and Intonation”, Phonetica 38: 341–352, 1981.

[4] Bald, W.-D., “English tag-questions and intonation”, in K. Schuhmann [Ed.], Anglistentag 1979: Vorträge und Protokolle, 263–91, Berlin: Technische Universität Berlin, 1980.

[5] Tottie, G. and Hoffmann, S., “Tag questions in British and American English”, Journal of English Linguistics 34(4): 283–311, 2006.

[6] Dehé, N. and Braun, B., “The prosody of question tags in English”, English Lang. & Linguistics 17.1: 129–156, 2013.

[7] Reese, B. and Asher, N., “Prosody and the interpretation of tag questions”, Proc. Sinn und Bedeutung 11: 448–462, Barcelona: Universitat Pompeu Fabra, 2006.

[8] Janota, P., “Personal Characteristics of Speech”. Praha: Academia, 1967.

[9] Janota, P. and Palková, Z., “Auditory evaluation of stress under the influence of context”, AUC Philologica 2/1974, Phonetica Pragensia, 4: 29–59, 1974.

[10] Volín, J., “Z intonace čtených zpravodajství: výška první slabiky v taktu”, Čeština doma a ve světě 1–2: 89–96, 2008.

[11] Volín, J. and Weingartová, L., “Idiosyncrasies in local articulation rate trajectories in Czech”, Proceedings of Perspectives on Rhythm and Timing, 67, Glasgow: UG, 2012.

[12] Romportl, J., “Statistical Evaluation of Prosodic Phrases in the Czech Language”, Proceedings of the Speech Prosody 2008 Conference, 755–758, Campinas, Brazil, 2008.

[13] Dankovičová, J., “Articulation rate variation within the intonation phrase in Czech and English”. Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, 1999.

[14] Kolář, J., Romportl, J. and Psutka, J. “The Czech speech and prosody database both for ASR and TTS purposes”, Proceedings of Eurospeech 2003, 1577–1580, Geneva: ISCA, 2003.

[15] Bartošek, J. and Hanžl, V., “Intonation Based Sentence Modality Classifier for Czech Using Artificial Neural Network”, Proc. NOLISP 2011, 162–169, 2011.

[16] Duběda, T. and Raab, J., “Pitch Accents, Boundary Tones and Contours: Automatic Learning of Czech Intonation”, Lecture Notes in Computer Science 5246: 293–301, 2008.

[17] Duběda, T., “Towards an inventory of pitch accents for read Czech”, Slovo a slovesnost 72: 3–12, 2011.

[18] Dombrowski, E., “Semantic Features of Accent Contours: Effects of F0 Peak Position and F0 Time Shape”, Proceedings 15th ICPhS, 1217–1220, Barcelona, 2003.

[19] Kohler, K., “Timing and communicative functions of pitch contours”, Phonetica, 62(2–4): 88–105, 2005.

[20] Niebuhr, O., “Perzeption und kognitive Verarbeitung der Sprechmelodie, Theoretische Grundlagen und empirische Untersuchungen”, Language, Context, and Cognition, Vol. VII, Berlin/New York: deGruyter, 2007.

[21] Niebuhr, O., “On the phonetics of intensifying emphasis in German”, Phonetica 67: 170–198, 2010.

[22] Niebuhr, O. and Zellers, M., “Late pitch accents in hat and dip intonation patterns”, in Niebuhr, O. and Pfitzinger, H. R. [Eds.], Prosdies: context, function, and communication, Berlin/New York: de Gruyter, 2012.

[23] Skarnitzl, R., “Prague Phonetic Corpus: status report”, AUC Philologica 1/2009, Phonetica Pragensia, XII: 65–67, 2010.

[24] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer” [computer program], version 5.3.35. Online: http://www.praat.org/.

[25] Kohler, K. J., “Modelling prosody in spontaneous speech”, in Y. Sagisaka, N. Campbell, and N. Higuchi [Eds.], Computing Prosody, Computational Models for Processing Spontaneous Speech, New York: Springer: 187–210, 1997.

[26] StatSoft, Inc. (2004). STATISTICA [computer program], ver. 7.

[27] Beňuš, Š., Gravano, A. and Hirschberg, J.: “Prosody, emotions, and... 'whatever'”, Proceedings of Interspeech 2007: 2629–2632, 2007.

Automatic Analysis of Emotional Prosody in Mandarin Chinese: Applying the Momel Algorithm

Ting Wang [1,2], *Hongwei Ding*[1,3], *Qiuwu Ma*[1], *Daniel Hirst*[4,1]

[1] School of Foreign Languages, Tongji University, Shanghai, China
[2] Department of Linguistics, University of Pennsylvania, Philadelphia, USA
[3]School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China
[4]Laboratoire Parole et Langage, UMR 7309 CNRS, Aix-Marseille University, France
{2011ting_wang; hongwei.ding; mqw}@tongji.edu.cn; daniel.hirst@lpl-aix.fr

Abstract

Based on the Momel algorithm, a set of acoustic parameters was analyzed automatically on Chinese emotional speech. Global prosodic features were calculated on the sentence level, which showed a concordance with the usual pattern reported in the literature. Local constraints were also considered on the syllable layer. An ANOVA showed that there were interactive effects among emotions, syllable positions and syllable tones on certain parameters. Further more, by examining the pitch movements, no significant difference was found between neutral speech and active emotional speech, which was different from the performance in non-tonal languages. However when reducing the tonal influence by using utterances composed of only tone 1 syllables, this inverse effect disappeared. Hence we posited an interpretation that due to the existence of lexical tone in Mandarin Chinese, the paralinguistic use of pitch movements has been reduced.

Index Terms: emotional prosody; the Momel algorithm; Mandarin Chinese; lexical tones

1. Introduction

Human speech communication conveys not only linguistic information, but also shows the speaker's age, gender, emotions and other paralinguistic cues, among which the importance of emotions in vocal speech has been recognized throughout history. Emotion-specific patterns of acoustic cues have been widely investigated in non-tonal languages, such as in [1], [2] and [3], and showed some common properties across languages. However, the acoustic realization of emotion in lexical tone languages like Mandarin Chinese does not seem to always work the same way. [4] posited that the presence of lexical tones significantly constrains the manipulation of f_0 in emotional speech. Therefore, language-specific attributes, especially in tone language, still need more attention.

In Chinese, the lexical tones and intonation are intertwined as one phonetic representation of the raw f_0 curve conveying both linguistic and paralinguistic functions. How lexical tones and intonation interact with each other is still an unsolved problem. Chao [5] was one of the pioneers who studied Chinese emotional prosody. He pointed out that the emotional intonation depends on the voice quality, stress, phrase pitch and tempo of speech. Yuan et al. [6] proposed that anger and fear are mainly realized by phonation; joy is mainly realized by f_0; whereas sadness is realized by both phonation and f_0. Zhang et al. [7] investigated f_0, duration and short-time amplitude on both sentential level and syllable level. Li et al. [8] investigated emotional intonations by analyzing mono-syllabic utterances. Results showed that the tonal space, the edge tone and the duration differ greatly across 7 emotions. Wang et al. [9] studied the cross-linguistic perceptual patterns of four basic emotions. However, many questions are still waiting to be answered. For instance, the interaction between global intonation and local constraints like tone and position is still not quite clear. Furthermore, we should also find a proper way of modelling the emotional melody to account for microprosody and noise in f_0 extraction.

In this paper, the Momel algorithm was adopted to represent the surface f_0 contour in Chinese emotional speech. Model based acoustic analysis has been conducted on both global layer and syllable base. Finally, the restriction of lexical tone and vocal emotion realization was examined.

2. Method

The emotions used for investigation were four basic emotions including happiness, fear, anger and sadness, supplemented by a neutral state for comparison. Corresponding to the discrete emotion theory [10], these four emotions are among the most commonly postulated basic emotions and are most frequently studied [11]–[13].

2.1. Corpus

Recording materials are 36 sentences. Each sentence contains eight syllables. Target words differing in tones are put at the beginning, middle and end of each sentence respectively. All sentences are proper to elite different emotions under certain scenario. Examples of one group of 12 sentences are given below. The other two groups of 12 sentences have the same syntactic forms but with different contents.

{wang1/liu2/ma3/wei4} *ling2 ming2 wan3 xiang3 hui2 xue2 xiao4.*
"Wang/ Liu/ Ma/ Wei Ling wants to go back to school tomorrow evening."
zhe4 shi4 luo2 min3 de0 ***{xin1/ nan2/ nv3/ jiu4}*** *peng2 you3.*
"This is Luo Min's new/ boy/ girl/ old friend."
ma1 ma1 jiao4 xiao3 ming2 qu4 mai3 ***{mao1/yang2/niao3/lu4}****.*
"Mom asked Xiao Ming to buy a cat/ sheep/ bird/ deer."

The recordings were made in the sound booth at 48 kHz sampling rate with a 16-bit resolution. Two professional actors (one male one female) were recruited from Cinema College of Tongji University. The speakers were asked to act each utterance with each of the four emotional states and in a neutral way as a contrast. Each emotional state was accompanied by a short scenario with a picture. The elicitation scenarios help to minimize the interpretation variations that

may differ from speaker to speaker [14]. In total, we obtained 360 utterances (36 sentences × 2 speakers × 5 emotional states).

2.2. Listening test

Listening tests were conducted to confirm that the intended emotions were accurately decoded. Ten Chinese listeners were asked to choose for each the most suitable emotion among angry, happy, fear, sad and neutral options in a forced-choice task. Finally 303 utterances (72 for anger, 50 for fear, 56 for happiness, 53 for sadness and 72 for neutral) were chosen for further analysis.

2.3. Automatic alignment

For acoustic-related researches, one fundamental but time-consuming step before *real* analysis is the correct segmentation and alignment of the speech with the orthographic transcription. Several tools have been developed to make this laboursome task automatic, such as HTK [15], Julius [16] and the P2FA [17].

Here we chose a recently developed tool, SPPAS [18], to implement automatic phonetisation and alignment of Chinese emotional speech. SPPAS generated four tiers in the TextGrid including *inter-pausal units, words, syllables and phonemes*. For our Chinese speech, only three tiers were used in later analysis. An example is shown below:

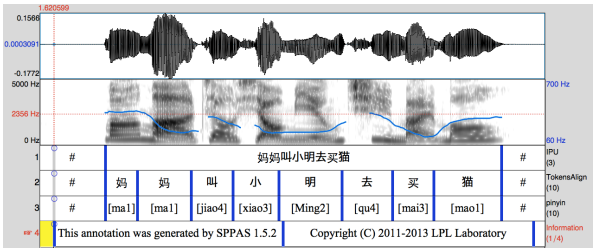


Figure 1: *SPPAS output example.*

2.4. The Momel algorithm

The Momel algorithm [19], [20], short for Modelling Melody, assumes that the raw f_0 curve is the interaction between two components: a global *macroprosodic* component determined by the underlying intonation pattern of the utterance, and a local *microprosodic* component caused by the articulatory constrains of segmental phonemes. Hence the underlying intonation pattern can be modeled as a continuous and smooth f_0 curve. In the Momel algorithm, this continuous curve is realized by a sequence of *target points* interpolated with a quadratic spline function.

The Momel algorithm has been applied to many languages including Standard Chinese, English, French, Korean, Italian, Catalan, Brazilian Portuguese, Venezuelan Spanish, Russian, Arabic and isiZulu [19]. In [21], this algorithm was first used on the lexical tone language, Standard Chinese, which showed its robust capability to model pitch contours.

The Momel algorithm is considered to be theory-neutral, or, better, *theory-friendly* [20], since it works as a phonetic representation of the intonation pattern with respect to speech production and speech perception. It can be compatible with some different theoretical approaches to describe speech prosody, and actually has been used as the first step in deriving the f_0 representations such as in the Fujisaki model [22], ToBI [23] and INTSINT [24].

Since this algorithm optimizes the modelling of speech prosody by taking the raw f_0 curves as input, and outputting the continuous and smooth macroprosodic components, it should be a powerful model to account for the rich local pitch changes in emotional speech.

2.5. Applying Momel algorithm to the corpus

The utterances in the emotional speech corpus were coded using the automatic Momel algorithm. Main steps were described in detail below.

2.5.1. Detect f_0

The quality of f_0 detection is a crucial fundamental step for all the pitch contour models. Unfortunately, most software like Praat will produce errors when extracting f_0 values if only the default parameters are used. Especially when there are a lot of pitch movements and for non-modal speech styles, f_0 extraction using the default parameters is not reliable. The most common errors are due to the inappropriate setting of minimum value and maximum value allowed for the f_0 analysis, that is, Pitch Floor and Pitch Ceiling in Praat.

The Momel plugin [19] on Praat provides an automatic f_0 detection algorithm to generate appropriate Pitch Floor and Pitch Ceiling parameters. This is a two-pass method. In the first pass, default parameters (50Hz for Pitch Floor, 700Hz for Pitch Ceiling) are used to calculate the f_0. Then, the first and third quartiles, named q1and q3, of the f_0 distribution are taken. In the second pass, Pitch Floor and Pitch Ceiling are recalculated respectively by the formula 0.75*q1 and 2.5*q3.

In this paper, we adopted this method by treating our emotional speech corpus with the automatic f_0 detection algorithm in Momel. The outputs were the corresponding *.Pitch* files. Instead of manual correction, this automatic way of detecting f_0 is preferable for the analysis of large speech corpus and is reproducible for further research.

2.5.2. Calculate Momel targets

Once we obtained the *.Pitch* files after the above step, a sequence of pitch target points was calculated by the Momel algorithm, which resulted in the *.PitchTier* files. Figure 2 shows an example of these pitch targets (black circles) detected in a happy utterance from our emotional speech corpus. The red curve is the modeled f_0 contour interpolated quadratically from the targets points with a quadratic spline function.

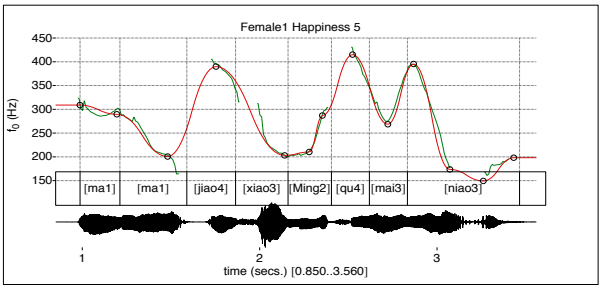


Figure 2: *Automatic Momel algorithm for the utterance "Mom asks Xiao Ming to buy a bird." Raw f_0 (green curves), pitch targets (black circles) and modeled f_0 (red curve).*

For best results, the output *.PitchTier* files containing pitch target points were manually checked using the *Correct Momel* interface. Any erroneous or missing target points were either deleted or added by hand.

3. Model-based acoustic analysis

Melody metrics [19], [20] were extracted by a Praat script based on the Momel outputs on the unit of sentence or syllable, including parameters of duration, intensity, f0 and their variance measures which are regarded as the basic parameters of speech prosody.

3.1. Global prosodic pattern

We first used sentence level as the analysis unit to capture the global prosodic pattern under each emotional state. The averaged results are listed in table 1.

Table 1: *Summary of the acoustic measurements*

	Anger	Fear	Happiness	Sadness	Neutral
Intensity (dB)	68.64	65.54	67.86	63.81	65.00
Speech rate (syl/s)	4.904	4.270	4.105	3.046	3.469
f_0 (octave in z-score)	-.043	.974	.784	-.238	-1.067
f_0 range (octave in z-score)	.261	-.481	.752	-.594	-.075
Mean abs slope (octave/s)	2.447	1.607	2.30	1.500	2.470

A one-way ANOVA with emotions (5 levels) as factor was conducted on the parameters described above.

For intensity, there was a significant main effect of emotions, F (4, 298) = 33.897, p = 0.000<0.05. The mean intensity of anger and happiness were significantly higher than others, while sadness is significantly lower, as indicated by post hoc multiple comparisons. For speech rate, there also appeared main effect of emotions, F (4,298) = 68.730, p = 0.000 < 0.05. The utterances spoken in an angry way were significantly faster, followed by fear and happiness. The speech rate under the sadness state was the lowest, and we only found inner-sentence pauses in sad speech. f_0 and f_0 range values were calculated on interpolated *.PitchTier* files from Momel. All values were first converted from Hz to octave, and then z-transformed with each speaker to reduce the inter-subject variability. A main effect of emotions was found for both parameters, F (4, 298) = 92.831, p = 0.000 < 0.05 and F (4, 298) = 21.436, p = 0.000 < 0.05 respectively. Post hoc tests showed that happiness and fear had significantly higher mean f_0, followed by anger, sadness and neutral. Anger had largest f_0 range, then happiness larger than neutral significantly. Fear and sadness had no significant difference.

Mean absolute slope was measured as absolute difference from the previous pitch point divided the by distance in seconds, which indicated whether there were a lot of pitch movements or not. A significant main effect was shown across different emotions, F (4, 298) = 21.381, p = 0.000 <0.05. In post hoc tests, we found that neutral, anger and happiness had significantly more pitch movements than fear and sadness as illustrated in Figure 3. What drove us to pay attention was that the mean absolute slope of neutral versus anger and happiness in Chinese showed a different pattern to that found in other non-tonal languages in [25], [26]. As reported in the literature, active emotional speech, such as anger and happiness, have much more pitch movements than neutral speech. However, we didn't find such an effect in our Chinese corpus. Either because speakers in our study used a different strategy to express these emotions, or because there's some underlying difference of pitch movements across emotions between Chinese and other non-tonal languages. Detailed discussion will be given later.

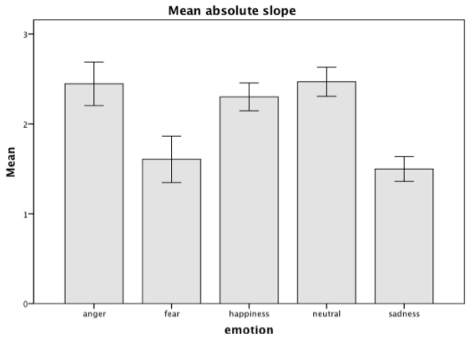


Figure 3: *mean absolute slope on sentence level.*

In summary, as for the sentence level, different emotional speech showed significantly different acoustic patterns, which are largely in concordance with the usual patterns reported in the literature, such as in [2] and [27]. However, the amplitude of pitch movements indicated by mean absolute slope showed different results, which required further analysis.

3.2. Emotions, positions and tones interaction

Although many of the acoustic studies focus on the global realization of prosody, studies on synthesis of emotional speech didn't work like this. A bottom-up method has been adopted. For synthesis purpose, different parameters were usually manipulated on the levels of syllables or even phonemes. Here we should draw attention to the local constraints. In this study, different tones and sentence positions were considered together with emotions on the acoustic analysis of target words/syllables in the utterances.

Table 2 shows the significant level for each of the acoustic parameters analyzed by ANOVA, with emotions (5 levels), positions (3 levels: beginning, middle and end of the sentence) and tones (4 levels: tone 1, tone 2, tone 3 and tone 4 in Mandarin Chinese) as factors. There were significant main effects of emotion, position and tone on most of the parameters except for mean slope. Significant effects of interaction between factors were also found on some parameters. Due to the limitations of space, detailed analysis of these results will be postponed for future publication, instead we focus on f_0 and mean absolute slope.

For mean f_0, the interaction between emotions and positions, emotions and tones were significant, F (8, 243) = 7.985, p = 0.000 < 0.05 and F (12, 243) = 2.519, p = 0.004 < 0.05 respectively. Figure 4 shows the mean f_0 performance in z-scored octave at different tones and emotions.

Across all emotions, tone 1 and tone 4 had higher f_0, which is in line with the fact that in Mandarin Chinese tone 1 and tone 4 begin as high tones. Interestingly, if look at the amplitude difference between tone 1, tone 4 and tone 2, tone 3, we found that compared with neutral speech, this difference is reduced in emotional speech. This phenomenon suggested the possibility that emotions, to some extent, restricted the distinction between the four tones in Chinese. This result seemed to provide an explanation for the unusual performance

of pitch movements described in 3.1. When speaking with emotions, pitch movements were reduced compared to neutral speech in Mandarin Chinese. To further investigate this phenomenon, we did an additional experiment.

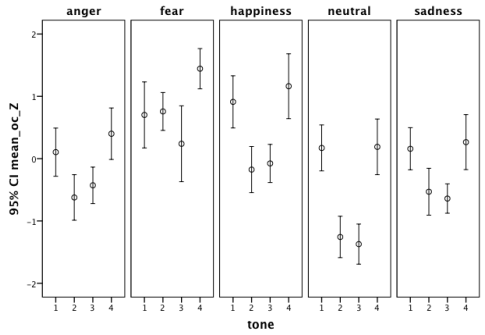


Figure 4: *mean f_0 value (z-scored octave) at different tones and emotions.*

3.3. Lexical tone and vocal emotion restriction

To test the hypothesis above, we recorded four groups, each for one tone, of monotone utterances in Mandarin Chinese using the same female speaker. The utterances are digital strings, which also contain eight syllables for each. The recording procedure was the same as the previous experiment, which yielded 189 utterances. One example is as follow:

yi1 ba1 ba1 san1 qi1 yi1 ba1 qi1.

"One eight eight three seven one eight seven."

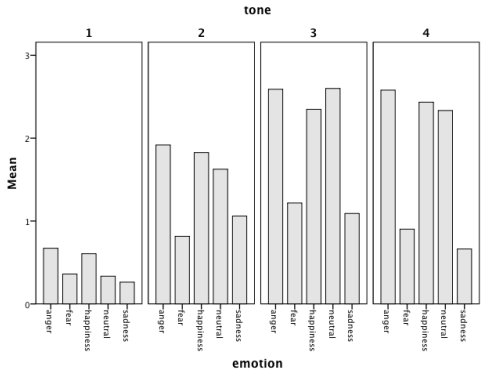


Figure 5: *mean absolute slope in monotone utterances.*

Mean absolute slope was calculated on Momel output as illustrate in Figure 5. In utterances with only tone 2 or tone 3 or tone 4, we found the same pattern as in Figure 4 that, neutral, anger and happiness had significantly more pitch movements than fear and sadness. However, in utterances with tone 1, anger and happiness had significantly more pitch movements than neutral, which was consistent with the pattern in non-lexical tone languages [25], [26]. Since tone 1 in Mandarin Chinese is a level tone (usually represented as 55), the utterances with only tone 1 syllables, to some extent, reduce the tonal influence on intonation compared to other tone combinations, and can be regarded as similar to a non-tonal utterance. From the evidence above, we came to the conclusion that the lexical tone in Mandarin Chinese has an influence on the realization of emotional prosody. It seems that the lexical tone restricts the paralinguistic use of pitch. This point has also been reported from other evidences in [28]. A similar result was also shown for Cantonese [29] which has a richer tone system than Standard Chinese.

4. Discussion and conclusion

Based on the Momel algorithm, a set of acoustic parameters has been analyzed automatically on Chinese emotional speech. Global prosodic features were calculated on the sentence level, which showed a consistency with the usual pattern reported in the literature. Given the local constraints, we further examined the acoustic performance on the syllable level. An ANOVA showed that there were interactive effects among emotions, syllable positions and syllables tone on certain parameters.

Although previous findings such as [30] showed evidence for the existence of universal patterns from vocal characteristics to specific emotions across cultures, the existence of language-specific paralinguistic features were still found in vocal emotion expression. By examining the pitch movements indicated by mean absolute slope, no significant difference was found between neutral speech and active emotional speech, which was quite different from the performance in non-tonal languages. However when reducing the tonal influence by using utterances composed of only tone 1 syllables, this inverse effect disappeared. Hence we posited an interpretation that due to the existence of lexical tone in Mandarin Chinese, the paralinguistic use of pitch movements has been reduced. This result served as a further proof of the finding in [28] and [29].

The automatic phonetic representation of the intonation pattern, with respect to both speech production and speech perception, by the Momel algorithm makes it possible to account for more tonal and non-tonal language comparison in the future. More over, we should take into consideration more local constrains such as narrow focus and tone sandhi.

5. Acknowledgements

The first author benefited from the Scholarship Award for Excellent Doctoral Student granted by Ministry of Education of China, and the scholarship from the China Scholarship Council. This research was also supported by the National Social Science Foundation of China (No.13BYY009) and Innovation Program of Shanghai Municipal Education Commission (No. 12ZS030).

Table 2: *Significance levels of ANOVA for each parameter. [--]:no significance,[*]=p¡0.05,[**]=p¡0.01, [***]= p¡0.001*

	emotion	position	tone	emotion*position	emotion*tone	position*tone	emotion*position*tone
Intensity	***	**	***	***	--	--	--
Duration	***	**	***	***	--	***	--
f_0	***	***	***	***	**	--	--
f_0 range	*	***	***	--	*	*	*
Mean slope	--	--	***	*	--	***	--
Mean abs slope	***	***	***	--	--	--	--

6. References

[1] Murray, I. R., & Arnott, J. L., "Toward the simulation of emotion in synthetic speech : A review emotion", *The Journal of the Acoustical Society of America*, 93, pp. 1097–1108, 1993.

[2] Johnstone, T., & Scherer, K. R., "Vocal communication of emotion", *Handbook of emotions*, 2000.

[3] Scherer, K. R., "Vocal communication of emotion: A review of research paradigms," *Speech communication, 40(1), 227-256*, 2003.

[4] Ross, E. D., Edmondson, J. A., & Seibert, G. B., "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice," *Journal of Phonetics*, 14(2), 283-302 1986.

[5] Chao, Y. R., "Tone and intonation in Chinese", *Bulletin of the Institute of History and Philology*, 4(2), 121-134, 1933.

[6] Yuan, J., Shen, L., & Chen, F., "The acoustic realization of anger, fear, joy and sadness in Chinese", in *INTERSPEECH*, pp. 2025–2028, 2002.

[7] Zhang, S., Ching, P. C., & Kong, F., "Acoustic analysis of emotional speech in Mandarin Chinese", *In International Symposium on Chinese Spoken Language Processing, pp. 57-66*, 2006.

[8] Li, A., Fang, Q., & Dang, J., "Emotional intonation in a tone language: Experimental evidence from Chinese", *ICPhS XVII, Hong Kong*, 2011.

[9] Wang, T., Ding, H., & Gu, W., "Perceptual Study for Emotional Speech of Mandarin Chinese", *in Speech Prosody 2012*, 2012.

[10] Darwin, C., "The expression of the emotions in man and animals". Oxford University Press, 1998.

[11] Ekman, P., "An argument for basic emotions", *Cognition & Emotion*, 6(3-4), 169-200, 1992.

[12] Ekman, P., "Are there basic emotions?" *Psychological Review*, 99(3), 550-553, 1992.

[13] Ekman, P., "Basic emotions", *Handbook of cognition and emotion*, 4, 5-60, 1999.

[14] Pell, M. D., "Influence of emotion and focus location on prosody in matched statements and questions", *The Journal of the Acoustical Society of America*, vol. 109, no. 4, p.1668, 2001.

[15] Young, S. J., & Young, S., "The htk hidden markov model toolkit: Design and philosophy", *Entropic Cambridge Research Laboratory, Ltd*, 1994.

[16] Lee, A., Kawahara, T., & Shikano, K., "Julius---an open source real-time large vocabulary recognition engine", *in EUROSPEECH* 2001, 2001.

[17] Yuan, J., & Liberman, M., "Speaker identification on the SCOTUS corpus", *Journal of the Acoustical Society of America, 123(5), 3878*, 2008.

[18] Bigi, B., & Hirst, D. J., "SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody", *Proc. of Speech Prosody*, 2012.

[19] Hirst, D. J., "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation", *In Proceedings of the XVIth International Conference of Phonetic Sciences*, pp. 1233–1236, 2007.

[20] Hirst, D. J., "The analysis by synthesis of speech melody: from data to models", *Journal of speech Sciences*, vol. 1, no. 1, pp. 55–83, 2011.

[21] Zhi, N., Hirst, D. J., & Bertinetto, P. M., "Automatic analysis of the intonation of a tone language. Applying the Momel algorithm to spontaneous Standard Chinese (Beijing)", in *INTERSPEECH* 2010, 2010.

[22] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", *In Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference, Vol. 3, pp. 1281-1284,* 2000.

[23] Maghbouleh, A., "ToBI Accent Type Recognition", *ISSUES*, 1998.

[24] Hirst, D. J., "La représentation linguistique des systèmes prosodiques: une approche cognitive", *Doctoral dissertation, Aix Marseille 1,* 1987.

[25] Paeschke, A., Kienast, M., & Sendlmeier, W. F., "F0-contours in emotional speech", *Proc. ICPhS*, 1999.

[26] Paeschke, A., & Sendlmeier, W. F., "Prosodic Characteristics of Emotional Speech : Measurements of Fundamental Frequency Movements", *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[27] Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P., "Emotional speech: Towards a new generation of databases", *Speech communication,* 40(1), 33-60, 2003.

[28] Hirst, D. J., "Melody metrics for prosodic typology: comparing English , French and Chinese", *INTERSPEECH 2013*, 2013.

[29] Hirst, D. J., Wakefield, J. & Li, H.T.Y. "Does lexical tone restrict the paralinguistic use of pitch? Comparing melody metrics for French, English, Mandarin and Cantonese". in *Proceedings of the International Conference on the Phonetics of the Languages in China*, Hong Kong, 2013.

[30] Scherer, K. R., Banse, R., & Wallbott, H. G., "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology, 32(1), 76-92*, 2001.

Prosodic Profiles of Social Affects in Mandarin Chinese

Yan Lu[1], Véronique Aubergé[2], Albert Rilliard[3]

[1] GIPSA Lab, CNRS, Stendhal University, Grenoble France;
[2] LIG Lab, CNRS, Grenoble France
[3] LIMSI-CNRS, Orsay, France

Yan.lu@gipsa-lab.grenoble-inp.fr, Veronique.Auberge@imag.fr, albert.rilliard@limsi.fr

Abstract

This work examines the production side of social affects in Mandarin Chinese, with the aim of extracting the more prominent patterns of acoustical variations. Results are then compared to previous perception data obtained on the same expressions. The F_0, intensity and duration characteristics of 76 utterances conveying 19 prosodic attitudes are statistically examined in this study. All attitudes are regrouped into 5 clusters according to their prosodic features. The result of the statistical analysis shows that the prominent differentiation between clusters is mostly related to F_0 and duration parameters; some similarities are noted between the clustering of attitudes from acoustic features and from perceptual confusions obtained in previous experiments; inside each cluster, some attitudes show typical characteristics in F_0 and duration.

Index Terms: prosodic attitudes, social affects, acoustic parameters, Mandarin Chinese

1. Introduction

Since the voice is considered as a carrier of affective signals in human speech, vocal cues, and especially prosodic cues, shall have an important role in the expression of affective nuances, which ensures some interaction functions like situation cues, mental states and processing, intentions, attitudes and emotions. [1] distinguished the socio-affective expressions (or expressions of attitude), which can be voluntarily controlled, from the expressions of emotion, which cannot be. Meanwhile, both emotional expressions and socio-affective expressions are often conveyed by the prosodic variations, which influence significantly the interpersonal interaction and social communication [2].

Many empirical studies demonstrated that people decode the acoustic signal conveying emotional expressions and attitudes with only voice samples (e.g. [3, 4, 5, 6]). On the other hand, many scholars have been engaged in finding out how affective signals are encoded in voice, with special focus on the acoustic measures of emotion encoding (e.g. [7, 8, 9, 10, 11]). The acoustic variables which have been widely measured in the literature are mostly fundamental frequency, energy (or intensity), duration (or speech rate), harmonics, stress, intonation, timbre, etc. Fundamental frequency, intensity and duration are the most classical acoustic parameters used as correlates of prosody [10], and it was commonly accepted that the fundamental frequency play an important role to signal affect, intention, or emotion [12].

Although [13] has claimed that the affective prosody universally exists in every language, there is also no denying that the expression of affective prosody varies from one language to another, and many studies have been conducted for the specific purpose of investigating the prosodic characteristics of affective speech specifically in Chinese (e.g. [14, 9, 11, 15, 16]).

Following the example of these previous studies, we will intend in the present work to identify the main prosodic profile of 19 social affects expressed in Mandarin Chinese by statistically separating them into different clusters. The characteristics of these expressions on F0, intensity and duration parameters, both at the sentence and the syllable levels will be took into consideration. Meanwhile, this work also aims at finding explanations for the perceptual confusions observed between these social affects during previous perception experiment [6], because acoustic proximity is thought to be possibly explicative of certain confusions because they remain important cues for encoding and decoding studies [17].

2. Method

2.1. Corpus of Chinese social affects

Based on research on attitudes in Chinese and other languages [18, 19, 20, 3, 4], 19 Chinese daily encountered attitudes have been selected for our study. The speech corpus of this work contains four sentences performed with these 19 attitudes by one native Chinese female speaker speaking an unmarked standard Mandarin Chinese. The corpus has been perceptually validated in [6], where almost all attitudes have been recognized over the chance level (except “confidence”). The sentences analyzed in this paper received the best average recognition score across all attitudes in each length (monosyllable, disyllable, 4-syllable and 9-syllable). These sentences can be considered as the most representative of a prototypical expression of the targeted attitude. Table 1 lists the sentences composing the corpus and Table 2 presents the 19 Chinese attitudes which will be analyzed in the present work.

Table 1. *Sentences composing the corpus.*

Chinese	Pinyin	English
树	shu4	tree
放学	fang4 xue2	School is over.
张医生来	Zhang1 yi1 sheng1 lai2	Doctor Zhang will come.
王医生他三姑妈休假	Wang2 yi1 sheng1 ta1 san1 gu1 ma1 xiu1 jia4	Doctor Wang’s third aunt will go on holiday.

Table 2. *Social affects and their abbreviation.*

Social affects and abbreviation	
admiration(ADMI)	authority(AUTH)
confidence(CONF)	contempt(CONT)
declaration(DECL)	disappointment(DISA)
doubt(DOUB)	irritation(IRRI)
infant-directed speech(IDS)	intimacy(INTI)
irony(IRON)	neutral surprise(NEU-S)
negative surprise(NEG-S)	obviousness(OBVI)
politeness(POLI)	positive surprise(POS-S)
question(QUES)	resignation(RESI)
seduction(SEDU)	

2.2. Measurements of prosodic features

According to some studies on emotional expressiveness in Chinese and other tonal language, the global intonation form of sentence is often linked to its expressive function [16; 21] and the variation of the initial and final movements of F_0 contour is more significant in characterizing the F_0 contour of different attitude than the movement in middle [22]. Therefore, we will observe in this work both global prosodic characteristics of social affects in sentence level and the specific cues at the beginning and end of sentence.

The majority of the values of prosodic parameters were automatically extracted from the 76 stimuli using the PRAAT software. Each sentence was previously hand-labeled at the phonemic level. The prosodic parameters measured are the fundamental frequency, the syllabic duration, and the intensity. For each sentence, six measures in F_0 were considered: mean F_0, standard deviation of F_0, maximum and minimum values of F_0 of the sentence, mean values of the first and the last syllable for F_0. Similarly, the durations of the first and the last syllable and of sentence were also measured, as well as the mean intensity of sentence. F_0 values were extracted by cross-correlation method and are expressed in semi-tone (with reference to 1Hz). The intensity values are expressed in decibel (dB). The decimal logarithm of duration (in millisecond) is used here [14]. Two other parameters were calculated manually in Excel:

- F_0 range (in semitones): the difference between the maximum and the minimum value of F_0 [16].
- F_0 slope (semitones/s): is here defined as the direction and rate of F_0 change. It is the difference of the mean F_0 of the last syllable to the mean F_0 of the first syllable divided by sentence duration [14].

The means for selected acoustic parameters of vocal utterances conveying 19 attitudes were calculated before being statistically analyzed. Table 3 shows the 10 prosodic variables extracted from the audio samples.

Table 5. *Matrix of rotated components.*

	Component 1	Component 2
F_0_range	0.95	0.17
F_0_mean	0.93	0.05
F_0_std	0.83	0.20
F_ F_0_mean	0.97	0.07
L_ F_0_mean	0.95	0.07
F_dur	0.20	0.90
L_dur	0.70	0.64
Sentence_dur	0.28	0.94
F_0_slope	-0.42	0.71
Intensity_mean	0.67	0.01

Table 3. *Prosodic parameters calculated on the acoustic measures of F_0, intensity, and duration.*

Parameter	Unit	Abreviation	Acoustic measure
F_0 range	semitones	F_0_range	F_0
F_0 register of sentence	semitones	F_0_mean	F_0
F_0 variation	semitones	F_0_std	F_0
F_0 register of the first syllable	semitones	F_ F_0_mean	F_0
F_0 register of the last syllable	semitones	L_ F_0_mean	F_0
F_0 slope	semitones/s	F_0_slope	F_0
Intensity register of sentence	dB	Intensity_mean	Intensity
Duration of the first syllable	$\log_{10}$(ms)	F_dur	Duration
Duration of the last syllable	$\log_{10}$(ms)	L_dur	Duration
Duration of sentence	$\log_{10}$(ms)	Sentence_dur	Duration

3. Results

3.1. Hierarchical clustering of attitudes based on principal component analysis

Combining a principal component analysis and an agglomerative hierarchical cluster analysis, this method has the advantage of clustering the individuals with less noise [23]. Therefore, a principle component analysis (hereafter referred to as PCA) was performed as a preprocessing step before the cluster analysis. After having measured the sampling adequacy with KMO & Bartlett's test (KMO = 0.614; Bartlett's Sig. < 0.001), the PCA was carried out on the dataset with 19 individuals (attitudes) and 10 prosodic variables. The measures were standardized during the procedure. The SPSS software was used to carry out the analysis.

The result of the PCA is presented in Table 4: two principal components were extracted, which explained cumulatively almost 82% of the variance. Table 5 shows the matrix of components after rotation (the "varimax rotation" method was used here). The main results indicate that the first principle component is more linked to F_0 parameters, intensity register and the duration of the last syllable, while the second one is linked to the duration parameters and the slope of F_0.

Table 4. *Proportion of variance explained by the first four components of PCA over all prosodic parameters.*

		1	2	3	4
Initial Eigenvalues	% of variance	59.00	22.95	9.03	5.23
	Cumulative %	59.00	81.96	90.98	96.21

The factor scores of the 19 attitudes on the two first components were then used as a new variable on which was performed a hierarchical cluster analysis based on an agglomerative procedure. The Ward's minimum variance criterion was used to calculate the distance between clusters, while the squared Euclidean distance of observations was

defined as their distance. To determine the optimal number of clusters from the hierarchical tree, we referred to the "elbow criterion" based on the variance explained by each clusters [24]. Figure 1 shows the dendrogram of the hierarchical clustering and the graph of the between-inertia in function of the number of clusters on the top right of corner. Observing the inertia graph, we thought it would be reasonable to consider 5 clusters.

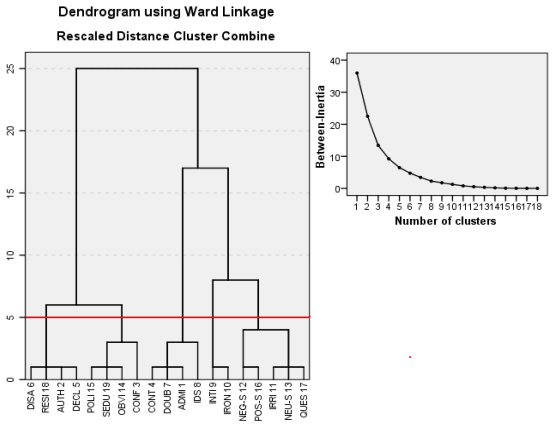


Figure 1. *Dendrogram resulting from a hierarchical clustering of 19 social affects. The red line marks the clusters resulting in the observation of inertia reduction presented in the graph on the top right corner.*

The last step of this analysis consists in visualizing the 5 clusters' position on the two principal components, with the aim of highlighting how these attitudes disperse on these two dimensions, and of looking at how the acoustic dimensions allow separating or grouping them together.

As showed in Figure 2, the majority of attitudes are projected along the axis in the plot, while certain ones are in the extremity of the axis: "positive surprise", "negative surprise" and "neutral surprise" are marked by their highest values on dimension 1, while "disappointment", "resignation" and "confidence" the lowest values; "infant-directed-speech" shows the highest values in dimension 2, at the opposite of "intimacy". The coordinates of some attitudes almost overlap on the graph, and it is the case of "positive surprise" and "negative surprise", "disappointment" and "resignation", "seduction" and "politeness", "authority" and "declaration". This proximity of distance between two attitudes implies the similarity of their prosodic features. That may be an important cue to explain some confusion patterns observed during the perceptual experiment.

As regards the 5 clusters obtained in hierarchical cluster, they are well separated on the two principal dimensions (cf. Figure 2):

- Cluster 1 groups "admiration", "infant-directed speech", "contempt" and "doubt". These attitudes show some high values in both F_0 and duration parameters.
- Cluster 2 groups "authority", "declaration", "resignation" and "disappointment". On the contrary to the first group, they have low values in both F_0 and duration parameters.
- Cluster 3 groups "obviousness", "seduction", "politeness" and "confidence". These attitudes show some similarities with the members of cluster 2 in F_0 phenomenon, but differ from them with their higher values in the dimension of duration.
- Cluster 4 groups "irony" and "intimacy", which are quite separated from other attitudes and characterized by their extremely low values in duration parameters.
- Cluster 5 groups "positive surprise", "negative surprise", "neutral surprise", "question" and "irritation". They look more similar to the first group in F_0 parameters, but quite different in duration dimension.

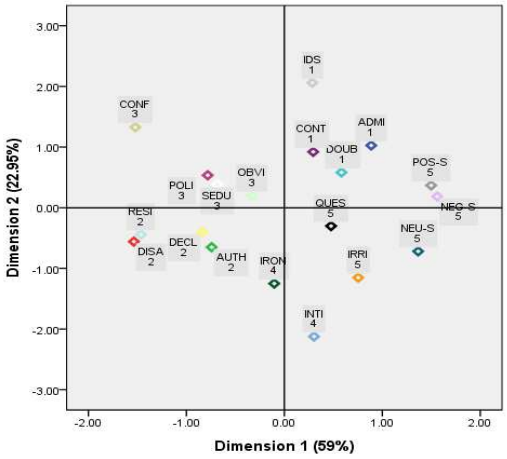


Figure 2. *Representation of the 19 attitudes clustered in 5 groups on the first two principal components of the PCA. Numbers under labels refer to the cluster they belong to.*

3.2. Differentiation of attitudes clusters

The analysis of the hierarchical clustering gives a global impression of the dispersion of attitudes and of their clustering according to their prosodic features. But it is also necessary to observe the prosodic characteristics of each cluster — in other words, how do the clusters differentiate prosodically one from another. Consequently, another analysis was done by comparing the means of each cluster across all variables. The results are detailed in Figure 3:

(1). For the F_0 measures (F_0 range, F_0 register of sentence, mean F_0 of the first and the last syllable, F_0 variation and F_0 slope), cluster 1 and 5 have higher values, while cluster 2 and 3 have the lower ones. Cluster 5 displays the highest values on F_0 range, F_0 register of sentence, mean F_0 of the first and the last syllable; cluster 1 has the highest value in F_0 variation; cluster 2 has the lowest values on almost all F_0 variables except F_0 slope. Cluster 4 has the highest negative value of F_0 slope. All clusters show slight differences in F_0 register of sentence. On the other hand, it can be found that the difference between cluster 1 and 5 is linked to the duration of the first syllable and of the sentence, with higher values for cluster 1 than for cluster 5. Cluster 3 and 2 also present some differences in duration measures, with higher values for cluster 3 than for cluster 2, and they are especially different in duration of the last syllable.

(2). For duration measures (duration of the first syllable and the last syllable, sentence duration), cluster 1 shows the highest values, while cluster 2 and 4 shows the lowest ones. Cluster 3 and 5 are in between and the latter has a very high value of duration of the last syllable just next below cluster 1. It is worth noting that cluster 4 and 2 does not show remarkable difference in duration measures neither in F_0 measures. Regarding the differences between cluster 3 and 5, we found that their differences mostly concern F_0 measures: the values of cluster 5 are apparently higher than that of cluster 3.

(3). Concerning intensity measure (intensity register of sentence), no clear patterns of differences between clusters were found.

4. Discussion and Conclusions

In the present work, we investigated how the attitudes could be clustered according to their prosodic features by observing the acoustic parameters of F_0, intensity and duration. The corpus of attitudinal speech contained 4 sentences of different length conveying everyone 19 Chinese attitudes and it had been perceptually validated during a previous experiment. Although the present study is still preliminary and the data involved is not large, some interesting and valuable results have been obtained.

First of all, the result of the hierarchical clustering ran on principal components gives a separation of attitudes into two main groups (cf. Figure 1) and this separation is basically based on the characteristics of fundamental frequency of attitudes (cf. Figure 2). One group is composed of attitudes which have high pitch level (e.g. “positive surprise” and “admiration”) and large pitch span (e.g. “question” and “doubt”); the other is composed of attitudes whose pitch level and pitch span are lower and narrower, (e.g. “declaration” and “politeness”). Such a higher and wider pitch span for surprise and admiration may be related to hypothesis of the “effort code” postulated by Gussenhoven [25]. A similar separation between the studied attitudes has been found in the perception experiment where the main distinction was observed between “assertive” attitudes and “interrogative” attitudes [6]. This observation implies an important role of F_0 in affective expression decoding. The attitudes regrouped in clusters 2 and 3 appear more homogeneous in terms of F_0 features, and that could help us to understand the perceptual confusions observed between “declaration” and some other “assertive” attitudes like “politeness” and “obviousness”, as well as the confusion between “disappointment” and “resignation”. Of course, some perceptual similarities remain unexplained with only these acoustic cues: for example, the confusions between “infant-directed speech” and “seduction”, “authority” and “irritation”. In these cases, one analysis of voice quality seems necessary, because voice quality, as the fourth dimension of prosody [26], is an important aspect of the affective expression.

The F_0 slope did not exhibit important difference across clusters, except for cluster 4 (“irony” and “intimacy”) which is distinguished by its highest value. All of the clusters are homogeneous in intensity register. Hence, we can summarise the salient prosodic features of each attitude cluster essentially in function of their F_0 and duration profile: the cluster 1 (“admiration”, “infant-directed speech”, “contempt” and “doubt”) shows a very long duration; on the contrary, the cluster 4 (“irony” and “intimacy”) shows the lowest duration values and the highest for F_0 slope. The cluster 5 (“positive surprise”, “negative surprise”, “neutral surprise”, “question” and “irritation”) is typically marked by large F_0 range and high F_0 level, while the cluster 2 (“authority”, “declaration”, “resignation” and “disappointment”) by lower F_0 values. Cluster 3 (“obviousness”, “seduction”, “politeness” and “confidence”) is characterized by low F_0 values and in particular the lowest F_0 slope.

Some differences inside clusters also deserve our attention: “infant-directed speech” shows the longest duration and “intimacy” the shortest duration; positive, negative and neutral surprises are marked by their high F_0 values while “resignation” and “disappointment” show the lowest F_0 values of all attitudes; “confidence” differs from the other attitudes of cluster 3 by lower F_0 values and a longer duration.

Another acoustic analysis about voice quality of the same audio samples is under way in order to investigate the potential influence of voice quality to the perception of the attitudes in question.

5. Acknowledgements

This work is partly supported by the Major Program for the National Social Science Fund of China (13&ZD189).

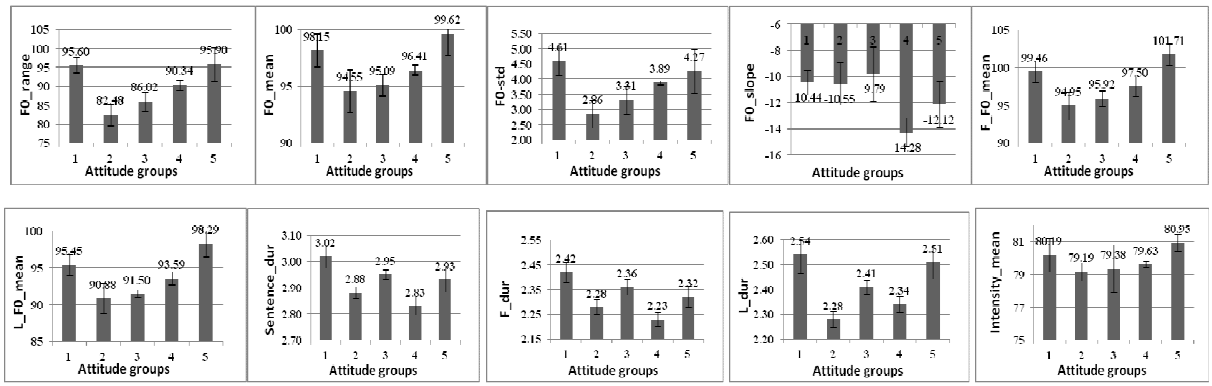


Figure 3. *Mean values of the five clusters across all prosodic parameters. Numbers represent the clusters; bars the mean values for a given parameter.*

6. References

[1] Aubergé, V., "A Gestalt Morphology of Prosody Directed by Functions", Speech Prosody 2002 Proc., 151-154, Aix en Provence, France, 2002.

[2] Scherer, K.R., "Interpersonal expectations, social influence and emotion transfer", in P.D. Blanck [Ed], Interpersonal expectations: Theory, research, and application, 316-336, Cambridge University Press, Cambridge, 1993.

[3] Shochi, T., Rilliard, A., Aubergé, V. and Erickson, D., "Intercultural Perception of English, French and Japanese Social Affective Prosody". in S. Hancil [Ed], The role of prosody in Affective Speech, 31-59, Linguistic Insights 97, Peter Lang AG, Bern, 2009.

[4] Mac, D. K., Aubergé, V. Rilliard, A., and Castelli, E., "How prosodic attitudes can be recognized and confused: Vietnamese multimodal social affects", SLTU, Penang, Malaysia, 2010.

[5] de Moraes, J. A., Rilliard, A., Alberto, B. and Shochi, T., "Production and perception of attitudinal meaning in Brazilian Portuguese". Speech Prosody 2010 Proc., Chicago, USA, 2010.

[6] Lu, Y., Aubergé, V. and Rilliard, A., "Do You Hear My Attitude? Prosodic Perception of Social Affect in Mandarin", Speech Prosody 2012 Proc., 685-688, Shanghai, China, 2012.

[7] Deller, J. R., Proakis, J. G. and Hansen, J. H. L., "Discrete-time processiong of speech signals", Macmillan Pub. Co., New York, 1993.

[8] Borden, G.J. and Harris, K. S., "Speech science primer: Physiology, acoustics and perception of speech (3rd ed.), Williams & Wilkins, Baltimore, 1994.

[9] Yuan, J., Shen, L. and Chen, F., "The acoustic realization of anger, fear, joy and sadness in Chinese", ICSLP 2002 Proc., 2025-2028, Denver, USA, 2002.

[10] Sherer, K.R. and Ellgring, H., "Multimodal Expression of Emotion, Affect Programs or Componential Appraisal Patterns?", Emotion, Vol. 7, No. 1, 158-171, 2007.

[11] Zhang, S., Ching, P. C. and Kong, F., "Acoustic Analysis of Emotional Speech in Mandarin Chinese", ISCSLP 2006 Proc., 57-66, Singapore, 2006.

[12] Ohala, J.J., "The frequency codes underlies the sound symbolic use of voice pitch", In L. Hinton, J. Nichols & J.J. Ohala [Ed], Sound symbolism, 325-347, Cambridge University Press, Cambridge, 1994

[13] Crystal, D., "The English tone of voice", St Martins Press, New York, 1976.

[14] Ross, E. D., Edmondson, J. A. and Seibert, G. B., "The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice", Journal of Phonetics (1986) 14, 283-302, 1986.

[15] Gu, W. and Lee, T., "Quantitative Analysis of *F*0 Contours of Emotional Speech of Mandarin", ISCA 2007 Proc., 228-233, Bonn, Germany, 2007.

[16] Lin, H. and Fon, J., "Prosodic and Acoustic Features of Emotional Speech in Taiwan Mandarin", Speech Prosody 2012 Proc., 450-453, Shanghai, China, 2012.

[17] Johnstone, T. and Schere, K., "Vocal Communication of Emotion", in M. Lewis and J. Haviland [Ed], Handbook of Emotions, 220-235, Guilford Press, New York, 2000.

[18] Gu, W., Zhang T., and Fujisaki H., "Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes", Proceedings of Interspeech 2011, Firenze, Italy, 1069-1072, 2011.

[19] Fónagy, Y., "La Vive Voix", Paris, Payot, 1991.

[20] Diaferia, M. L., "Les Attitudes de l'Anglais : Premiers Indices Prosodiques". Master thesis in Cognitive Science. National Polytechnique Institut of Grenoble, France, 2002.

[21] Li.A., Fang, Q. and Dang. J., "Emotional Intonation in a Tone Language: Experimental Evidence From Chinese", ICPhS XVII, Hong Kong, 17-21, 2011.

[22] Mac. D.K., "Génération de parole expressive dans le cas de langues à tons", PhD Thesis, Grenoble University, 2012.

[23] Husson, F., Josse, J. and Pagès, J., "Principal component methods – hieraichical clustering – partitional clustering: Why would we need to choose for visualizing data?", Technical report – Agrocampus Ouest. Online: http://foactominer.free.fr/docs.HCPC_husson_josse.pdf, 2010.

[24] Soni Madhulatha, T., "An Overview on Clustering Method", IOSR Journal of Engineering, vol 2 (4), 719-725, 2012.

[25] Gussenhoven, C., "The phonology of tone and intonation", Cambridge Univ. Press, Cambridge, 2004.

[26] Campell, N. and Mokhtari, P., "Voice quality: The 4th prosodic dimension", The 15th International Congress of Phonetic Sciences Proc, 2417–2420, 2003.

Prosodic cues for emotion: analysis with discrete characterization of intonation

Houwei Cao[1], *Štefan Beňuš*[2,3], *Ruben C. Gur*[1], *Ragini Verma*[1] *and Ani Nenkova*[1]

[1] University of Pennsylvania, United States
[2] Constantine the Philosopher University, Nitra, Slovakia
[3] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia

Houwei.Cao@uphs.upenn.edu, sb513@nyu.edu, gur@mail.med.upenn.edu,
Ragini.Verma@uphs.upenn.edu, nenkova@seas.upenn.edu

Abstract

In this paper we study the relationship between acted perceptually unambiguous emotion and prosody. Unlike most contemporary approaches which base the analysis of emotion in voice solely on continuous features extracted automatically from the acoustic signal, we analyze the predictive power of discrete characterizations of intonations in the ToBI framework. The goal of our work is to test if particular discrete prosodic events provide significant discriminative power for emotion recognition. Our experiments provide strong evidence that patterns in breaks, boundary tones and type of pitch accent are highly informative of the emotional content of speech. We also present results from automatic prediction of emotion based on ToBI-derived features and compare their prediction power with state-of-the-art bag-of-frame acoustic features. Our results indicate their similar performance in the sentence-dependent emotion prediction tasks, while acoustic features are more robust for the sentence-independent tasks. Finally, we combine ToBI features and acoustic features together and further achieve modest improvements in sentence-independent emotion prediction, particularly in differentiating fear and neutral from other emotion.
Index Terms: ToBI, emotion, automatic prediction

1. Introduction

Despite clearly perceived connection between emotional meanings and prosody, a satisfactory model linking the two has been elusive. Both prosody and emotion are complex phenomena and the adopted framework for their representation vastly influences the interpretability of discovered relationships between the two. The main question for prosody analysis in emotional speech research is how to represent the melody contour. Most research relies on continuous features extractable automatically from the acoustic signal for studying the relationship between emotions and prosody [10]. Discrete representations of prosody in the ToBI [20] or Tilt representations [21] on the other hand mark perceptually salient properties of the utterance tune.

Given that the nature of the precise mapping between the underlying phonological representation of intonation and its phonetic implementation is not known, such labeling might tap into a different type of information regarding the relationship between emotion and prosody. Mozziconacci [14] reviewed several studies including Scherer et al. [19] and Mozziconacci [13] arguing that both the representations of F0 contour using models traditionally considered phonological as well as those representing phonetic implementation of pitch in terms of levels and ranges, offer independent, and possibly additive information for the perception of emotions. Furthermore, Liscombe [12] explored the usefulness of categorical intonation labels in emotion classification in a subset of the EPSAT corpus [11], in which actors read 4-syllable semantically neutral phrases (numbers) in 15 emotions. Ten of the emotions balanced for valence (angry, anxious, bored, confident, encouraging, friendly, frustrated, happy, interested, and sad) plus a neutral utterance from 4 speakers (N=44 utterances) were then subsequently selected for rating of perceived level of emotion by 40 subjects. The relationship between ToBI labels and perceived emotion was then analyzed. Liscombe found a significant effect of pitch accent type on emotion rating for confident, happy, interested, friendly, and bored with L+H* accent showing the greatest disambiguation potential favoring the first five emotions and in general positive valence and high activation while disfavoring boredom. A plateau contour (H*H-L%) was associated with boredom and, in general, H-L% boundaries tended to be associated with negative affect while low boundaries (L-L%) were not.

In addition to the potentially relevant role of phonological description of contour, some studies on very short phrases have shown that knowledge gained from phonological coding of F0 is comparable to information obtained from phonetic features simulating this coding. For example, Benus et al.[4] reported correlation between the backchannel function of cue words and rising F0, which was reflected both in linguistic ToBI labels (H-H% boundaries) and in (stylized) pitch slope extracted over the entire token, its second half, or the last 200ms. These findings suggest that ToBI-like labels offer a feasible description of F0 contours for general analyses, as well as for data with low-quality audio signal.

Our interest in categorical representations of emotion in voice is also motivated by the huge success of such an approach in facial analysis of emotion. For emotion recognition in face, Darwin proposed that a cluster of discrete facial configurations (action units) (such as nostrils raised, mouth compressed, furrowed brow, eyes wide open) need to occur to facially express and interpret emotions [23]. His ideas were later further developed and widely promoted in Ekman's influential work [24], leading to the development of accurate systems for automatic detection of action units on the face [25]. Our interest is in asking if discrete "action units" in voice may similarly be related to emotion expression. A long-term goal would be to develop systems to detect such action units in voice and use them as the basis for emotion prediction, possibly in combination with standard low-level acoustic descriptors.

In this paper we present experiments on a dataset of 433 utterances conveying five basic emotions and manually annotated prosody in the ToBI framework. We analyze the distribution

of discrete prosodic labels in the emotion classes and find several prosodic markers of emotion. Then we present a series of machine learning experiments based on discrete representations of prosody, which have not been done in prior work where the dataset was usually too small to allow for learning. Finally we compare the predictive power of categorical ToBI features and conventional acoustic features, and further combine these two types of features together.

2. Emotional Data

Our data contains recording of emotional utterances produced by 91 professional American actors acting pre-selected sentences under the supervision of a director. The actors were asked to act out a given sentence in a specific target emotion until the director's approval. The dataset contains examples of six basic emotions (*anger, disgust, fear, happy, neutral, sad*) [8] on the following three sentences: [TAI] *The airplane is almost full.*; [TIE] *That is exactly what happen.*; [ITH] *I think I have a doctor's appointment.*

Each of the recorded utterances was classified by ten subjects as expressing one of the possible emotions. The dataset we analyze in this paper contains only utterances which were clearly recognized as the intended emotion by more than five of the of ten raters. We aimed to select 25 examples for each of the six emotions, for each sentence. If more than 25 candidates remained after the perception test, we selected the 25 examples with the highest human recognition rate. Otherwise all validated clear utterances were selected.[1]

3. Annotation of ToBI Labels

All stimuli for a given carrier sentence were randomized across emotions and no information about the emotion label was preserved. They were labeled by an experienced qualified annotator with linguistic and phonetic background, using the ToBI framework [3, 2] within Praat [5].

ToBI labels encode the underlying phonological representation of an utterance primarily in terms of perceived pitch targets (H)igh and (L)ow and disjunctures between words (breaks 0-4 from minimal to strong). Perceptually prominent syllables, primarily due to pitch excursions but also lengthening and intensity, are associated with pitch accents that could consist of single tonal targets (H, !H*, L*), or bi-tonal combinations, most commonly L+H*, L*+H, H+!H*; !H represent a target downstepped from a preceding H target, and "*" corresponds to the tone aligned with the stressed syllable. To provide a rough estimate of pitch range, ToBI includes HiF0 label that indicates the point of the highest F0 in a given intermediate or intonational phrase. For prosodic chunking, breaks 0 and 1 correspond to regular fluent word transitions, 2 to a perceived disjuncture with no salient tonal marking, 3 marks an intermediate phrase associated with H-, L-, or !H- targets, and 4 signals the strongest disjuncture corresponding to the intonational phrase boundary representing a combination of the three phrase accents, marked with dashes and listed above, and a L% or H% tone: H-H%, L-H%, etc. Only minimal adjustments to the ToBI guidelines were employed. Regarding the break indices, diacritics describing uncertainty and disfluency were not used since the nature of the data elicitation minimized these phenomena and we wanted to mitigate data sparsity. Additionally, break 0 was not used. The full set of tonal events was employed.

[1] Specifically there were 23 sad samples of TAI, 19 happy for TIE, 24 disgust and 17 sad for ITH. All other emotions and sentences are represented by 25 different stimuli each.

4. Pitch events across emotions

We now turn to our utterance-level analysis of categorical prosody cues for emotional speech. We consider five sets of ToBI features: types of breaks, intermediate phrase accents, boundary tones in the end of the utterances, pitch accents, and the position of HiF0 in the utterances. For each set of features, we compute the distribution of possible pitch events across different emotion. The results are shown in Table 1. For ease of comparison and interpretation, the first column lists the distributions on *neutral* speech, while the last column gives the average across six emotions. These can be interpreted as benchmarks, providing an insight on how speech associated with a particular emotion differs from neutral speech and the overall average.

The table shows that different emotion classes have different distribution of ToBI features. *Anger* and *disgust* utterances are more likely to contain salient disjunctures between words as shown by higher frequency of 2 and 3 breaks. In particular, *anger* and *disgust* utterances have higher occurrence of !H-L% and L- in the middle of utterance, respectively. Similarly, *fear* and *sad* are characterized by increase in the use of H-L%. *Neutral* utterances on the other had contain only H- intermediate phrasal tones. Hence, emotions with negative valence and high activation tend to be associated with utterances 'chunked' into more prosodic units despite their relatively short duration. This finding is in line with findings that perceived negativity of the word 'whatever' correlated with a prosodic boundary between the first two syllables [4]. Similarly, plateau boundary tones correlate with negative valence, which corroborates the findings discussed in section 1 [12].

Most of the examples in our data are uttered with declarative ending of L-L% final boundary tones. The exceptions are *fear* and *happy* and to a lesser extent also *sad*. They have fewer declarative endings of L-L% and are characterized by a somewhat higher rate of !H-L% and H-L% boundary tones. The association of !H-L% boundary with a positive valence *happy* has not been previously reported and is connected to emotional meaning of H*!H-L% contour that we discuss later.

Compared with *neutral*, emotional speech has more accented words, particularly for *anger*. We can also see the differentiation of emotions on pitch accent distributions. Apart from *neutral*, all other emotions have less occurrence of downstepped (!H*) pitch accents. *Fear* and *sad* have extremely high frequency of H* accents, while *happy, disgusted* and *fear* are associated with the use of L+H* accents. The L+H* association with *happy* and *fear* is a novel surprising finding given the observation of Liscombe's dataset [12] in which mostly positive valence emotion were associated with this accent.

Finally, emotions are very different in terms of the placement of the highest F0 in the prosodic phrase. For example, most of *neutral* and *disgust* utterances start with high pitch, while *fear* and *happy* utterances tend to have highest F0 in the end of the utterances. This HiF0 placement in these two emotion is another indication for the special status of extra-high last pitch accent associated with (L+)H*!H-L% contour for these two emotions, which we discuss later.

In addition to the analysis of individual ToBI labels, we also explore bi-grams of pitch accents. These capture the intonation contour of two concatenated accented words. Final boundary tones were also involved in the bi-gram analysis. Table 1 also lists the most frequent bi-grams in terms of pitch accents and

Table 1: *Distribution of ToBI labels on neutral (first columns) and emotional speech for different types of pitch events and pitch event bigrams*

	NEU	ANG	DIS	FEA	HAP	SAD	Ave.
breaks (%)							
b1	76.0	68.0	69.8	79.3	78.6	74.1	74.2
b2	2.0	7.2	6.6	0.5	0.5	0.9	3.0
b3	3.4	3.7	3.3	0.7	2.4	3.0	2.8
b4	18.6	21.1	20.3	19.5	18.5	22.0	21.0
phrasal tones – intermediate (%)							
proportion of utts have intermediate phrasal tones							
prop.	20.0	34.7	25.7	8.0	14.5	29.2	21.9
L-L%	0.0	3.8	15.8	0.0	0.0	10.5	6.3
!H-L%	0.0	26.9	5.3	0.0	10.0	0.0	9.5
H-L%	0.0	0.0	10.5	33.3	0.0	26.3	9.5
H-H%	0.0	15.4	0.0	16.7	0.0	10.5	7.4
H-	100.0	38.5	26.3	33.3	60.0	15.8	43.2
!H-	0.0	7.7	0.0	16.7	10.0	26.3	9.5
L-	0.0	7.7	42.1	0.0	20.0	10.5	14.7
boundary tones – end of utts (%)							
L-L%	96.0	97.3	82.4	64.0	68.1	72.3	80.4
!H-L%	0.0	2.7	8.1	21.3	20.3	6.2	9.7
H-L%	1.3	0.0	6.8	10.7	5.8	15.4	6.5
L-H%	2.7	0.0	1.4	2.7	4.3	1.5	2.1
H-H%	0.0	0.0	0.0	1.3	1.4	1.5	0.7
pitch accent (%)							
proportion of accented words							
prop.	50.6	56.8	51.4	53.5	51.5	51.8	52.6
H*	45.1	57.1	47.3	71.5	53.8	66.1	56.8
!H*	35.3	21.6	20.7	12.1	12.3	16.1	19.8
L+H*	9.8	20.4	22.7	11.6	25.7	5.7	16.2
L*	2.5	0.0	4.9	0.0	3.6	4.0	2.4
H+!H*	7.4	0.9	3.9	4.7	4.6	6.9	4.6
occurrence of HiF0 at different position of utts (%)							
begin	73.9	46.5	67.8	38.8	23.7	60.3	52.4
middle	18.2	32.3	21.8	17.5	25.0	16.7	22.2
end	8.0	21.2	10.3	43.8	51.3	23.1	25.4
bigrams - pitch accent (top 10 tokens) (%)							
H*, !H*	28.1	18.5	13.3	13.6	8.3	19.3	16.9
H*, H*	15.6	28.4	17.8	50.7	32.6	31.1	29.4
!H*, !H*	13.3	3.1	4.4	2.1	0.8	2.5	4.4
H*, H+!H*	7.4	0.6	3.7	4.3	1.5	6.7	4.0
L+H*, !H*	7.4	8.6	10.4	2.1	9.1	0.0	6.3
!H*, H*	3.7	9.3	7.4	7.1	3.8	8.4	6.6
L+H*, H*	3.0	9.9	4.4	4.3	6.1	2.5	5.0
H*, L*	1.7	0.0	5.0	0.0	0.0	2.5	1.5
H*, L+H*	0.7	7.4	8.1	5.0	7.6	0.8	4.9
L+H*,L+H*	0.7	4.9	5.2	3.6	6.1	0.0	3.4
bigrams - pitch accent with final boundary tones (top 5) (%)							
H*, L-L%	21.3	50.7	29.7	46.7	30.4	38.5	36.3
!H*,L-L%	56.0	37.3	23.0	6.7	18.8	15.4	26.6
L+H*,L-L%	1.3	5.3	1.4	8.0	13.0	3.1	5.30
H*, !H-L%	0.0	1.3	4.1	14.7	18.8	6.2	7.40
H+!H*,L-L%	12.0	2.7	2.7	2.7	0.0	9.2	4.80

the combination of a pitch accent and a boundary tone.

Compared with the analysis of individual ToBI features, the bi-gram features give us insights of more global and contextual intonation patterns associated with emotions. For example, *fear* can be characterized by increased occurrence of two adjacent H* accents, *happy* and *fear* are associated with pitch accent of H*, followed by down-stepped !H-L% boundary in the end, while *anger* utterances are more associated with H* followed by low (L-L%) boundary. In general, bigrams support the observation about the tendency for avoiding down-stepped pitch accents in emotion speech compared to emotionally neutral speech. One note regarding H*!H-L% contour is that the target for H* pitch accent was commonly extremely high, which gave a particular contour especially for *happy* utterances. Since standard ToBI does not have a dedicated label for this situation, we could not determine if disambiguation between *happy* and *fear* might be facilitated with this additional information. In future work we can consider adaptations of ToBI for emotional speech which would account for this difference.

5. Classification with ToBI features

Our analysis clearly indicates that different emotions exhibit differences on discrete intonation patterns.Now we turn to investigate the effectiveness of these ToBI features for automatic emotion classification. No prior work has tested the applicability of ToBI labels for automatic prediction.

Each utterance is represented by 41 discrete prosodic features which are a combination of all sets of individual ToBI features and bi-gram features in Tables 1. We use SVM classifiers with radial basis kernel constructed using the LIBSVM library [7]. We performed one-versus-all emotion classification tasks: recognition of each of the six emotions versus the other five emotions. For example, one of the tasks was to recognize if an utterance conveys *anger* versus some other emotion among *disgust, fear, happy, sad,* and *neutral*. Since the number of utterances for class *all* is much higher than the one for the target emotion, we performed down-sampling to equal size classes.

To investigate the effect of the sentence structure and context information, we performed experiments on both within-sentence and cross-sentence classification. For the within-sentence classification task, one-versus-all emotion recognition was performed on each of the three selected sentences separately. We perform 10-fold cross-validation on all renditions of the same sentence in different emotions. There are about 145 renditions of each sentence. The accuracy of prediction for each sentence is shown in the top section of Table 2.

Table 2: *One-versus-all accuracy for ToBI features*

	NEU	ANG	DIS	FEA	HAP	SAD	Ave.
within-sentence classification rate (%)							
TAI	82.4	81.0	81.4	77.6	80.4	73.9	79.5
TIE	78.8	79.6	76.4	76.8	81.6	75.1	78.1
ITH	82.0	76.2	74.3	76.8	74.8	77.9	77.0
Ave.	81.1	78.9	77.4	77.1	78.9	75.6	78.2
cross-sentence classification rate (%)							
TAI	77.0	69.0	69.8	77.0	75.4	57.9	71.0
TIE	69.3	68.0	68.4	64.8	75.9	55.7	67.0
ITH	75.1	67.2	64.1	66.0	51.5	70.8	65.8
Ave.	73.8	68.1	67.4	69.3	67.6	61.5	67.9

The results show that the sentence-specific ToBI features represent a promising line of research for predicting the target emotion. The performance is reasonable on all emotions for all three sentences, with average classification rate of 78.2%. Interestingly at the same time there is a noticeable performance difference for different sentences. For instance, we obtain much higher classification rate of 80.4% and 81.6% on TAI and TIE respectively for *happy*, while the lowest one on ITH of 74.8%.

In our second set of experiments, we turn to the analysis of the cross-sentence performance of emotion recognition. Here we train the model on all data from two sentences and test on the data from the remaining sentence. The results are shown in the bottom part of Table 2, each row representing results when

the given sentence was used as a test set. Compared to the results of within-sentence prediction, cross-sentence accuracy degrades considerably, as can be expected. In contrast to the within-sentence validation, the average classification rate drops from 78.2% to 67.9% and we observe consistent degradation of around 10% on all emotions. This indicates that the ToBI features are highly related to the carrier sentence.

Neutral was the emotion for which classification rate was highest in both types of experiments. It was also the emotion for which there was least degradation in accuracy between the two types of experiments. This finding suggests that the intonation patterns in terms of ToBI features are more robust on *neutral* utterances than on sentences expressing emotions. The worst performance in both experiments is on *sad*, which further corroborates the complex nature of emotions with low-activation.[2]

6. Classification with Acoustic Features

To compare the prediction power of categorical prosody cues and conventional bag-of-frame acoustic features for emotion classification, we conduct similar 1-vs-all emotion classification experiments with a set of state-of-the-art acoustic features.

We use the openSMILE feature extraction library [9] to obtain a comprehensive set of standard acoustic features. The openSMILE library extracts 26 low-level descriptors including intensity, loudness, F0, F0 envelope, probability of voicing, zero-crossing rate, 12 MFCCs, and 8 LSFs. We also use the first order delta coefficients for these features, as well as 19 summary functions for a total of 988 features.

Table 3 lists the corresponding performance of acoustic features in the within-sentence and sentence-independent emotion classification tasks. In the within-sentence emotion classification tasks with conventional acoustic features achieved average accuracy of 78.5%, which is comparable to the 78.2% we obtained with ToBI features in Table 2. However, the prediction power of acoustic features and ToBI features vary among different emotions. For instance, acoustic features show the highest prediction power on *anger*, while ToBI features work better on differentiation of *neutral* and emotional utterances.

Unlike ToBI features, raw acoustic features appear to be less sensitive to the carrier sentence and lead to practically identical accuracy on within- and cross-sentence prediction.

Table 3: *Classification rate of one-versus-all emotion classification using acoustic features.*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
within-sentence classification rate (%)							
TAI	77.0	90.2	72.6	70.2	68.2	79.3	76.3
TIE	84.6	90.0	72.6	74.6	80.0	84.0	81.0
ITH	76.4	90.0	76.2	76.2	71.8	79.4	78.3
Ave.	79.3	90.0	73.8	73.7	73.3	80.9	78.5
cross-sentence classification rate (%)							
TAI	75.4	89.1	71.4	77.4	79.0	77.2	78.3
TIE	87.7	90.6	73.4	67.6	77.6	77.9	79.1
ITH	81.3	88.8	75.1	69.7	78.4	77.4	78.5
Ave.	81.5	89.5	73.3	71.6	78.3	77.5	78.6

[2]In experiments that fall out of the scope of this paper, we also conducted experiments with automatically derived ToBI annotation using AutoToBI [17]. The classification results were poor, and the most predictive ToBI features were practically never recognized correctly, probably because of their rare occurrence in the non-emotional data on which the system is trained.

Finally, we turn to discuss the combination of acoustic and categorical prosodic cues for emotion classification. We apply early stage feature fusion and train SVM classifiers with one feature vector concatenating ToBI features and acoustic features. We consider the sentence-independent task and list the corresponding emotion classification performance of the fusion classifiers in Table 4. Compared with the best single classifiers with standard acoustic features, we can obtain modest improvement by including categorical prosodic ToBI information, where the final average classification accuracy increases from 78.6% to 79.3%. The combined feature representation does achieve noticeable improvement on *neutral, fear,* and *anger*, which is consistent with what we found in cross-sentence results in Table 2 that ToBI features also show higher prediction power on these three emotion classes.

Table 4: *Classification rate of one-versus-all emotion classification by fusion of acoustic and ToBI features*

	NEU	**ANG**	**DIS**	**FEA**	**HAP**	**SAD**	Ave.
sentence-independent classification rate (%)							
TAI	78.2	89.1	71.0	78.2	78.6	75.6	78.5
TIE	88.5	93.0	73.4	69.3	76.0	77.9	79.7
ITH	83.8	88.8	75.1	71.4	81.3	78.2	79.8
Ave.	**83.5**	**90.3**	73.2	**73.0**	78.6	77.2	79.3

7. Conclusions

We have presented a study of the relationship between prototypical emotion expression and discrete, perceptually based characterization on prosody. Our corpus is much larger than any of those used in prior work addressing a similar question. We show promising results both in descriptive and predictive power of ToBI features. Our findings reveal targets for continuous feature extraction that can capture the relevant prosodic phenomena.

In addition, we compare the prediction power of categorical prosodic ToBI features with state-of-the art bag-of-frames acoustic features. We find that ToBI features and acoustic features show comparable performances on within-sentence emotion predictions. Unlike ToBI features, conventional acoustic features are very robust to different carrier sentences. It is nevertheless notable that the much smaller and interpretable set of ToBI features conveys rich information about emotional state.

Finally, we achieved further improvements by integrating ToBI features and acoustic features. This suggests that categorical prosodic ToBI representations can provide complementary information to the conventional acoustic features in prediction of emotion. Our work presents evidence that discrete characterizations of intonation have the potential to inform future feature development for emotion recognition and may lead to overall improved performance.

8. Acknowledgment

This work was supported in part by the following grants: NIH R01-MH073174, NIH P50-MH096891, and NIH R01-MH084856. Some work results from project implementation: Research and development of new information technologies for forecasting and mitigation of crisis situations and safety, ITMS 26240220060 supported by the Research & Development Operational Programme funded by the ERDF.

9. References

[1] Bänziger, T., Scherer, K., "The role of intonation in emotional expressions," *Speech Communication*, vol. 46(3-4), pp. 252-267, 2005.

[2] Beckman, M. E., Ayers, E., Guidelines for ToBI labelling, version 3.0, 1993.

[3] Beckman, M. E., Hirschberg, J., Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", In S.-A. Jun [ed.], *Prosodic Typology – The Phonology of Intonation and Phrasing*, 2005.

[4] Benus, S., Gravano, A., Hirschberg, J., "The prosody of backchannels in american english," in *Proceedings of ICPhS*, pp. 1065-1068, 2007.

[5] Boersma, P., Weenink, D., Praat: doing phonetics by computer [Computer program], Version 5.3.42, http://www.praat.org, 2013.

[6] Busso, C., Lee, S., Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", *IEEE Trans. Audio Speech Language Process.*, Vol. 17(4) pp. 582-596, 2009.

[7] Chang, C.-C., Lin, C.-J., LIBSVM: a library for support vector machines, http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[8] Cowie, R., "Describing the emotional states expressed in speech", In *Proc. of the ISCA Workshop on Speech and Emotion*, pp 11-18, 2000.

[9] Eyben, F., Wöllmer, M., Schuller, B., "openSMILE: The Munich versatile and fast open-source audio feature extractor", in *Proc. of the International Conference on Multimedia*, pp. 1459-1462, 2010.

[10] Laukka, P., Juslin, P. N., "Similar patterns of age-related differences in emotion recognition from speech and music", *Motivation & Emotion*, vol. 31, pp. 182–191, 2007.

[11] Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J. Emotional prosody speech and transcripts, Linguistic Data Consortium, Philadelphia.

[12] Liscombe, J., *Prosody and Speaker State: Paralinguistics, Pragmatics and Proficiency*, PhD thesis, Columbia University, 2007.

[13] Mozziconacci, S. J. L., *Speech variability and emotion: production and perception*, Ph.D. thesis, Technical University Eindhoven, 1998.

[14] Mozziconacci, S. J. L., "Prosody and Emotions", 2002.

[15] Nakatani, C., Hirschberg, J. and Grosz, B., "Discourse Structure in Spoken Language: Studies on Speech Corpora", in *Proc. of AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[16] Pang, B., Lee, L., "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol. 2(1-2), 1-135, 2007.

[17] Rosenberg, A., "Autobi ? a tool for automatic tobi annotation", in *Proc. of Interspeech*, 2010.

[18] Russ, J. B., Gur, R.C., Bilker, W. B., "Validation of affective and neutral sentence content for prosodic testing", *Behavior Research Methods*, vol. 40(4), pp 935-939, 2008.

[19] Scherer, K. R., Ladd, D. R., Silverman, K., "Vocal cues to speaker affect: testing two models", *Journal of the Acoustic Society of America*, vol. 76(5), pp 1346-1356, 1984.

[20] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., "Tobi: a standard for labeling English prosody", in *Proc. of ICSLP*, pp. 867-870, 2002.

[21] Taylor, P., "The tilt intonation model", in *Proc. of ICSLP*, pp. 1383-1386, 1998.

[22] Wiebe, J., Wilson, T., Cardie, C., "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, Vol. 39(2-3), pp. 165-210, 2005.

[23] Darwin, C., The expression of emotion in man and animals, New York: Oxford University Press. (Original work published 1872)

[24] Ekman, P., "Facial expression of emotion: New findings, new questions", Psychological Science, vol.3, pp. 34–38, 1992.

[25] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J., "Fully Automatic Facial Action Recognition in Spontaneous Behavior", in *Proc. of FGR 2006*, pp. 223–230, 2006.

"Young" and "Old" Voice: the prosodic auto-transplantation technique for speaker's age recognition

Massimo Pettorino, Elisa Pellegrino, Marta Maffia

Department of Literary, Linguistic and Comparative Studies,
University of Naples "L'Orientale", Italy
{mpettorino, epellegrino, mmaffia}@unior.it

Abstract

The present study is intended to figure out the extent to which prosody and intonation affect listeners' ability to estimate the speaker's age. The performance of a 40-year old anchorman and another by the same speaker at the age of 80 were spectro-acoustically analyzed in order to identify the prosodic features of a "young" and an "old" voice. The results of the analysis have shown significant differences between the two voices on a suprasegmental level. To test the effects of these differences on a perceptual level, through the prosodic transplantation technique, the F0 values and the durations of segments and silences were transferred from the "young" to the "old" voice and viceversa. Two age recognition tests, based on original and transplanted voices, were administered to Italian listeners. The results of perceptual tests have confirmed the strict relationship between some rhythmic and prosodic features and the speaker's age and have demonstrated the effectiveness of the transplantation technique. With advancing age, articulation rate and speech rate slow down, voice register rises and tonal range widens. Moreover, the "old" voice is also characterized by a higher percentage of vocalic portion which determines a shift of the Italian rhythm towards the isomoraic pattern.

Index Terms: Prosodic correlates of speaker's age, Speaker's age recognition, Prosodic Transplantation Technique.

1. Introduction

The relationship between the speaker's age and his/her own voice has been investigated in several experimental studies. Acoustic [1], [2], [3] and perceptual researches [4], [5], [6] have underlined that the voice changes with age, at both segmental and supra-segmental levels, because of the numerous anatomical and physiological modifications of the respiratory mechanism and of the phonatory apparatus. For example, lung tissue loses its elasticity, the thorax tends to stiffen, muscles weaken and vital capacity decreases [7]. In the elderly, a process of calcification of laryngeal cartilages also occurs and the vocal folds become thinner, stiffer and less elastic [8]. Previous studies on the effects of aging on acoustic parameters of voice have demonstrated that older voices are likely to undergo progressive tonal lowering [9], lowering of speech rate [10], [1], increasing of jitter and shimmer [8], [11], [12], lowering of formant frequencies [13], lengthening of vowels and stop consonants [14], increasing of standard deviation of F0 [13], [15], [16].

Nevertheless, it is important to underline that these data result from studies that analyze different kinds of corpora (spontaneous or read speech), use different techniques of speech data collection and involve different languages [2]. The considerable methodological heterogeneity across these kinds of studies certainly does not guarantee that the changes undergone by older voices are dependent exclusively on the speaker's age. Besides the chronological age, other relevant variables, such as the idiosyncratic characteristics of a speaker's voice, the contextual situation, the kind of speech and the subject matter could influence the oral performance of the speakers and affect data comparability.

In order to control all these variables, and thus, to be able to exclusively assess the effect of aging on the acoustic parameters of voice, it would be very effective to record the same speakers uttering the same words in the same communicative situation, but at different ages.

2. The study

The objective of this study is to verify the role played by prosody and intonation in the listener's ability to evaluate the speaker's age. To achieve this, a particular corpus of Italian speech was collected. The read speech of a 40-year old anchorman, Piero Angela, was extracted from a 1968 TV news broadcast and orthographically transcribed. In 2007, the 79-year old Piero Angela, who still worked as a RAI journalist, read the same 1968 script again, acting as if he were hosting a real TV news broadcast. The recording was taken at RAI TV studios in Rome, in order to maintain the same communicative situation.

Preliminary spectro-acoustic analyses conducted on the 1968 and 2007 corpora showed differences between the two voices, both on the segmental and suprasegmental levels [17]. The "old" speech was clearly more isochronous, exhibited wider tonal range and presented longer and more frequent silent pauses than the "young" voice.

In order to investigate further the role played by the specific acoustic parameters of the "young" and "old" voices, three utterances drawn from the 1968 TV news broadcast and the corresponding utterances of the 2007 corpus were manipulated through the prosodic transplantation technique. This procedure is based on the PSOLA (Pitch-Synchronous Overlap and Add) algorithm and it is implemented in Praat [18], [19], [20]. The six utterances were segmented and annotated into four tiers:

- phones
- syllables
- intervals of consecutive consonants and vowels
- intervals between two consecutive vowel onset points (VtoV).

In order to apply the transplantation procedure, the "phones" tiers were duplicated and modified, so that each segment of the 1968 utterance had a corresponding segment in the 2007 utterance. Since the transplantation procedure requires that the TextGrids of donor's and receiver's voices contain the same number of elements [21], a micro segment was inserted to avoid mismatch between the 1968 and 2007 utterances. Thanks to the transplantation procedure, the rhythmic and prosodic features of the donor's voice were

transferred to the receivers' voice. As a result, in the present study, the utterances produced in 1968 by the 40-year old Piero Angela were made sound "older" by transferring the pitch contour on them, the durations of phones and silences of the corresponding 2007 utterances. By contrast, the latter were made to sound "younger" since they acquired the prosodic features of the 1968 utterances.

The resulting corpus was, thus, made up of three original utterances produced in 1968, three original utterances produced in 2007, three 1968 utterances with 2007 prosody and three 2007 utterances with 1968 prosody.

2.1. Perception test

In order to assess on a perceptual level the effect of the acoustic differences between the "old" and "young" voices, 70 university Italian students, ranging in age from 23 to 26, were administered a perception test. Following the experimental protocol used in a previous research [6], the test was divided into two main phases. In the first phase, participants listened to 16 pairs of utterances. Each pair was composed of two voices reading the same news. For each pair, listeners were asked to rate if the youngest speaker was the "1st voice" or the "2nd voice", or if the two speakers were of the "same age". The test was designed with the following voice combinations:

- 2007 voice - 2007 voice (Old –Old; henceforth O-O);
- 1968 voice -1968 voice (Young-Young; henceforth Y-Y);
- 2007 voice – 1968 voice (O-Y);
- 1968 voice - 2007 voice with the transplanted prosody of 1968 voice (Y-tY);
- 2007 voice - 2007 voice with the transplanted prosody of 1968 voice (O-tY);
- 1968 voice - 1968 voice with the transplanted prosody of 2007 voice (Y-tO);
- 2007 voice - 1968 voice with the transplanted prosody of 2007 voice (O-tO);
- 2007 voice with the prosody of 1968 voice – 1968 voice with the prosody of 2007 voice (tY-tO).

The test material also included a control pair (cO-cY), consisting in a 85 year-old voice and a 27 year-old voice reciting the same utterance: "le foglie diventano gialle, l'albero muore" (Engl. "when the leaves turn yellow, the tree dies"). In the second phase of the perception test, the same subjects were asked to listen to 14 single utterances and to indicate the speaker's age, choosing between seven age bands: 26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86-95.

2.2. Results of perception test

2.2.1. First test

Overall results of the perception test show listeners' ability to accurately recognize the speaker's age from speech sample alone. As a matter of fact, when exposed to the pairs based on the same voice O-O, Y-Y, these pairs are properly rated as being of "the same age", respectively in 100% and 96% of cases. When the pairs were composed of the 2007 voice and the 1968 voice (O-Y), 84% of listeners rate the second voice as the youngest.

The results obtained by the pairs composed by original and transplanted voices confirm the role of prosody and intonation in the perception of speaker's age. At the same time, the listeners' judgments highlight the effectiveness of the prosodic transplantation technique in making voices sound "older" or "younger". As for the aging effect, it was tested both with respect to the 2007 "old" voice and to the 1968 "young" voice (figs 1 and 2). As it is shown in figure 1, when the aged voice (tO) was paired with the original old voice (O), on average, the 60% of listeners judged the two items as being of the "same age".

Comparing the rates given to the O-tO pair with those assigned to the O-Y pair, the percentage of the answer "2nd voice" decreases from 84.2% to 30.8%, while the answer "same age" increases by about 50%, from 13 to 60.3%.

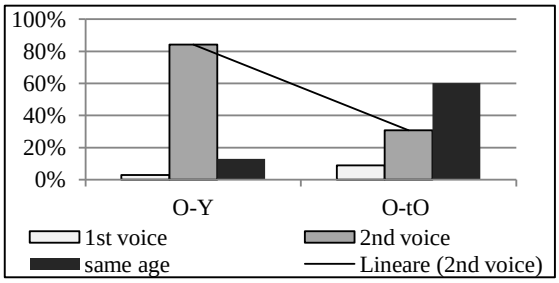


Figure 1: Aging effect with respect to the 2007 old voice. The rates given to pairs O-Y and O-tO are significantly different ($p<0.01$).

The data in figure 2 show the results obtained from the comparison between the young and the aged voice (Y-tO). 60% of listeners rate the "1st voice" as the youngest. The aging effect is particularly evident if one compares the answer "same age" obtained with the Y-Y pair and that of the Y-tO pair. In the former case, the third option is chosen by 96.8% of listeners, while in the latter, this percentage decreases to 38%.

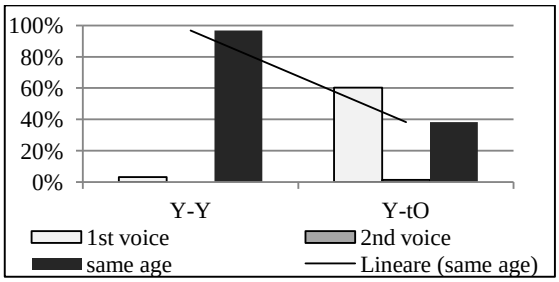


Figure 2: *Transplantation effect with respect to the 1968 young voice. The rates given to pairs Y-Y and Y-tO are significantly different ($p<0.01$).*

As regards the effect of making the voice sound "younger" (fig. 3), it is possible to claim that it produces changes in the "old" voice but its influence on listeners is not as effective as the aging effect. In the pair O-tY, the transplanted voice is recognized as the youngest by the 31.8% of listeners.

The data in figure 4 show that the rates given to the two pairs Y-O and Y-tY do not undergo a significant variation ($p>0.05$). However, in the Y-tY pair the percentage of "1st voice" decreases from 84.2 to 67. By contrast, the percentage of "same age" increases from 13 to 20.5 and in 11.5% of cases the synthesized young voice is even recognized as the youngest. In the pair composed of both transplanted voices,

tY-tO, 56% of listeners judge it as if it was composed of voices of the same age.

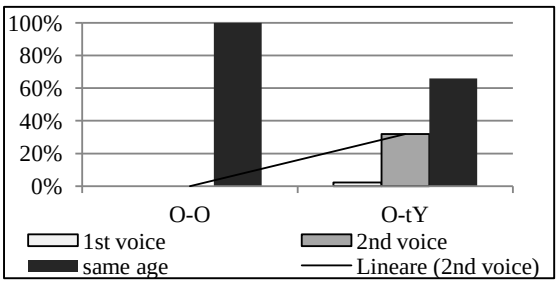


Figure 3: *Transplantation effect in respect to the 2007 old voice. The rates given to pairs O-O and O-tY are significantly different (p<0.01)*

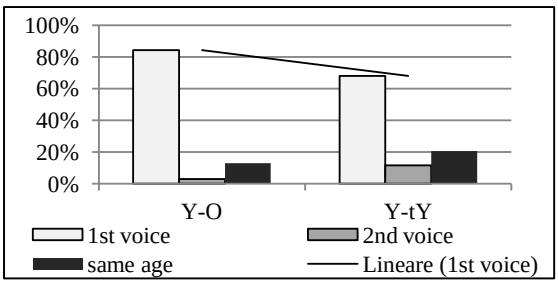


Figure 4: *Transplantation effect in respect to the 1968 young voice.*

2.2.2. Second test

In the second perception test, listeners were asked to estimate the speaker's age, choosing from 7 age bands. The data in figure 5 show that the young control voice (actual age 27) is recognized as belonging to the first age band (26-35) by 98% of listeners; while the control old voice (actual age 85) is judged by the 60% as ranging in the band 76-85. On a perceptual level, the 1968 voice varies from the 2nd to the 4th band, with the highest percentage of rates given to the 3rd band (46-55). On the contrary, the rates given to the 2007 voice are distributed on higher age-bands (4-6), with 44% of listeners choosing the 5th band.

As for the synthesized voices, they occupy an intermediate position between the original young and old voices. The age bands chosen by the highest number of listeners were the 4th and the 5th. These results indicate that the prosodic transplantation technique had a strong effect on the recognition of speaker's age in both directions: the aged voice is perceived as older than the original 1968 voice, as well as the synthesized young voice being rated as younger than the original old voice. From the data relative to the age bands, the weighted average of the perceived age was calculated, taking into account the recognition percentage scored by each band. Figure 6 plots the perceived and actual ages of the original, control and transplanted voices. The comparison between the perceived and actual ages of the speakers suggests different considerations. First of all, the variance between the two original voices decreases from the actual 40 years to perceived 20 years. The manipulated voices are perceived as produced by 66 and 65-year old speakers respectively.

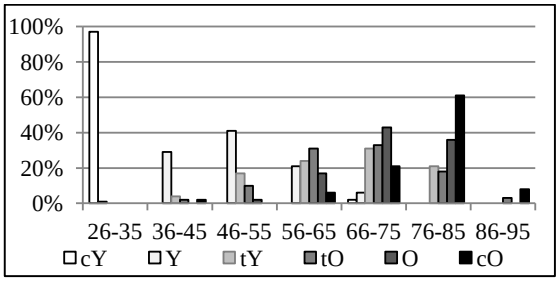


Figure 5: *Perceived age of control, original and transplanted voices.*

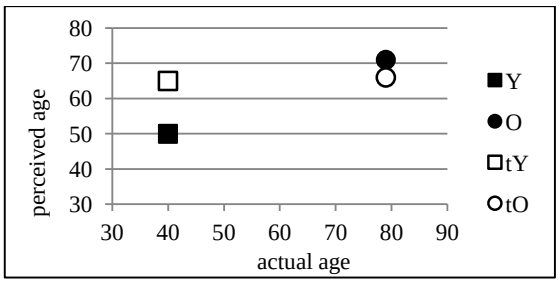


Figure 6: *Perceived and actual age of the original and transplanted voices.*

2.3. Spectro-acoustic analysis

The results of perception have shown that the 70 subjects were able to accurately judge speaker's age relying solely on single voice excerpts. This datum leads us to assume that the perception of the speaker's age is greatly influenced by the prosodic and intonational features of the utterance. To test this hypothesis, the whole corpus was spectro-acoustically analyzed. On the basis of the speech segmentation in phones, syllables, vocalic and consonantal intervals, VtoV intervals, the following measurements were taken: duration of phones, syllables, silent and filled pauses, length of vocalic and consonantal intervals, duration of VtoV intervals. Additionally, for every utterance the minimum, the maximum and the mean F0 values were measured.

On the basis of these measurements, the following rhythmic and prosodic parameters were calculated:

- articulation rate (AR) (syll/s.), i.e. the ratio between the number of syllables really uttered and phonation time;
- speech rate (SR) syll/s), i.e. the ratio between the number of syllables really uttered and total time of the utterance, including silent and non silent pauses;
- fluency (F) or frequency of silences in the utterance;
- tonal range (Hz);
- F0 register (Hz);
- speech time composition in terms of syllable, silence and disfluence percentage;
- utterance composition in terms of vocalic and consonantal percentage (%V and %C);
- mean duration of VtoV intervals (s).

In order to figure out which of the above-mentioned prosodic parameters could have influenced the listener's discrimination between the "young" and the "old" voice more

profoundly, for every prosodic parameter the average values obtained in the 1968 corpus and 2007 corpus were calculated.

Table 1 shows the data regarding AR and SR. Unsurprisingly, the 2007 voice exhibits a decrease both in terms of AR and SR, due to the slower mobility of the articulators. However in both the 1968 and 2007 voices, the SR is 0.8 syll/s slower than AR, and this is imputable to the same portion of silent pauses (13%) occurring in the two corpora. It is worth underlining that, despite the duration of the silent portion being stable between the two voices, the frequency of silences increases considerably with advancing age: in 1968 silences are more rare, on average 1 out of 14 syllables, while in 2007 they are more frequent, 1 out of 7.6 syllables.

Table 1. *AR and SR.*

	AR	SR
1968	6.3	5.5
2007	5.4	4.6

As for pitch contour, figure 7 shows that with advancing age the voice register and tonal range become higher. In the 2007 voice the average pitch reaches 146,8 Hz, while it is lower in the 1968 voice (68 Hz). The tonal range widens by about 17% from 1968 to 2007.

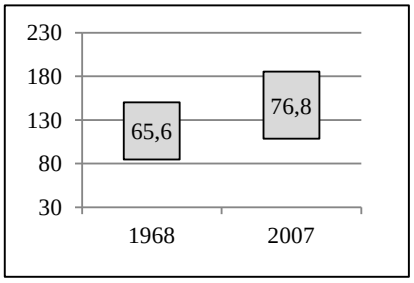


Figure 7: *Tonal range and Register*.

The other parameter under investigation was the vocalic and consonantal percentage in the utterance. Likewise the pitch movements, as well as the vowel percentage in the utterance increases with advancing age. In the corpus produced when Piero Angela was 40 years old, the vocalic portion amounts to 46%, while in the speech of the 80-year old speaker it reaches 51%. This datum enables us to clarify the components – vowels rather than consonants - on which the AR slowing down mostly depends. This phenomenon can be explained by the greater articulatory stability that characterizes vowels rather than consonants. The AR slowing down is, therefore, mainly due to the maintenance of static positions rather than to a variation of articulatory dynamics.

The variations of vocalic portion deserve more attention since this parameter plays a significant role in the rhythmic classification of languages. Indeed, according to a number of studies conducted on different languages, the traditional division in three rhythmic groups (syllable, stress, mora-timed languages) is related to vowel percentage and to another parameter, that according to [22] corresponds to ΔC (standard deviation of consonantal portions) and according to [23] to VtoV. In order to process the data of this corpus in the research framework of rhythmic organization of languages, the average %V and VtoV values obtained in Piero Angela's corpus were plotted with the ones obtained by [23]. Figure 8 shows the overall data relative to the different corpora. As it is clearly shown in the graph, the utterances produced by the 40-year old speaker lie in correspondence to the Italian TV news broadcast. On the contrary the mean VtoV and %V values of the utterances produced by the 80-year old speaker move to the right side of the graph, in the area occupied by Japanese, a mora-timed language. Thus, with advancing age, Italian speech seems to assume values more similar to mora-timed languages.

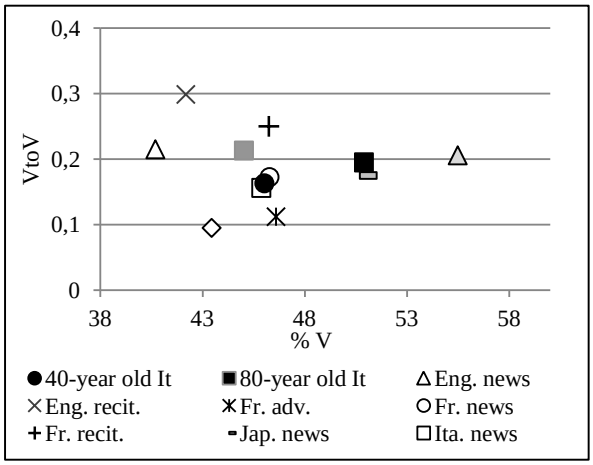


Figure 8: *%V and VtoV of different languages*.

This datum does not contrast the results of the above mentioned studies on the rhythmic features of languages, since neither the speakers involved in [22] nor the ones considered in [23] were 80 years old. Preliminary data on the speech produced by Japanese speakers of different age seem to confirm the trend which emerged for Italian: %V increases with advancing age (from 50.7 to 55.4%).

3. Conclusions

The spectro-acoustic analysis carried out on the corpus of read TV news broadcasts, produced by the same speaker at the age of 40 and 80 has demonstrated the existence of a relation between some rhythmic and prosodic features and the speaker's age. The old voice was characterized by a slowing down of AR and SR, by the rising of voice register and the widening of tonal range. The old voice also presents a higher percentage of vocalic portion (%V), and this increase makes Italian rhythm become more similar to the pattern of isomoraic languages. Thanks to the prosodic transplantation technique, the effect of these differences on a perceptual level was assessed. The judgments given to the transplanted voices suggest different considerations. Firstly, listeners are able to recognize the speaker's age relying solely on specific prosodic features. Secondly, on a perceptual level, the aging effect is more effective than that of making voices sound younger. In addition, the transplantation technique is therefore an effective procedure to the purpose of manipulating a speaker's age.

The data regarding the %V increase with advancing age suggesting the opportunity to extend the research on other languages belonging to other rhythmic groups.

4. References

[1] Hollien, H., Shipp, T., "Speaking fundamental frequency and chronological age in males", *Journal of Speech and Hearing Research*, 15, 155-159, 1972.

[2] Russell, A., Penny, L. and Pemberton, C., "Speaking fundamental frequency changes over time in women: A longitudinal study", *Journal of Speech and Hearing Research*, 38, 101-109, 1995.

[3] Schötz, S., "Acoustic Analysis of Adult Speaker Age", in C. Müller [Ed.], Speaker Classification I, Lecture Notes in Computer Science, 88–107, Springer, 2007.

[4] Horri, Y., Ryan, W. J., "Fundamental frequency characteristics and perceived age of adult male speakers", Folia Phoniatrica, 33, 227-233, 1981.

[5] Schötz, S., "Perception, analysis and synthesis of speaker age", *Travaux de l'Institut de Lund*, 47, 1-186, 2006.

[6] Pettorino M., Giannini A., "The speaker's age: a perceptual study", *Proceedings of the 17th ICPhS*, Hong Kong, 1582-1585, 2011

[7] Awan, S.N., "The aging female voice: acoustic and respiratory data", *Clinical Linguistics & Phonetics*, 20/2-3, 171-180, 2006.

[8] Linville, S.E., "The aging voice", *The American Speech-Language-Hearing Association (ASHA) Leader*, 12-21, 2004.

[9] Lindblad, P., *Rösten*, Lund, Studentlitteratur, 1992.

[10] Amerman, J.D., Parnell, M.M., "Speech timing strategies in elderly adults", *Journal of Phonetics*, 20, 65-76, 1992.

[11] Ramig, L.A., Ringel, R.L., "Effects of physiological aging on selected acoustic characteristics of voice", *Journal of Speech and Hearing Research*, 26, 22-30, 1983.

[12] Dehqan, A., Scherer, R. C., Dashti, G., Ansari-Moghaddam, A., & Fanaie, S., "The effects of aging on acoustic parameters of voice", *Folia Phoniatrica et Logopaedica*, 64, 265-270, 2012.

[13] Linville, S.E., "Acoustic-perceptual studies of aging voice in women", *Journal of Voice*, 1, 44-48, 1987.

[14] Ptacek, P.H., Sander, E.K., "Age recognition from voice", *Journal of Speech and Hearing Research*, 9, 273-277, 1966.

[15] Jacques, R., Rastatter, M., "Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners", *Folia Phoniatrica*, 42, 118–124, 1990.

[16] Traunmüller, H., van Bezooijen, R., "The auditory perception of children's age and sex", *ICSLP*, 1171-1174, 1994.

[17] Giannini, A., Pettorino, M., "L'età della voce", *La Fonetica Sperimentale: Metodo e Applicazioni, Atti del IV Convegno Nazionale AISV 4*, 165-178, 2009.

[18] Charpentier, F. and Moulines, E., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication* 9, 453-467, 1990.

[19] Boersma, P., "Praat, a system for doing phonetics by computer", *Glot International* 5:9/10, 341-345, 2001.

[20] Yoon, K. "Imposing Native Speakers' Prosody on Non-native Speakers' Utterances: The Technique of Cloning Prosody", *Journal of the Modern British & American Language & Literature* 25(4): 197-215, 2007.

[21] Pettorino, M. and Vitale, M. "Transplanting native prosody into second language speech", in M. G. Busà and A. Stella [Eds], *Methodological Perspectives on Second Language Prosody*. Papers from ML2P 2012, 11-16, Padova: CLEUP, 2012

[22] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition* 73, 265-292, 1999.

[23] Pettorino, M., Maffia ,M., Pellegrino, E., Vitale, M. and De Meo, A., "VtoV: a perceptual cue for rhythm identification" in Mertens, P. & A.C. Simon [Eds], *Proceedings of the Prosody-Discourse Interface Conference 2013* (IDP-2013). Leuven, September 11-13, 2013, 101-106, 2013.

Expressive *vs* neutral prosody in reading aloud: from descriptive binary to continuous features

Julien Magnier[1], *Maya Gratier*[2], *Anne Lacheret*[3]

[1, 2] Laboratoire Ethologie, Cognition, Développement, Université Paris Ouest Nanterre, France
[3] Laboratoire Modyco, Université Paris Ouest Nanterre, France

jmupon@gmail.com, gratier@gmail.com, anne@lacheret.com

Abstract

In this paper, we propose to compare expressive and neutral oral renditions of a children's tale in French by examining the segmentations performed by twelve high level French readers. We used a software dedicated to this kind of analysis (Analor) which takes into account different parameters (pause, pitch gesture, pitch jump) and their relative strength to determine pertinent prosodic units (phrases). The extraction of these phrases and their features enables us to observe the influence of both the type of oralisation (expressive or neutral) and punctuation signs on the organization of speech flow. Results show there are more prosodic phrases in the expressive readings (specially at comma locations), that their boundaries are more clearly demarcated, and that they have more varied contours than in neutral readings.

Index Terms: expressive prosody, neutral prosody, narrative, oral segmentation, phrasing

1. Introduction

Vocal expressivity, which we define as the capacity to express feelings, intentions and attitudes, is an important issue for research on oral communication and lies at the interface of computational linguistics, phonetics and psycholinguistics [1]. A number of prosodic resources can be used to impart expressivity to written text. [2] Some of these are universal features of all languages (variations in melody, rhythm and intensity, etc…) while others are language-specific (the nature of the variations, the use of melodic rather than temporal cues or vice versa). We focus in this paper on the prosodic process of grouping and phrasing in French performed by readers in a narrative task during a controlled study of oralized reading (Section 2). The aim of this study was to measure the variations in phrasing and prosodic grouping, and the nature of their correlation with punctuation signs[1], which distinguish two modes of oralisation, a neutral reading and an expressive reading.

Several studies confirm the influence of punctuation in the marking of breaks in oralized text [3], [4] and the variable durations of these breaks according to the type of punctuation sign [5] or the level of closeness of the sequence [6], [7]. However, other studies highlight the role of vocal expressiveness in the placement of these breaks, produced sometimes in places not oriented by punctuation marks or by discourse structure, but rather by the emotional qualities of the text [8], [9], [10]. Reading aloud is an activity which requires an emotional and physical involvement on the part of the reader [11], and we may assume that the structuring of breaks and intonation groups varies according to the degree of this involvement.

It is thus relevant to question the role of vocal expressiveness in the processes of segmentation in reading. Would the prosodic structuring of a text be less marked by readers who've been asked to read a story in a neutral manner ?

To answer this question, we tested 3 hypotheses: i) the number and location of phrase boundaries should be greater in the expressive reading condition; ii) the relative strength of phrase boundaries should be greater in the expressive condition; and iii) expressive reading should involve more pitch variation, i.e. specific contour types.

2. Corpus

2.1. Participants and task

Participants were 12 adults (6 men and 6 women ranging from 20 to 55 years in age) who considered themselves to be 'good readers'. They were asked to read aloud a children's tale (*the laughter of the frog* – 426 words, 23 full stops, 42 commas) in two styles: an expressive style, varying the tone of voice to convey emotion, and a neutral one, i.e. without expressing emotion.

(…) Dans sa folie des grandeurs, la grenouille avait asséché toute la planète. Les êtres vivants mourraient de soif et commençaient à suffoquer. (…)

Example 1.*Text extract from the tale*

The text was presented on a white sheet of paper in a single block, without paragraph breaks. Each reader was given the time he/she needed to familiarize him/herself with the tale and once he/she was ready, performed each of the 2 tasks (reading styles) as often as necessary until he/she was satisfied. The readings were recorded by means of a digital audio device (Roland BR600) with a sampling frequency of 44.1 kHz. The final corpus included 24 recordings (12 neutral readings and 12 expressive readings) with a total duration of 62 minutes.

2.2. Prosodic Segmentation

Each sample was segmented semi-automatically into prosodic phrases with the help of a software program called Analor[2] based on a method that has previously been tested [12]. The algorithm used by Analor is based on a global and multiparametric approach. The segmentation procedure is as follows: only melodic variations in time and breaks are used for segmentation into phrases, regardless of any segmental and syntactic data. Each break of at least 0.1 sec. is assigned a

[1]These markers co-exist with others (such as intonation contour variation, prototypicality of contours, melodic and tempo variation or vocal quality). They are considered easily observable indices for a model of vocal expressivity that precedes the phonetic level of analysis.

[2]http://www.lattice.cnrs.fr/analor

temporary marking and becomes a potential candidate for a phrase boundary. A break is a necessary but not a sufficient marker to locate a potential phrase boundary. In other words, the approach is global because the localization of a boundary can be envisaged only with respect to the combination of several parameters. Two other criteria are also used: (i) the detection of an ample melodic gesture; the trait [±ample] is fixed according to the melodic interval, measured in semitones, between the last extreme F0 value (before the boundary break) and the average F0 over the whole segment preceding the break; and (ii) the detection of a melodic jump which corresponds to the melodic interval which separates the points of F0 before and after the break (melodic resetting). An interesting characteristic of the algorithm is that the decision to place a boundary and its relative strength do not depend on the thresholds of each parameter taken independently but on the interaction between a set of parameters (Table 1, Figure 1).

PAUSE (seconds)	GESTURE (semi tones)	JUMP (semi tones)	Strength of parameter
x < 0,25	x < 2	x < 2	- 1
0,25 < x < 0,6	2 < x < 5	2 < x < 3,5	0
0,6 < x < 1	5 < x < 8	3,5 < x < 7	1
1 < x	8 < x	7 < x	2

Table 2.*Grid of the parameter values used to segment speech into phrases. The strength of the phrase boundary depends on the sum of the strength of each parameter. Each potential phrase that has a general index ≥ 0 must be validated by the experimenter.*

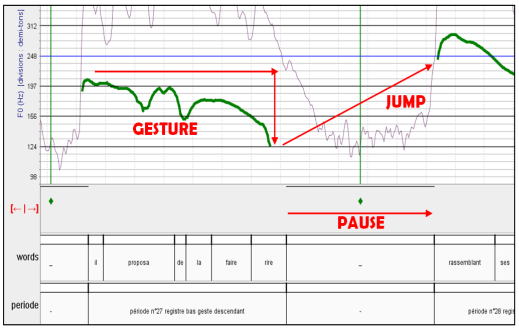


Figure 1: Screenshot of the software Analor and schematization of the criteria used for phrase segmentation

3. Data analysis

3.1. Analysis criteria

The comparisons between expressive and neutral readings were conducted, first, on the basis of the number of phrases produced by the subjects in the readings. We also compared the readings with regard to phrase boundaries taking into account text-based punctuation signs, i.e. 23 full-stops (excluding the final stop) and 42 commas. The tale contains 5 other punctuation marks, but these were excluded from analysis. The relative strength of phrase boundaries was also compared across all of the renditions, taking into account the punctuation signs. Finally, we quantified various types of general contours of phrases, according to the contour description provided by the Analor algorithm (flat, rise, fall). The contour of a phrase is considered flat if the amplitude of the gesture does not reach a minimal threshold (2 semitones (Table 1)) and it is considered as rising or falling if the absolute value of amplitude exceeds the given threshold.

Given the size of our sample, the non parametric Mann Whitney test was used to compare the averages.

3.2. Number and location of phrases and phrase boundaries

Figure 2 shows that, on average, readers produced many more phrases in the expressive reading (m=55.16, sd = 14.10) compared with the neutral reading condition (m=33.16, sd = 10.65). This difference is significant (U=11, p< 0.001). When we take into account the location of phrases and their correspondence with punctuation marks, it appears that full-stops almost systematically correlate with prosodic phrase boundaries in both conditions (m(exp) = 21.91, sd = 0.28, et m(neu) = 20.16, sd = 1.40). However, we find a significant difference between the expressive and neutral renditions at the comma level (U = 14, p< 0.001). The boundaries afforded by this punctuation mark are much less respected in neutral reading (m=8, sd = 6.68) than in expressive reading (m=21.66, sd = 7.61). Lastly, participants marked on average a slightly higher number of boundaries at unpunctuated locations in expressive reading (m=7.25, sd = 7.67) that in neutral reading (m=2.16, sd = 3.35) but this difference is not significant (p = 0.07).

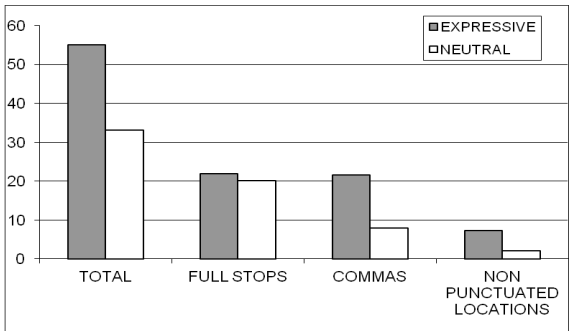


Figure 2: Mean number of phrases produced according to the versions and with respect to their locations

3.3. Strength of phrase boundaries

Based on the multiparametric index of boundary strength provided by Analor, it appears that the marking of boundaries is significantly stronger in the expressive readings (Figure 3) for all phrases (m(exp) = 2.40, sd = 0.46 and m(neu) = 1.41,sd = 0.54, U = 15, p< 0.001), with regard to full-stops (m(exp) = 2.83, sd = 0.63 and m(neu) = 1.65, sd = 0.66, U = 16, p < 0.01), and with regard to commas (m(exp) = 2.15, sd = 0.59,and m(neu) = 0.67, sd = 0.54, U = 4, p< 0.001). In both versions, boundaries are more strongly marked at full-stop locations than at comma locations (U(exp) = 31, p< 0.05 and U(neu) = 19, p< 0.01).

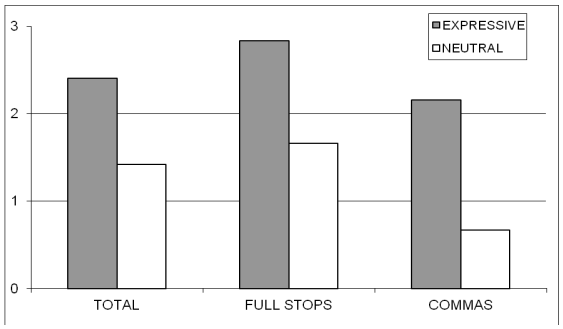


Figure 3: Mean index of cut strength according to the versions and with respect to their locations

3.4. Phrase contours

As shown in figure 4, the mean proportions of contour types are different in the two conditions. A higher proportion of flat contours is produced in neutral reading (m(exp) = 0.19, sd = 0.11, and m(neu) = 0.48,sd = 0.28,U = 26.5, p< 0.01), whereas a higher proportion of rising contours is produced in expressive reading (m(exp) = 0.17, sd = 0.12,and m(neu) = 0.04, sd = 0.05, U = 29.5, p> 0.05). Furthermore, a higher proportion of falling contours is produced in expressive reading, but this difference is only a trend (m(exp) = 0.63, sd = 0.12, and m(neu) = 0.46, sd = 0.27, p = 0.15).

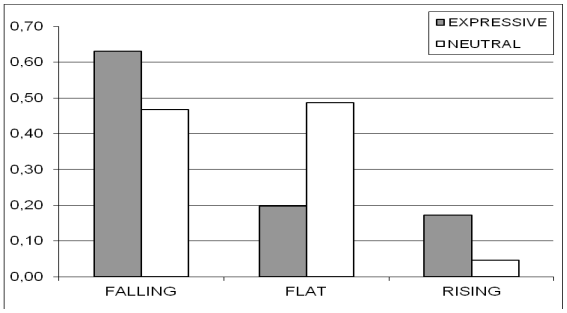


Figure 4: Mean proportions of contour types in the two versions

4. Conclusion and discussion

The aim of this study was to compare the prosodic grouping processes involved in expressive and neutral reading.

Our findings reveal a set of general factors that determine phrase segmentation in both conditions. Phrase boundaries, for instance, essentially coincide with punctuation marks and the full-stop is almost always a phrase boundary indicator. With regard to the strength of boundary markers, full-stops more frequently entail a segmentation than commas. This set of results was highly predictable as it illustrates the facilitative role of punctuation in segmenting a text into semantic units during oralisation. In addition, more than 45% of the phrases produced by the readers have a falling contour in both conditions. The high proportion of this contour type can be explained by the natural phenomenon of declination associated with the assertive modality of most of the clauses in the text.

However, our findings reveal a number of interesting differences between the expressive and neutral renditions of the text. It appears, first, that half of the commas give rise to prosodic boundary markers in the expressive condition as opposed to a little less than a quarter in the neutral reading condition. Readers thus use punctuation as a resource for conveying meaning and emotion in the oral rendering of written text. Furthermore, boundaries are more clearly marked (in terms of strength), whether they coincide with full-stops or commas, in the expressive reading condition.

These two observations point to a general principle associated with neutral reading, that of a minimal segmentation into prosodic units. The more frequent and marked segmentation in expressive reading, which we find in other genres such as political speeches, oratory or sermons, suggests that rhythm is used in oral discourse to support intersubjective engagement and emotional bonding. In effect, the neutral style can be considered an artefact of the experimental situation, and to be ecologically invalid, especially in a narrative genre where the plot itself pushes the reader to tell the story with emotion[1]. For this reason, the major implication of our findings is that expressive segmentation should be treated not in binary terms but rather as a continuous process of combining variable prosodic features.

Finally, the low frequency of flat contours in expressive reading compared to neutral reading shows that the prosodic structure of a text is made richer and more contrasted in the expressive rendition. These features may then facilitate the on line processing of prosody by the receiver and, it can be hypothesized, its comprehension and memorization. Further research should be undertaken to test this idea.

[1]Because lexical and syntactic properties of text carry intrinsic expressivity independently of prosody [13].

5. References

[1] Suciu, I., Kanellos, I. and Moudenc, T., "Expressivité et synthèse vocale. Isotopies expressives, cohérence discursive et structures prosodiques," *Nouveaux Cahiers de Linguistique Française*, vol. 28, pp. 199-206, 2007.

[2] Patel, R. & Mc Nab, C., "Displaying prosodic text to enhance expressive oral reading," *Speech Communication*, vol. 53, pp. 431-441, 2011.

[3] Rossard, B. & Cosnier, J., « Etude des pauses dans la lecture orale, » *Psychologie Française*, vol.26, pp. 54-67, 1981.

[4] O'Connell, D. & Kowal, S., "Use of punctuation for pausing : Oral readings by German radio homilists," *Psychological Research*, vol. 48, pp. 93-98, 1986.

[5] Martin, P., "Ponctuation et structure prosodique," *Langue Française*, vol. 172, pp. 99-114, 2011.

[6] Zvonik., E. & Cummins, F., "Pause duration and variability in read texts," in *Proc. 2002 International Conference on Spoken Language Processing*, 2002.

[7] Smith, C. , "Topic transitions and durational prosody in reading aloud : production and modelling," *Speech Communication*, vol.42, pp. 247-270, 2004.

[8] Campione, E. & Véronis, J., « Etude des relations entre pauses et ponctuations pour la synthèse de la parole à partir de texte », in *Proc. 2002 Congrès TALN*, 2002.

[9] Wang, X., Li, A., Yuan, C., "A preliminary study on silent pauses in Mandarin Expressive Speech", in *Proc. 4th Conference on Speech Prosody*, 2008.

[10] Viola, I. & Madureira, S., "The roles of pause in speech expression", in *Proc. 4th Conference on Speech Prosody*, 2008

[11] Sterponi, L., "Reading as involvement with text : Insights from a study of high functioning children with autism", *Rivista di Psicolinguistica Applicata*, vol. 3, pp. 87-114, 2007.

[12] Lacheret, A. & Victorri, B., "La période intonative comme unité d'analyse pour l'étude du français parlé : modélisation prosodique et enjeux linguistiques," *Verbum*, vol.24, pp. 55-73, 2002.

[13] Lacheret, A. & Legallois, D., "Expressivité vocale et grammaire : comment le symbolique construit le prosodique ? " in Gaudemar, M. (eds.), *Les plis de la voix*, pp. 45-54, 2013.

Challenges for Robust Prosody-based Affect Recognition

Heather Pon-Barry, Arun Reddy Nelakurthi

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University, Tempe, Arizona, USA
ponbarry@asu.edu, anelakur@asu.edu

Abstract

Prosody-based affect recognition has great potential impact for building adaptive speech interfaces. For example, in intelligent systems for personalized learning, sensing a student's level of certainty, which is often signaled prosodically, is one of the most interesting states to interpret and respond to. However, robust uncertainty recognition faces several challenges, including the lack of gold-standard labels, and differences in expressivity among speakers. In this paper we explore the intersection of these two issues. We have collected a corpus of spontaneous speech in a question-answering task. Three kinds of certainty labels are associated with each utterance. First, speakers rated their own level of certainty. Second, a panel of listeners rated how certain the speaker sounded. Third, an externally crowdsourced difficulty score is generated for each stimulus (the question). We present a word-level prosodic analysis of individual speaking styles, as they relate to these three different measurements of certainty. Our results suggest that instead of learning one-size-fits-all prosodic models of affect, we might find improvement from learning multiple models corresponding to different speaking styles.

Index Terms: Uncertainty, affect recognition, affect labels, speaking style.

1. Introduction

An exciting goal in human-computer interaction is that of adding human-level emotional behavior to intelligent systems, that is, the ability to perceive a user's emotional state and adaptively respond to it [1]. In speech systems in particular, there has been a lot of work in recent years on detecting a broad spectrum of affective states in speech, including basic emotions [2, 3, 4], frustration [5], charisma [6], uncertainty [7, 8, 9], sleepiness and intoxication [10], and interpersonal stance [11].

There are multiple ways of measuring a speaker's level of certainty. In the existing work on automatic emotion recognition, the most common approach is to measure *perceived* emotion, as annotated by one or more human listeners, producing labels that are by definition subjective [12]. While we treat these labels as a gold standard, we understanding that the subjectivity makes for a challenging classification problem [13]. On the other hand, we can consider *self-reported* certainty, when speakers are asked to rate their own level of certainty. In our prior work, we found that perceived certainty was often *higher* than self-reported certainty [9]. In the same vein, related work on interpersonal stance (friendliness, flirtatiousness, etc.) found that in conversation dyads, self-reported affect was not strongly correlated with perceived affect [11]. In applications such as spoken dialogue systems for tutoring students, we are most interested in knowing a student's *internal* level of certainty.

Prior work has not addressed the question of whether the annotator perceptions or self-reports are an accurate reflection of internal certainty. There is no way to precisely measure internal certainty, but we attempt to address this issue by eliciting speech in a question-answering setting with materials that we hypothesize to be consistently easy or difficult for all individuals. We then generate difficulty scores for each stimulus via crowdsourcing.

In this paper, we present an exploratory analysis of the prosodic characteristics of individual speakers. We find that some speakers produce consistent prosodic expressions of their certainty level, mostly in their pitch, while other speakers show highly inconsistent patterns of speech. Our methodology involves extracting short audio segments of individual words from spoken answers to questions that varying in the speaker's level of certainty. Similar to recent work that has applied principal components analysis (PCA) to large sets of low-level acoustic-prosodic features [14], we identify a set of 10 principal components from a large set of word-level prosodic features. We then use the smaller set of prosodic features to learn several decision trees for each speaker and analyze the manner and consistency of prosodic expression as a way to gauage individual speaking styles.

2. Harvard Uncertainty Speech Corpus

The speech data that we use in this experiment comes from the Harvard Uncertainty Speech Corpus. This section gives an overview of the Harvard Uncertainty Speech Corpus (Section 2.1), the speech elicitation process (Section 2.2), and the methods of annotating and approximating speaker certainty from the hearer's perspective (Section 2.3), the speaker's perspective (Section 2.4), and according the difficulty of the question (Section 2.5).

2.1. Speech Data from Uncertainty Corpus

The Harvard Uncertainty Speech Corpus contains spoken utterances and level of certainty annotations from three question-answering domains [15]. In this paper, we use the *handwritten digit* section of corpus. The utterances were recorded in a lab, in a question-answering setting. The questions and answers are of the form below.

Q: Which train leaves Los Angeles and at what time does it leave?

A: Train seven leaves Los Angeles at 1:27.

In this experiment, we examine the first two words of such utterances, for example, "train seven" or "train two".

The Harvard Uncertainty Speech Corpus contains the audio corresponding the to answers (not the questions). A notable

feature of the utterances in the corpus is that when a speaker is uncertain, the uncertainty can be attributed to a particular *word or phrase* in the utterance. The entrire corpus contains 1700 utterances, roughly 150 minutes of speech. The handwritten digit section of the corpus contains 1100 utterances, about 90 minutes of speech. Detailed descriptions of the corpus are available in previously published works [9, 16].

2.2. Background on Method of Speech Elicitation

The speech elicitation materials are designed in a way that controls the difficulty of the stimulus. This is achieved by asking participants to engage in a task that necessitates speaking a spontaneous utterance that incorporates reading handwritten digits that vary in how legible they are. The digit images are drawn from the MNIST database of handwritten digits [17]. The materials for eliciting speech are designed so that participants would speak the selected MNIST digit aloud in the context of answering a question. The handwritten digit images are embedded in illustrations of train routes connecting two U.S. cities, where the handwritten digits indicate the train number. An example illustration is shown in Figure 1.

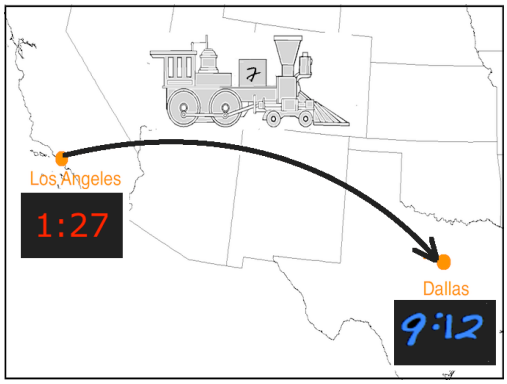


Figure 1: Example speech elicitation illustration featuring an ambiguous handwritten digit image, the train number.

We collected speech from twenty-two native English speakers. At the start of the data collection experiment, participants read a task scenario explaining why they are deciphering handwritten train conductor notes and answering questions about them. For each train route illustration, participants are asked a single question. The participants respond aloud, speaking spontaneously. Their word choice is influenced by a warm-up task where they are given answers to read aloud. This lets us have influence over the length and lexical content of the utterances without the participant explicitly reading aloud.

The method for eliciting uncertain speech is a modification the method used in a previous collection of affective speech [9]. In that work, we did not attempt to control the speaker's level of certainty. As a result, there was no way to verify whether a speaker's self-reported level of certainty was an accurate reflection of his or her actual certainty.

2.3. Certainty Labels from Hearer's Perspective

Each utterance in the corpus is annotated with the level of certainty from a hearer's perspective. We collected annotations from a panel of six human judges. Every annotator listened to and rated the entire set of 1100 utterances. They rated level of certainty on a 1 to 5 scale (1 = very uncertain, 5 = very certain). They did not see any contextual information such as the handwritten images. For each utterance, we consider the mode (average) of the six annotator labels to be the certainty label from the hearer's perspective. The distribution of certainty labels from the hearer's perspective in the corpus is shown in Figure 2.

The agreement among the six annotators highlights the subjective nature of the hearer-centric affect labeling paradigm. Across all pairs of annotators, we find an average pairwise agreement of 54.3%, average Cohen's kappa of 0.235, and average Spearman correlation coefficient of 0.494. If we look only at the pair of annotators with the highest agreement, we see much higher values: pairwise agreement of 74.1%, Cohen's kappa of 0.407, and Spearman correlation of 0.62.

2.4. Certainty Labels from Speaker's Perspective

Each utterance in the corpus is annotated with the level of certainty from the speaker's perspective. The speakers are asked, "How certain were you about the answer you just gave?" during the speech elicitation process. They rate their level of certainty on a 1 to 5 scale (1=very uncertain, 5=very certain). The distribution of certainty labels from the speaker's perspective in the corpus is shown in Figure 2.

2.5. Certainty Labels from Image Difficulty Score

We attempt to control the speaker's actual level of certainty by designing stimuli that are uniformly difficult or easy and we then use crowdsourcing to obtain a difficulty score for each stimulus. Each utterance in the corpus has a legibility score associated with the handwritten digit (the train number) that was used to prompt the question. We used Amazon's Mechanical Turk [18, 19] to collect human judgements from which we generate image legibility scores. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks* (HITs). We showed Turkers a digit image and instructed them to identify the digit using a drop-down menu. Each digit was labeled by 100 human workers. Details of the HIT design are available in previously published work [16].

The legibility score for each image is defined as 1 minus the Shannon entropy of the human label distribution:

$$\text{Legibility} = 1 - \big[- \sum_{i=1}^{N} P(x_i) log P(x_i) \big]$$

Thus, scores fall in the range [0,1]. A score of 1 has an entropy of 0 and indicates high legibility (all 100 people choose the same label). The handwritten digit in Figure 1 has a legibility score of 0.75. The distribution of difficulty scores (legibility scores) for the stimuli used in eliciting the speech data is shown in Figure 2.

3. Experiment and Results

In this experiment we analyze the speaking styles of individual speakers. We explore whether these individuals are prosodically expressive regarding level of certainty, and if they display consistency in their prosodic expression. Because of design of the corpus, each speaker utters phrases such as "train one" or "train two" multiple times, with differing levels of certainty. Figure 3 shows the spectrograms of three utterances from the same speaker saying "train two" while feeling uncertain, neutral, and certain. Because the corpus contains such sets of lexically-identical phrases, in this experiment we compare

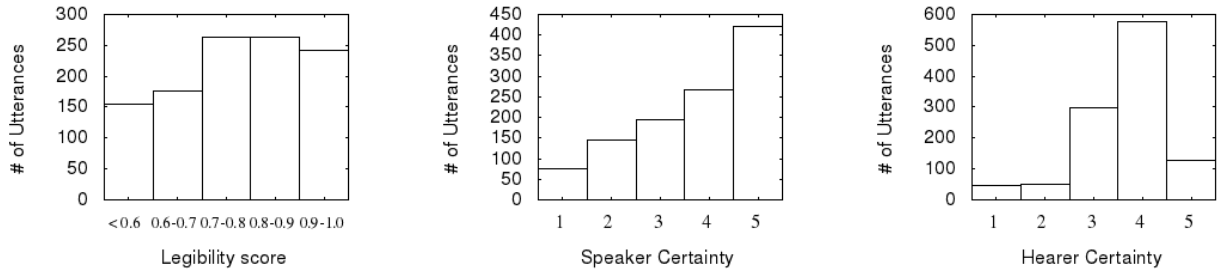


Figure 2: Three histograms: (left) distribution of difficulty scores for the stimuli that prompted the utterances in the corpus, (middle) distribution of certainty labels from the speaker's perspective, (right) distribution of certainty labels from the hearer's perspective.

word-level prosodic features. However, of the ten possible digits, only some are repeated with enough frequency to be analyzed. From the larger corpus, we identify a set of speakers and digits such that,

- each speaker utters each digit 3 or more times, *and*
- for each speaker-digit combination, the certainty labels are distributed among certain, neutral, and uncertain.

This yields a set of 408 utterances representing eight speakers and six digits. Table 1 shows the utterance counts for these eight speakers. For example, in our corpus, speaker a says "train one" five times and says "train two" eight times.

Table 1: Number of utterances that contain the phrases, "train one", "train two", "train three", "train five", "train seven", and "train nine", for a subset of speakers in the corpus.

	Num instances of "*train…*" per speaker						
Speaker	*one*	*two*	*three*	*five*	*seven*	*nine*	**Total**
a	5	8	4	7	8	8	**40**
b	4	8	6	6	9	10	**43**
c	4	11	4	6	6	12	**43**
d	3	8	5	6	8	12	**42**
e	3	10	5	6	7	8	**39**
f	4	7	7	5	7	11	**41**
g	3	6	5	6	9	8	**37**
h	6	7	3	7	9	7	**39**

3.1. Unit of Analysis

Because the corpus contains repeated instances of specific words, spoken with different levels of certainty, we perform prosodic analysis at the *word level*. The segments of interest are the train numbers, which correspond to the MNIST handwritten digits. The word-level audio segments are generated semi-automatically. We use the CMU Sphinx speech recognition toolkit to automatically transcribe each utterance and generate word alignments. The audio segments are manually verified and errors are manually corrected.

3.2. Prosodic features

Initially, we extract 230 prosodic features from each audio segment. We use the openSMILE feature extraction toolkit [20] with the `emobase` config file. The features include low-level descriptors (F0, F0-envelope, intensity, loudness, voice quality, and zero-crossing rate), functionals, and delta regression coefficients for smoothed feature contours.

We use principal component analysis to identify 10 principal prosodic components of the digit word segments in our data (using the entire corpus—word segments from all utterances of all speakers). PCA is performed using the WEKA toolkit [21]. The ranked results are aggregated and a set of 10 principal components are identified for further analysis. The resulting 10 features are listed below (delta features indicated by ^{d}).

1. F0 average^d
2. F0 range^d
3. F0 slope^d
4. F0 skewness^d
5. F0 envelope max^d
6. Intensity average^d
7. Intensity skewness^d
8. Intensity minimum
9. Probability of voicing^d
10. Zero-crossing rate^d

3.3. Speaker analysis

In order to understand how these features are related in predicting the level of uncertainty in utterances, we have made use of decision tree learning. Considering certainty labels from the speaker's perspective (3 classes), we learn separate decision tree classifiers for each speaker-word combination. That is, we learn a decision tree for [$speaker = a$, $word =$ "one"], [$speaker = a$, $word =$ "two"], and so on. In total, we learn six decision trees for each speaker. The maximum depth of decision trees is 3. We used the WEKA toolkit [21] implementation of C4.5 algorithm (J48).

For each speaker, we evaluate whether the learned decision criteria are consistent across all six words. In other words, we ask: for speaker a, are the informative prosodic features consistent for the word "one", the word "two", the word "three" and so on. We then do the same analysis for certainty labels from the hearer's perspective, and for the difficulty score approximation of certainty. Table 2 shows the speaker-specific consistency results. The 10 prosodic features are collapsed into three groups: pitch (#1-#5), intensity (#6-#8), and voice (#9-#10). Separate results are shown for the three approximations of certainty: labels from the speaker's perspective, labels from the hearer's perspective, and difficulty of the question (legibility

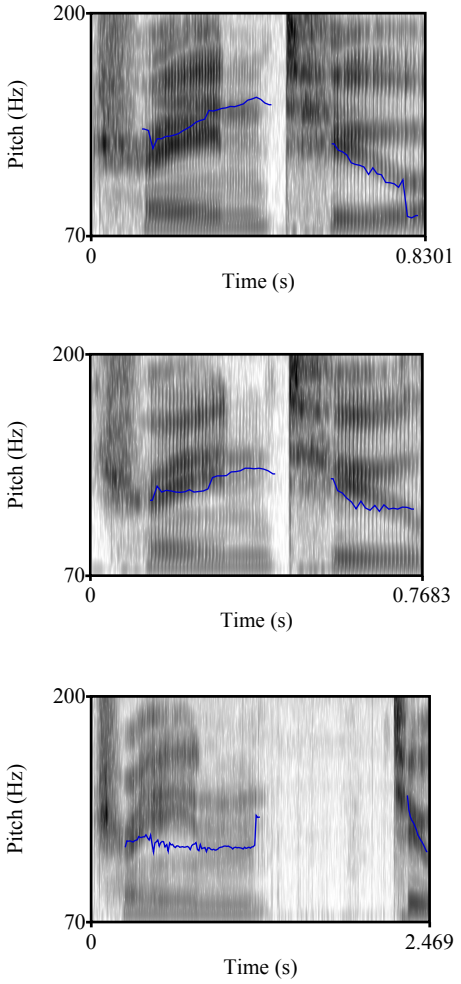


Figure 3: Three instances of a single speaker saying "train two" with varying affect: certain (top), neutral (middle) and uncertain (bottom). The pitch estimate (blue line) is overlaid atop the spectrogram.

score). For speakers that used consistent modes of prosodic expression, checkmarks indicate the class of prosodic feature that distinguished the certain, neutral, and uncertain words.

4. Discussion

We see two primary observations from this exploratory analysis. First, among the prosodic features that we analyzed, features related to pitch are the strongest differentiators between certain and uncertain affect, voice features are second strongest. Second, this analysis, though preliminary, suggests that some speakers consistently display their certainty through their prosody while others are inconsistent. We hypothesized that inconsistent speakers would be inconsistent across all three certainty metrics (speaker, hearer, and legibility). The results show only a small amount of support for this: speakers f and h are inconsistent under both the speaker and hearer metrics. The fact that there is no overlap between the inconsistent speakers in the bottom section of Table 2 and the other two sections indicates that the difficulty score metric may be too coarsely defined

Table 2: Speaker-specific prosodic modes for conveying uncertainty. Certainty is approximated in three ways: the speaker's perspective, the hearer's perspective, and the difficulty of the question.

	Speaker							
	a	b	c	d	e	f	g	h
Certainty labels: *speaker*								
Pitch	✓		✓	✓	✓			
Intensity					✓			
Voice	✓		✓		✓			
Inconsistent		✗				✗	✗	✗
Certainty labels: *hearer*								
Pitch		✓	✓		✓		✓	
Intensity								
Voice			✓					
Inconsistent	✗			✗		✗		✗
Difficulty of question								
Pitch	✓	✓		✓		✓	✓	✓
Intensity								
Voice								
Inconsistent			✗		✗			

(with two binary classes), or that it may not be as aligned with speaker and hearer certainty labels as we had posited.

We see that some speakers, e.g., speakers a and c, have similar manners of conveying certainty. Our next steps involve a clustering analysis to explore whether natural clusters can explain the variation seen among those speakers who convey their certainty in their prosody.

5. Conclusion

This paper presents an exploratory analysis of the prosodic characteristics of individual speaking styles, as they relate to three different measurements of certainty. We find that some speakers have consistent ways of conveying their level of certainty prosodically, while other speakers are inconsistent. Among the prosodic signals, pitch-related features are the strongest. Across the three different measures of certainty: speaker's perspective, hearer's perspective, and item difficulty, we find more varying speaker behaviors, suggesting the need for further analysis.

This work is of broad relevance to researchers studying affect recognition. Robustly recognizing affect, and especially subtle affective-cognitive states such as uncertainty, faces many challenges. It is not surprising that speakers have different ways of prosodically expressing affect. In this paper, we show that some speakers produce consistent prosodic signals of certainty, mostly in their pitch, while other speakers show highly inconsistent patterns of speech. Instead of using the same techniques for detecting affect in all speakers, there is great potential utility in adaptive affect detection. For example, if a person is very inconsistent in their speech signals, then an intelligent, multimodal system should direct its inference efforts toward signals from other modalities such as lexical content or facial expressions. On the other hand, if a speaker is prosodically expressive, adaptive systems in the future may dynamically determine which prosodic signals to weight more strongly.

6. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, January 2001.

[2] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] R. Fernandez and R. Picard, "Classical and novel discriminant features for affect recognition from speech," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 473–476.

[4] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[5] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 2037–2040.

[6] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proceedings of Interspeech*, 2005, pp. 513–516.

[7] J. Liscombe, J. Hirschberg, and J. Venditti, "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Interspeech*, Lisbon, Portugal, 2005, pp. 1837–1840.

[8] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech*, Brisbane, Australia, 2008, pp. 74–77.

[9] H. Pon-Barry and S. M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 251753, 2011.

[10] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proceedings of Interspeech*, 2011, pp. 3201–3204.

[11] R. Ranganath, D. Jurafsky, and D. A. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech and Language*, vol. 27, no. 1, pp. 89–115, 2012.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[13] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

[14] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea, 2012, pp. 198–206.

[15] H. Pon-Barry, S. M. Shieber, and N. Longenbaugh, "Eliciting and annotating uncertainty in spoken language," in *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC)*, 2014.

[16] H. Pon-Barry, "Inferring speaker affect in spoken natural language communication," Ph.D. dissertation, Harvard University, 2013.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[18] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, 2010.

[19] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, vol. 44, pp. 1–23, 2011.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia, MM '10*, 2010.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

Going ba-na-nas: Prosodic analysis of spoken Japanese attitudes

Dominique Fourer[1], *Takaaki Shochi*[2], *Jean-Luc Rouas*[1],
Jean-Julien Aucouturier[3], *Marine Guerry*[4]

[1]LaBRI - CNRS UMR 5800, Univ. Bordeaux 1, France
[2]CLLE-ERSSàB UMR5263 CNRS, Bordeaux, France
[3]IRCAM - CNRS UMR9912 - UPMC, Paris, France
[4]Univ. Bordeaux Montaigne, France

[1]firstname.lastname@labri.fr, [2]Takaaki.Shochi@u-bordeaux3.fr, [3]aucouturier@ircam.fr

Abstract

The aim of this paper is to examine cues for prosodic characterization of attitudes in Japanese. This work is based on previous studies where 16 communicative social affects were defined. The audio signal parameters (fundamental frequency, amplitude and duration) of previously recorded Japanese attitudes, are statistically analyzed. Interesting interactions among the parameters, the gender and the expression of specific attitude (e.g. politeness) were found, and we report on which parameters most significantly characterize each attitude.
Index Terms: speech, prosody, attitude, social affect, emotional speech, Japanese language

1. Introduction

The prosodic expressions of social affects, or attitudes as defined by [1], are a mean used by speakers to drive the illocutionary force of their intended speech acts [2] in face-to-face communication. Such choices are partly linked with the speaker's own proficiency in the spoken language, her/his personality, gender and the communication context which are also constrained at the linguistic level. Thus, each language has specific formulae or conventional prosodic variations for specific interaction contexts. Usually, studies investigating such kinds of prosodic variations rely on stereotypic stimuli [3, 4, 5, 6]. One common difficulty in studies which aim to compare the prosody of social affects is linked to a trade-off between the high sound quality required for acoustic analysis, the need for a neutral lexical content of the studied sentences (ideally identical sentences for all the studied affects), the search for spontaneity of the expressions, and a clear labeling of the communicative goals of the speaker. Most of the cited studies use laboratory corpora. Typically, adhoc sentences are recorded by speakers trying to read a sentence and reproduce a given expressivity. To enhance the spontaneity of these expressions and to facilitate the speaker's task, [7] proposes to place target sentences in affectively loaded texts. Similarly, [6] recorded attitudinally-neutral sentences embedded into dialogues that prepare the speaker to perform an adequate expression for the target sentence. The approach used during this research builds on these works. In order to study the expressive strategies used by speakers of varying linguistic backgrounds, communicative situations have been set-up so they can be plausibly used in different languages. The analysis method used in this paper also shares similar objectives to be applied to any language. We thus decided to use only low-level prosodic descriptors: the values of the fundamental frequency (F_0), the Root Mean Square (RMS) amplitude (computed on vowels) and the syllable duration. Using these features for statistical analysis with Repeated-Measure ANalysis Of VAriance (RM-ANOVA), we are able to determine to which extent each feature may contribute to the differentiation of attitudes. This paper is organized as follows: the framework we used for a speaker express social affects is described in Section 2. The recording set-up procedure and the feature extraction methods are respectively described in Sections 3 and 4. Finally, the results from the statistical analysis are presented in Section 5 and discussed in Section 6.

2. Social contexts for expression of attitudes

In order to immerse subjects in the context, a scenario was set up for each attitude, and the subject was requested to engage in a short dialogue that would lead to the production of target sentences with the native speaker. For the current experiment, 16 contexts have been selected, corresponding to a set of attitudes used in [8, 9] for different languages. Some of these contexts do not have lexical equivalents in all languages, as the corresponding communication situations have not been conventionalised in that particular culture. It is the case for example of the Japanese notion of *kyoshuku*, described by [10] as "corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker consciousness of the fact his/her utterance of request imposes a burden to the hearer". For instance, *Kyoshuku* has no lexical equivalent in English. Meanwhile, "walking on eggs" corresponds to a certain extent to this concept. The following 16 social affects were used in the present corpus: Admiration (ADMI)), Arrogance (ARRO), Authority (AUTH), Contempt (CONT), Doubt (DOUB), Irony (IRON), Irritation (IRRI), Neutral declarative sentence (DECL), Neutral question (QUES), Obviousness (OBVI), Politeness (POLI), Seduction (SEDU), Sincerity (SINC), Surprise (SURP), Uncertainty (UNCE), Walking on eggs (WOEG). They are defined by prototypical situations with the social relationship of the two interlocutors specified – as well as the communicative goal of the speaker (see [11] for details). For all situations, a short neutral target sentence has been used to record the respective prosodic expressions: "A banana". In order to elicit these target sentences in each context, small dialogues were written (cf. [6]), that take place in the prototypical context described above, and that end with the target sentence. During the recordings, each speaker (A) has an active interlocutor (B) who interacts with her/him in order to enhance the naturalness of the communication situation, and to ease the production of realistic expressions. Speakers are indeed not asked to produce an isolated

sentence with an identified attitude (e.g. seduction or authority), but rather to immerse in a scenario. For instance, the situation is the following for "walking on eggs":

- Your boss (Speaker B) has asked you (speaker A) to be in charge of setting up a room for a big conference. Your boss is a super compulsive guy who needs to have everything done just right, and gets easily angered if things are not perfect. Your boss walks into the room where the big conference is to be held, and in the wastebasket, there is a half-eaten banana. He is furious.

Currently, these situations have been adapted to three languages: American English, Japanese and French. The present paper focuses on the Japanese results, as performed by native speakers.

3. Recording procedure

A set of 19 Japanese native speakers (11 females, 8 males) have been recorded. Most speakers were recruited amongst university students and were paid for their performance. The recordings took place in a sound-treated room at Waseda University, Japan. The sound was captured by an *Earthworks QTC1* omnidirectional microphone, placed at one meter from the mouth of the speaker (this distance was chosen to limit the influence of the speaker movements on the sound level). The microphone level was calibrated before each recording session using a *Bruel & Kjaer* acoustical calibrator, thus the sound pressure level can be corrected after recording to a level comparable across all speakers. The target sentence "banana" was then manually searched for across the recorder corpus, isolated and extracted into individual files. Any speech utterances from speaker B occurring during the expressive gesture of speaker A performing the target sentence were removed from the sound track (none overlapped with their speech). Due to the interactive nature of the recording, some spontaneous changes were observed on the target sentences: typically "banana" sentence with interjections, such as "hmm", "er", "oh", etc., together with the target sentences. Each speaker recorded one utterance of the word for each of the 16 attitudes, resulting in a total of 304 stimuli. These were stored as 16 kHz, 16-bit WAV files. Each stimulus was trimmed to discard the beginning and the ending silence. The wave file of each stimulus was hand-labeled at a phonetic level using the PRAAT software [12].

4. Feature extraction

We characterized each stimulus with its F_0, amplitude and duration parameters.

4.1. Fundamental frequency

The F_0 parameter measured in Hertz is estimated using the SWIPE algorithm [13] with a 10 ms sampling interval. For statistical analysis (see Section 5), only F_0 values at time indices corresponding to the vocalic phonemes are kept, and averaged to give a mean F_0 value per vowel (3 values per stimulus). We also applied the MultiDimensional Scaling algorithm (MDS) [14] to the complete series of F_0 values normalized by the mean F_0 of the speaker (see Section 5.4). Thus, MDS allows a graphical interpretation of the distance between the attitudes which depends on the computed correlation between the F_0 series.

4.2. Amplitude

The amplitude is estimated using the RMS function where the signal is windowed to result in 20 ms frames with 50% overlap. Thus, the RMS amplitude is computed on each frame from the normalized values (in $[-1;+1]$) of the signal samples. As before, for statistical analysis, only RMS values at time indices corresponding to the vocalic phonemes are kept, and averaged to give a mean RMS amplitude value per vowel (3 values per stimulus).

4.3. Duration

Additionally, for statistical analysis, each syllable duration measured is computed as the sum of its consonant and vocalic parts, in milliseconds, based on the manual segmentation of the stimuli.

5. Results of the statistical analysis

The statistical significance of prosodic differences is measured between attitudes and gender, separately for the F_0, amplitude and duration parameters. For each parameter, each stimulus is considered as a series of three repeated measures, corresponding to the mean parameter value of the stimulus for the three successive vocalic phonemes. For each parameter, we conducted a RM-ANOVA, with attitude (16) $\times$ phoneme (3) as within-subject factors and gender (2) as between-subject factor. For the remainder of this section, $F(a,b)$ denotes the computed statistic which is assumed to follow a Fisher probability density function of parameters a and b. p denotes the p-value which results from the RM-ANOVA and M is used to denote the mean value of the considered parameter.

5.1. Fundamental frequency

The RM-ANOVA on F_0 values reveals a strong statistical interaction of attitude $\times$ time: $F(30,210)=8.12$, $p<0.001$, indicating that the temporal profile of F_0 varies significantly across attitudes. As shown in Figure 2(a), the interaction is most noticeable between F_0 values of the first and second vowel of each utterance: while the majority of attitudes are associated with lower-pitch in the second vowel ("V-shape"), attitudes AUTH, DECL, QUES and (to a lesser extent) SINC are associated with higher-pitched second vowels ("inverted V-shape").

To further characterize this behavior, we extract the normalized F_0 ratio between the second and first vowel of each stimulus (in % increase) - see Figure 2(d). A RM-ANOVA (with attitude (16) as a within-, and gender as a between-subject factor) reveals a main effect of attitude on this ratio: $F(15,225)=13.6$, $p<0.001$. Attitudes AUTH(+26%), DECL (+17%), QUES (+34%) show an increase in F_0 that is more important than the other measures ($p<0.001$, Bonferroni-corrected). Attitude SINC (+7%) also shows an increase in F_0 values more important than for CONT (-14%) ($p=0.01$, Bonferroni-corrected).

Most interestingly, the F_0 differences between attitudes have a marked interaction with gender: $F(30,210)=1.95$, $p=0.003$, indicating that different F_0 patterns are used by male and female to convey the same attitude. Remarkably, this interaction with F_0 is not characterized by the first (attitude $\times$ gender: $F(15,240)=1.36$, $p=0.16$), second (attitude $\times$ gender: $F(15,240)=1.83$, $p=0.03$) or third vowel individually (attitude $\times$ gender: $F(15,240)=0.38$, $p=0.98$), but rather on the difference between values on each vowel.

As seen in Figure 2(d), there are significant gender dif-

ferences in particular in the second-to-first F_0 ratio, both overall (attitude × gender: $F(15,225)=2.58$, $p=0.001$) and for individual attitudes ADMI (male: +5%, female: -11%, p=0.03), AUTH (male=+10%, female: +37%, $p=0.0008$) and QUES (male=+49%, female=+23%, $p=0.001$; all: Fisher LSD posthocs).

5.2. Amplitude

The RM-ANOVA on amplitude values reveals a main effect of gender: $F(1,17)=24.5$, $p=0.00012$, indicating that recordings by female speakers are louder than males; a main effect of time: $F(2,34)=88.51$, $p<0.001$, with first ($M=.00071$) and second vowels ($M=.00066$) both twice louder than the third vowel ($M=.00034$);and a main effect of attitude: $F(15,255)=1.99$, $p=0.015$, showing that some attitudes are generally louder than others, regardless of vowel and gender.

Attitudes ADMI ($M=.00063$), POLI ($M=.00060$), SINC ($M=.00060$) and WOEG ($M=.00059$) are louder than AUTH ($M=.00054$), CONT ($M=.00052$), IRRI ($M=.00052$), QUES ($M=.00054$), SEDU ($M=.00053$), SURP ($M=.00056$) and UNCE ($M=.0052$) at the $p<.05$ level (Fisher LSD test, non Bonferroni-corrected).

There is an interaction of time × gender: $F(2,34)=3.51$, $p=0.04$ - second vowels were more quiet for males than females -, but no interaction of attitude × gender: $F(15,255)=0.88$, $p=0.58$.

Most remarkably, there is a strong interaction of attitude × time: $F(30,510)=6.57$, $p<.001$, showing that, as for F_0, different temporal profiles of amplitudes are used to convey different attitudes (see Figure 2(b)), and, as for F_0, differences are particularly notable in the ratio of amplitude at the first and second vowels: while second vowels were quieter than first for the majority of attitudes, they were louder for attitudes AUTH, DECL and QUES.

However, contrary to F_0, there was no interaction of this effect with gender: $F(15,255)=0.33$, $p=0.99$ (attitude × gender interaction, RM-ANOVA on second-to-first amplitude ratio). Similar amplitude patterns are used by male and female speakers to convey the same attitudes (see Figure 2(e)).

5.3. Duration

The RM-ANOVA on syllable durations reveals a marginal main effect of gender: $F(1,17)=4.71$, $p=0.04$ - females were slower ($M=137$ ms) speakers than males ($M=117$ ms); a main effect of time: $F(2,34)=95.6$, $p<0.001$, with final syllables twice as long ($M=188$ ms) as the first ($M=99$ ms) and second syllables ($M=98$ ms); and a main effect of attitude: $F(15,255)=13.0$, $p<.001$, showing that some attitudes were generally slower than others, regardless of syllable and gender.

Attitudes AUTH ($M=162$ ms), CONT ($M=166$ ms), DECL ($M=171$ ms), DOUB ($M=155$ ms), QUES ($M=147$ ms) and SURP ($M=165$ ms) were slower than the others ($p<.05$, Bonferroni-corrected). This main effect of attitude was seen both on durations of the first syllable: $F(15,255)=11.09$, $p<.001$, second syllable: $F(15,255)=9.48$, $p<.001$ and third syllable: $F(15,255)=13.08$, $p<.001$.

There was no interaction of time × gender: $F(2,34)=1.57$, $p=0.22$, and no interaction of attitude × gender: $F(15,255)=1.17$, $p=0.29$. However, as mentioned below, there was a strong interaction of attitude × time: $F(30,510)=12.07$, $p<.001$, showing that, as for F_0 and amplitude, different patterns of successive syllable durations were used to convey different attitudes (see Figure 2(c)).

As for F_0 (but not for amplitude), this interaction is significantly related with gender: $F(30,510)=2.43$, $p=<.001$. The gender difference was mainly seen on the duration of the third syllable: $F(15,255)=1.89$, $p=0.02$ (see Figure 2(f)); gender differences were only marginal for the duration of the first syllable: $F(15,255)=1.68$, $p=0.05$, and not significant for the second syllable: $F(15,255)=1.04$, $p=0.4$.

5.4. Multidimensional scaling

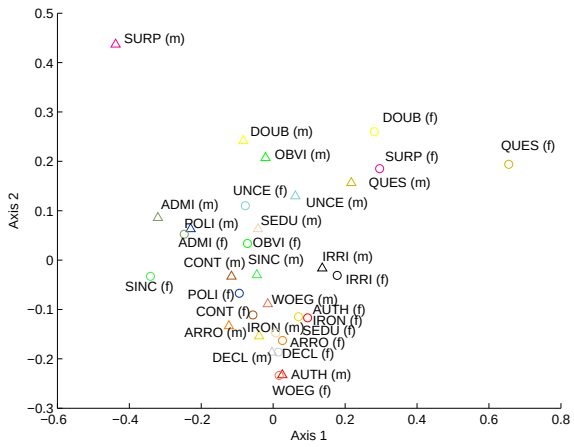
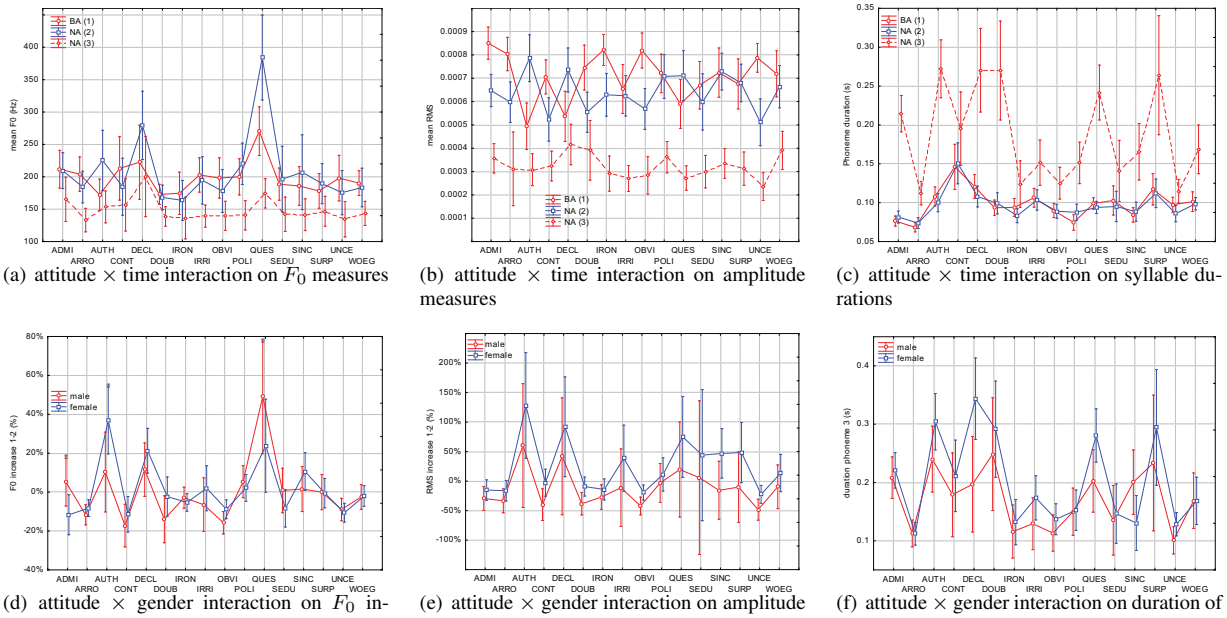


Figure 1: Result of the MDS algorithm based on the correlation distance of the F_0 profiles (circle: female, triangle: male).

The result of the MDS algorithm [14] applied on the normalized F_0 profiles associated to the entire phrase "banana" and computed on the corpus is presented in Figure 1. This figure shows the center of mass for each attitude where male and female are separated. The distance between each point is associated to the correlation distance computed between the profiles associated to the attitude of each speaker. For the normalization, each estimated F_0 profile is divided by the averaged F_0 related to the speaker to obtain a modulation function centered on the unity. For the duration, all the estimated profiles are rescaled to a series of 100 frames fitting the mean duration of the corpus which is 531.25 ms

6. Discussion

This study on Japanese spoken language investigated acoustic characteristics of various social affects in order to identify what are the similar and different prosodic patterns for various attitudes. For this study we used the F_0 and the amplitude parameters which are objective signal parameters used to describe the prosody of each expressed attitude. Statistical results show that the interaction between attitudes and gender is observed only for the F_0 parameter. It indicates that the gender differences of attitudinal expressivity are more characterized by F_0 rather than amplitude or duration. However, we have observed that female expressiveness is not only characterized by a higher pitch but also by a higher amplitude and longer syllable durations than males. We also identified that Japanese politeness expressions (ADMI, POLI, SINC, WOEG) are louder than others, including impoliteness expressions (AUTH, CONT and IRRI). Concerning the duration parameter, the Japanese language is



(a) attitude × time interaction on F_0 measures

(b) attitude × time interaction on amplitude measures

(c) attitude × time interaction on syllable durations

(d) attitude × gender interaction on F_0 increase

(e) attitude × gender interaction on amplitude increase

(f) attitude × gender interaction on duration of syllable 3

Figure 2: Differences in F_0 temporal profiles between attitudes 2(a) and gender 2(d). 2(a): mean F_0 value (in Hz) for each vowel of (b)a-(n)a-(n)a for the 16 attitudes. 2(d): difference of mean F_0 between the second and first vowels (in % increase) for the 16 attitudes and for both genders. Error bars in 2(a) and 2(d) correspond to 95% confidence intervals. Asterisks in 2(d) mark significant gender difference ($p<0.05$, Fisher LSD posthoc). Differences in RMS amplitude temporal profiles between attitudes 2(b) and gender 2(e). 2(b): mean amplitude value for the three vowels of (b)a-(n)a-(n)a for the 16 attitudes. 2(e): difference of mean amplitude between the second and first vowels (in % increase) for the 16 attitudes and for both genders. Error bars in 2(b) and 2(e) correspond to 95% confidence intervals. Differences in syllables duration between attitudes 2(c) and gender 2(f). 2(c): duration of each of the three syllables ba-na-na for the 16 attitudes. 2(f): duration of the third syllable for the 16 attitudes and for both genders. Error bars in 2(f) and 2(f) correspond to 95% confidence intervals.

well known to be mora-timed which means that each mora (taking account of two moras for long vowels, geminate obstruents and nasal /N/) is of similar duration because of the compensation of segmental variation of duration inside the mora-structure [15, 16]. However these results show an important temporal variation among attitudinal expressions. A variation between the two first vowels and the 3rd vowel was especially observed (i.e. 3rd vowel was almost twice as long as 1st and 2nd vowel). It suggests that social affects may change duration adjustment of Japanese rhythmic structure, and it confirms a previous work [17] which finds that the duration of this 3rd vowel is an important factor to make a difference in various attitudinal expressions. Moreover, a correlation between attitude and time based on the F_0 values was observed. Standard Japanese is described as a pitch accent language which is characterized by a rising F_0 at the beginning of the phrase, and an important pitch fall after an accented vowel [18]. The recorded sentence "Banana" has an accentual nucleus on the 1st mora. Thus, we expect to observe F_0 rising on the 1st vowel with a fall at the end of the 1st vowel. However results show that the phrasal initial rising of F_0 until the peak (F_0 max) vary according to attitudes [19]. Our assumption is that the timing of F_0 peak whether it comes on the 1st or the 2nd vowel is correlated with speech rate and amplitude. According to our data , it seems that some attitudinal expressions (AUTH, QUES and DECL) where F_0 peak comes later (on the 2nd vowel) are related with slower speech rate and with an amplitude peak which comes later (on the 2nd vowel) as opposed to the majority of attitudes which are associated with lower pitch in the second vowel ("V-shape"). MDS analysis for F_0 values of all attitudes identified 3 different categories: polite expressions, impolite expressions and dubitative expressions. The first category consists of 4 impolite expressions (AUTH, ARRO, IRON, IRRI) plus 2 other attitudes (DECL, WOEG). The second category is composed of 4 polite expressions (ADMI, SINC, POLI, SEDU (males only) plus 2 attitudes (OBVI, UNCE). It is important to note that WOEG which is akin to the Japanese polite expression of *Kyoshuku* is located in the category of impolite expressions. These results confirm previous work on Japanese politeness expressions [9, 20]. The expression of WOEG may be differentiated from impolite expressions because it does not have the same voice quality characteristics to the Japanese (polite) expression of *Kyoshuku* [9]. Contrary to the similarity of F_0 values of SEDU for males and females given from RM-ANOVA (see Figure 2(d)), MDS analysis identified that female values for this attitude are quite different from the male ones according to Figure 1, the female speakers' circle is located in the category of impolite expressions, which is different from the male one which is located in the politeness category. Although the pitch of SEDU is similar to impolite expressions, the voice quality of this attitude is softer, and therefore SEDU may not be perceived as impolite expression. For future work, a comparison with non-native speakers will also be done to examine prosodic similarities among different languages.

7. Acknowledgments

Thanks to Donna Erickson for her useful comments which helped us to improve this paper. Thanks to Mariko Kondoa and Sylvain Detey for the recording of the Corpus at the University of Waseda. This research was partly supported by the French ANR PADE project and the Bordeaux PEPS IDEX/CNRS project.

8. References

[1] A. Wichmannn, "The attitudinal effects of prosody, and how they relate to emotion," in *Proc. ISCA Workshop on Speech and Emotion*, 2000, pp. 143–148.

[2] I. Fonagy, E. Bérard, and J. Fonagy, "Clichés mélodiques," *Folia Linguistica*, vol. 17, pp. 153–185, 1984.

[3] H. Fujisaki and K. Hirose, "Analysis and perception of intonation expressing paralinguistic information in spoken Japanese," in *Proc. ESCA Workshop on Prosody*, Sep. 1993.

[4] Y. Morlec, G. Bailly, and V. Aubergé, "Generating prosodic attitudes in French: Data, model and evaluation," *Speech Communication*, vol. 33, no. 4, pp. 357–371, 2001.

[5] J. A. de Moraes, "The pitch accents in Brazilian Portuguese: analysis by synthesis," in *Proc. Speech Prosody*, 2008, pp. 389–397.

[6] G. Wentao, T. Zhang, and H. Fujisaki, "Prosodic analysis and perception of Mandarin utterances conveying attitudes," in *Proc. Interspeech*, Aug. 2011, pp. 1069–1072.

[7] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, no. 3, pp. 414–429, 2012.

[8] T. Shochi, A. Rilliard, V. Aubergé, and D. Erickson, *The role of prosody in affective speech*. Peter Lang, 2009, vol. Linguistic Insights 97, ch. Intercultural perception of English, French and Japanese social affective prosody, pp. 31–59.

[9] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson, and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and speech*, vol. 52, no. 2–3, pp. 223–243, 2009.

[10] T. Sadanobu, "A natural history of Japanese pressed voice," *Journal of the Phonetic Society of Japan*, vol. 8, no. 1, pp. 29–44, 2004.

[11] A. Rilliard, D. Erickson, T. Shochi, and J. A. D. Moraes, "Social face to face communication - American English attitudinal prosody," in *Proc. Interspeech*, Aug. 2013.

[12] P. Boersma and D. Weenink. (Version 5.3.32 retrieved 17 October 2012 from http://www.praat.org/) Praat: doing phonetics by computer [computer program].

[13] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 124, pp. 1638–1652, 2008.

[14] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Generalized multidimensional scaling: a framework for isometry-invariant partial surface matching," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 5, pp. 1168–1172, 2006.

[15] M. S. Han, "Acoustic manifestations of mora timing in japanese," *Journal of acoustical society of America*, vol. 96, pp. 73–82, 1971.

[16] Y. Sagisaka and Y. Tohkura, "Phoneme duration control for speech synthesis by rule," *IEICE Trans.*, vol. 67, no. 7, pp. 629–636, 1984.

[17] K. Maekawa and T. Kagomiya, "Influence of paralinguistic information on segmental articulation," in *Proc. 6th International Conference on Spoken Language Processing*, Oct. 2000, pp. 349–352.

[18] H. Fujisaki and H. Sudo, "A model for the generation of fundamental frequency contours of japanese word accent," *Journal of acoustical society of Japan*, vol. 27, no. 9, pp. 445–453, 1971.

[19] K. Maekawa, "Production and perception of 'paralinguistic' information," in *Proc. Speech Prosody*, Mar. 2004, pp. 367–374.

[20] T. Shochi, V. Aubergé, and A. Rilliard, "How prosodic attitudes can be false friends: Japanese vs. French social affects," in *Proc. Speech Prosody*, May 2006, pp. 692–696.

On the Role of Pitch in Perception of Emotional Speech

Noam Amir[1], *Eitan Globerson*[2]

[1] Department of Communication Disorders, Tel Aviv University, Israel
[2] Academy of Music and Dance, Israel

noama@post.tau.ac.il, gleitan@zahav.net.il

Abstract

Two experiments investigated the role of intonation in perception of basic emotions. In the first experiment, pitch contours of stimuli from a corpus containing portrayals of anger, joy, fear and sadness were manipulated with respect to range, mean and smoothness. In the second experiment, pitch contours of identical words portraying different emotions were exchanged. In each experiment, the emotional category and intensity of the original and manipulated stimuli were evaluated by two separate groups of 20 participants. Results of the first experiment show mainly that pitch mean and range should vary congruently to portray activation correctly, and demonstrate the interaction in varying these two parameters. Results of the second experiment show that a pitch contour conveying high activation is not sufficient in conveying the appropriate emotion, if the other paralinguistic cues are not also in accordance. A pitch contour indicating low activation, on the other hand, is apparently a more powerful cue and thus less reliant on other cues.

Index Terms: emotion, speech, intonation, perception

1. Introduction

It has long been acknowledged that pitch has a central role in production and perception of emotional speech [1], with more recent studies indicating that pitch mean and range are strong indicators of Activation or Arousal [2]. Many studies have shown, however, that other prosodic parameters such as intensity, timing and voice quality, are also involved in this process [3, 4,5]. It is therefore interesting to try and isolate the contribution of each of these factors to perception of emotion. Several previous studies have explored this issue using resynthesis, manipulating various prosodic or spectral properties and examining the effect on recognition scores [5, 6,7]. This is of interest for producing emotional speech in text-to-speech systems, however such tools can also be used studying the acoustic cues themselves.

In this study we present two experiments which attempt to investigate the contribution of pitch to emotion perception in a controlled manner. However we chose to perform manipulations on recorded emotional speech rather than text-to-speech synthesis. Both experiments employ a corpus of emotional speech which has been analyzed extensively in several other studies [8]. It is composed of short sentences and single-word (nonsensical and meaningful) emotional utterances with neutral content, recorded from 4 speakers and conveying 4 emotions: Anger, Joy, Fear and Sadness.

In experiment 1, the pitch contours of the different utterances were manipulated to exaggerate or diminish the pitch movements, in order to determine how this influenced the detected emotion and its intensity. In the second experiment, pairs of pitch contours from identical words uttered in different emotional expressions were exchanged, in order to examine which cues would be dominant: the pitch contour or the remaining acoustic cues.

The next section describes in some detail the corpus employed here, followed by detailed descriptions of each experiment, and a conclusion.

2. Emotional Corpus

A full description of the emotional corpus, detailing how it was recorded and evaluated can be found in previous work by the present authors [7]. In brief: stimuli were recorded in a professional recording studio by four professional actors (two female, and two male). The stimuli included nonsense monosyllabic utterances, nonsense polysyllabic words, Hebrew words and Hebrew sentences. None of the words and sentences had any linguistic emotional content. The stimuli represented four basic emotions: anger, joy, fear and sadness. The stimuli were validated by a panel of 20 independent judges, receiving an overall average recognition rate of 79.0% (SD=15.9%).

3. First Experiment

3.1. Objectives

The objectives of this experiment were to determine the effect of several properties of the pitch contour on the perception of emotion: 1) *Small pitch movements*: normally produced pitch contours contain major pitch movements known to signify both pragmatic and affective information, such as a final rise indicating a question. However, there are many smaller pitch movements which may have a significant but less obvious role [9]. Removing the small pitch movements, while retaining the large ones, may enable a controlled investigation of their perceptual importance. 2) *Pitch range*: this parameter has long been considered an important property of the pitch contour, usually considered to be an indicator of Activation or Arousal [2]. By stretching or compressing the pitch contour, the pitch range can be shifted considerably, without damaging its overall shape. 3) *Pitch mean*: This is also considered an important indicator of emotion, though it is not clear whether it is a global feature, independent to some degree of the speaker's normal mean, or must be normalized to each speakers individual mean. Once again, shifting the pitch mean can help determine its perceptual value.

3.2. Methods

In order to keep the listening task manageable, 92 utterances were selected from the original emotional speech corpus. The selection was balanced to ensure nearly the same number of utterance per emotional category (23 or 24 per category) and per utterance type (21-24). All the selected stimuli were originally identified correctly at a rate of over 68%. Each of these utterances was then manipulated in four ways:

1. Smoothing: removing any small pitch movements. This was carried out using Praat software's "stylize" manipulation with a parameter of 2 semitones.
2. Increase of pitch range by a factor of 2.
3. Decrease of pitch range by a factor of 2
4. Shifting of the pitch mean: Globally, pitch means for anger and sadness were low and means for joy and fear were high, both for men and women. Their overall averages were 260Hz for women's utterances, 167Hz for men's utterances. Thus, all stimuli were shifted to these mean values, for women and men respectively.

All manipulations were performed in Praat software. The net result was a collection of 460 stimuli: the original 92, and four manipulations of each.

The experiment was implemented as a Matlab GUI. For each stimulus, participants had to choose one of five emotions: anger, sadness, joy, fear or neutral, and rate the emotional intensity on a scale of 1 to 3. Despite the fact that the corpus did not contain neutral stimuli, the neutral category was included to account for cases in which the emotional content was possibly lost due to the above manipulations.

The participants were a group of 20 women, aged 20 to 30 (M=26). All participants had no reported hearing problems, 12-16 years of education, and were native Hebrew speakers.

3.3. Results

In the interest of brevity, only the main results of this experiment are presented here.

3.3.1. Identification of emotions

Table 1 shows the average recognition score per stimuli, for each emotion and each type of manipulation, in percents. Figure 1 shows the same results graphically. Evidently, baseline recognition scores (i.e. for non-manipulated stimuli) were very similar, over 80%, for all the emotion types.

Observing each row, it becomes obvious that the different types of manipulation had different effects on recognition of the various emotions.

The main findings from Table 1 are as follows: **Anger** recognition was reduced significantly ($p<0.05$) by shifting the pitch (in this case to higher levels) and *reduction* of the pitch range. **Joy** recognition was reduced significantly by pitch shifting (to lower levels in this case) and *reduction* in pitch range. **Sadness** recognition was reduced significantly by *increasing* the pitch range. **Fear** recognition was lowered greatly by pitch shifting (to lower levels).

*Table 1. Average **recognition** scores (%) per emotion and manipulation type. Asterisks mark statistically significant differences.*

Emotion	Original	Stylized	Shifted	0.5 x range	2 x range
Anger	84	86	73*	74*	80
Joy	81	74*	69*	53*	84
Fear	81	77	51*	76	71
Sadness	84	86	84	87	72*

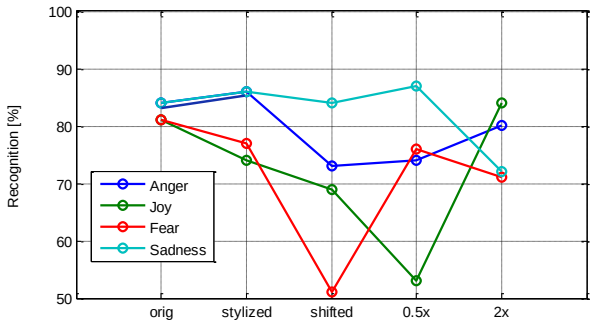


Figure 1: *Recognition scores of Table 1*

3.3.2. Ratings of emotional intensity

Table 2 shows mean intensity scores per emotion, for each manipulation. Anger and Joy scores fell when pitch range was reduced. Joy scores increased when pitch range was increased, whereas this had the opposite effect on Sadness. Finally, pitch shifting reduced intensity scores for Joy and Fear. This is in line with the fact that it reduced their recognition scores also.

*Table 2. Average emotional **intensity** scores per emotion and manipulation type (normalized to a 10-point scale). Asterisks mark statistically significant differences.*

Emotion	Original	Stylized	Shifted	0.5 x range	2 x range
Anger	6.1	6.3	6.1	5.4*	6.
Joy	5.7	5.6	5.2*	4.8*	6.6*
Fear	6.9	7	5.3*	6.9	6.5
Sadness	6.5	6.3	6.4	6.4	5.7*

3.4. Discussion

The results above show several interesting trends. In the original stimuli, the two emotions with high activation, Anger and Joy, demonstrated opposing tendencies with regard to average pitch. Thus, shifting them to the overall average (reducing pitch in Joy and increasing it in Anger) caused recognition rates to fall for both. As activation is commonly acknowledged to be associated with high pitch range [Banziger], recognition rates of both these emotions also fell when pitch range was reduced.

Sadness and Fear also employed opposing tendencies in the original corpus, with regard to average pitch: Fear being high and Sadness low. However, shifting the average pitch had a large effect on Fear, but no effect on sadness. Nevertheless, since these emotions have relatively low activation, *increasing* the pitch range for these emotions resulted in lower recognition rates.

Pitch stylization had some interesting effects also. It affected Joy most significantly, reducing recognition rates. This may indicate that recognition of Joy relies to some extent on small pitch movements as well as large ones. On the other hand, stylization increased recognition of Anger and Sadness slightly, though not significantly. Possibly this type of manipulation removed some small pitch movements that might in fact cause the emotion to sound more ambiguous in these cases.

The trends for intensity scores are roughly similar to those found for recognition scores. Emotions with high activation (Anger and Joy) were adversely affected by *reducing* pitch range, though less affected by *increase* in range. Pitch shifting had a similar effect on intensity as on recognition, though not in all emotions.

Overall, this experiment reveals that pitch mean and pitch range behave somewhat independently. Range appears to signify activation. Thus, reducing the range for emotions with high activation (Anger and Joy) appears to confound the listeners. However, reducing the pitch range for emotions with low activation (Sadness and Fear) does not have an identical effect, only marginally changing the recognition of both.

Increasing the range had no effect on emotions that had a large pitch range to start with, but reduced the recognition rates for emotions with low activation (Fear and Sadness) where a large pitch range was not expected.

Shifting the pitch mean adversely affected nearly all emotions, probably because shifting was done towards the overall average, thus reducing the degree of emotional content. This raises the question whether listeners were able to normalize their expectations of average pitch to the speakers, or whether speakers are preconditioned towards some overall global average in this respect. Interestingly, Sadness was not at all affected by this manipulation.

To summarize, the results show that the parameters of pitch range and mean and their interactions have important significance in production and perception of emotions, which are highlighted in this experiment.

4. Second Experiment

4.1. Objective

The objective of this experiment was to examine the degree to which the specific pitch contours associated with each emotional production were responsible for the correct perception of the emotion.

4.2. Methods

The original corpus was scanned for utterances which could be "paired": i.e. words or nonsense words uttered by the same speaker while expressing two or three different emotions. 104 such utterances were found. For these utterances, the pitch contour of one emotion could be synthesized onto the utterance the other one or two remaining utterances. The final corpus was thus composed of:

1. 104 original utterances
2. For pairs of utterances uttered in *two* different emotions, the pitch contour of one utterance was synthesized onto the other utterance, and vice versa, producing two new stimuli.
3. For triplets of emotions uttered in *three* different emotions, each combination of cross synthesizing pitch contour and utterance were created, producing six new stimuli

The net result was a corpus of 268 stimuli: 138 nonsense words (54 original and 84 synthesized), and 130 Hebrew words (50 original and 80 synthesized).

All manipulations were performed in Praat software. This experiment was also implemented as a Matlab GUI. For each stimulus, the participant had to choose one of five emotions: anger, sadness, joy, fear or neutral, and rate the emotional intensity on a scale of 1 to 3.

The participants were a group of 20 women, aged 20 to 30 (M=26.4), separate from the group that had participated in the first experiment. All participants had no reported hearing problems, 15-19 years of education, and were native Hebrew speakers.

4.3. Results

The average recognition score for all non-manipulated stimuli was 15.2 out of the possible 20, which is slightly lower than in experiment 1. The entire confusion matrix for these stimuli appears in Table 3. All values on the diagonal are in the vicinity of 15, with a relatively even distribution of errors. This indicates that on average, the emotions in the original, non-manipulated stimuli were fairly easy to recognize.

Table 3. *Confusion matrix of recognition scores for the original stimuli (%)*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	75	7	2	1	14
Joy	12	77	2	1	8
Fear	1	1	80	18	1
Sadness	7	0	10	72	11

For the manipulated stimuli, two confusion matrices can be calculated: one for identification of the stimuli according to the original emotion of each stimulus, and the second according to identification of the emotion represented by the synthesized pitch contour of each stimulus. These two matrices are presented in tables 4 and 5.

Table 4. *Confusion matrix of recognition scores for the manipulated stimuli, identified in accordance with the original emotions (%).*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	36	15	21	17	11
Joy	13	36	14	15	22
Fear	5	4	47	34	11
Sadness	7	5	28	46	13

Table 5. *Confusion matrix of recognition scores for the manipulated stimuli, identified in accordance with the synthesized pitch contour (%).*

Perceived / Produced	Anger	Joy	Fear	Sadness	Neutral
Anger	10	19	25	24	22
Joy	23	18	25	21	13
Fear	12	14	39	29	6
Sadness	18	11	20	36	16

Table 4 indicates that on average, changing the pitch contour caused recognition rates to fall, though they remained above the chance level (4/20) for all the emotions. Emotions Fear and Sadness remained more immune to changes of the pitch contour.

The first two values on the diagonal of Table 5 are very low. This indicates that imposing a pitch contour of Anger or Joy on utterances that originally conveyed other emotions, had very little success in causing these utterances to "shift" towards Anger or Joy. On the other hand, the second two values on the diagonal of this table are much higher, well above chance. This indicates that imposing the pitch contour of Fear or Sadness had a clear effect in shifting the perceived emotions towards these two.

Finally a more detailed analysis was carried out. Each combination of original and F0 emotion was analyzed separately, scoring perception according to 1) the original utterance; 2) the pitch contour. Results are presented in Table 6 and Figure 2.

Table 6. *Recognition scores for each cross-manipulation of original emotion and F0 contour emotion (%). Four blocks of rows are denoted with shading: hi activation to hi, hi to low, low to hi and low to low*

Row #	Original emotion	F0 emotion	Score according to original	Score according to F0	N
1	Anger	Joy	48	26	16
2	Joy	Anger	49	14	16
3	Anger	Fear	23	40	17
4	Anger	Sadness	38	28	11
5	Joy	Fear	35	32	11
6	Joy	Sadness	23	28	14
7	Fear	Anger	49	5	17
8	Fear	Joy	52	10	11
9	Sadness	Anger	50	12	11
10	Sadness	Joy	37	13	14
11	Fear	Sadness	40	52	13
12	Sadness	Fear	52	43	13

Some interesting observations can be made based on this table. For example, rows 1 and 2 show that emotions with high activation (Anger and Joy) are not easily "taken over" by swapping each other's pitch contours. However, rows 11 and 12 show that the opposite is true for emotions with low activation. Rows 2 - 6 indicate that Anger and Joy are more readily taken over by pitch contours of Fear and Sadness. Rows 7 – 10 show that Fear and Sadness are *not* easily taken over by Anger and Joy.

4.4. Discussion

The observations regarding the confusion matrices strengthen the notion that pitch is an important cue of emotion, though not the sole one. Imposing a "wrong" pitch contour reduced recognition score of the original emotion by approximately half, though emotions with high activation (Anger and Joy) appear to be more severely affected than emotions with low activation (Fear and Sadness).

Conversely, when looking at the same manipulated stimuli in order to observe when the imposed pitch contour "took over" the perceived emotion, a gross asymmetry can be observed. On average, the pitch contours of the highly active emotions had no discernible effect at all in taking over the perceived emotion. On the other hand, pitch contours of the emotions with low activation were moderately successful in taking over the perceived emotion.

Our conjecture is therefore that a pitch contour conveying high activation is far from being sufficient in conveying the appropriate emotion, if the other paralinguistic cues are not congruent. A pitch contour indicating low activation, on the other hand, is apparently a more powerful cue and thus less reliant on other cues.

The detailed analysis in Table 6 also raises some interesting asymmetries, showing that pitch contours can more easily draw one emotion to the other than the reverse. Line 3 shows that Anger is rather easily turned into fear, for example, but line 7 shows that the opposite is not true. Joy is easily converted to Sadness (line 6), but again the opposite is not true (line 10).

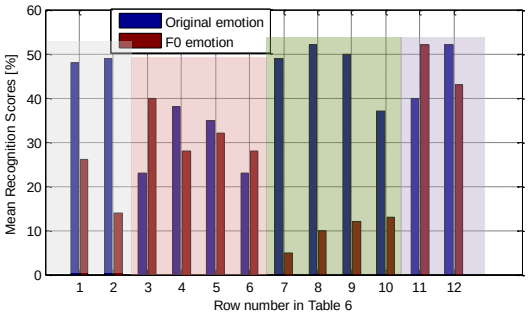


Figure 2: *Recognition scores of Table 6.* Groups of bars are color coded to fit the rows in the table.

5. Conclusions

Arriving at a clear and definite picture of the cues behind the production and perception of emotions is a daunting and likely impossible task. However, studies such as the present one can provide insight into the relative perceptual importance of these cues. Pitch has been long been acknowledged to be a central cue in emotion recognition, however this study quantifies its role to a certain extent, and underlines the fact that it has different influence on the perception of different emotions. The corpus used here offers a unique opportunity to study such effects in isolation, an opportunity which is not afforded by many emotional corpora. Presumably, manipulations of further cues, such as voice quality, could lead to results with even a higher degree of refinement in separating the effects of different cues to emotional perception.

6. Acknowledgements

The authors would like to thank Nina Gilad, Amit Lavi, Michal Rubinstein and Ravit Tahar for their assistance in running the experiments.

7. References

[1] Murray, I.R., Arnott, J.L., "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", Journal of the Acoustic Society of America. 93(2), 1097- 1108, 1993

[2] Banziger, T., Scherer, K.R., "The role of intonation in emotional expressions", Speech Communication, 46(3), 252-267, 2005

[3] Banse, R., Scherer, K.R., "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, 70(3), 614-63, 1996.

[4] Yanushevskaya, I., Gobl, C., & Ní Chasaide, A., "Voice quality and loudness in affect perception", Proceedings of Speech Prosody 2008, Campinas, Brazil

[5] Gobl, C., Ni Chasaide, A., "The role of voice quality in communicating emotion, mood and attitude", Speech Communication 40(1), 189-212, 2003.

[6] Lindh, J., "A model based experiment towards an emotional synthesis", Proceedings of FONETIK 2005, Goteborg.

[7] Burkhardt, F., "Emofilt: the simulation of emotional speech by prosody-transformation", Proceedings of Interspeech 2005, Lisbon

[8] Globerson, E., Amir, N., Golan, O., Kishon-Rabin, L., Lavidor, M., "Psychoacoustic abilities as predictors of vocal emotion recognition", Attention, Perception and Psycophysics, 75(8), 1799-1810, 2013

[9] Lieberman, P., & Michaels, S. B., "Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech", Journal of the Acoustical Society of America 34, 922–927, 1962

Politeness, culture, and speaking task – paralinguistic prosodic behavior of speakers from Austria and Germany

Sven Grawunder[1], *Marianne Oertel*[2], *Cordula Schwarze*[3]

[1]Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[2]Department of Speech Science, Martin Luther University Halle-Wittenberg, Halle, Germany
[3]Department of German Studies, University of Innsbruck, Austria
grawunder@eva.mpg.de

Abstract

This paper tests previous findings for polite speech of low pitch, low intensity, higher number of hesitation markers and filled pauses against those parameters in a different socio-cultural background. Two similar groups of (19+13) participants, from Austria and from Germany, were recorded. The adopted experimental approach used 16 tasks aiming at different speech acts in situations that evoke either polite or informal speech. The analyzed acoustic and electroglottographic signals reveal main effects for lower pitch, lower intensity and HNR only for the German group. Open quotient values differ only for female speakers. In both groups significantly lower word rate and lower speaking rate as well as higher rates of filled pauses and hesitation markers are found in formal (polite) conditions. However individual speakers can show indifferent or opposing behavior for a given parameter with compensatory utilisation of other parameters in order to express politeness (formality).
Index Terms: polite speech, voice quality, pauses, hesitations

1. Introduction

While looking at expressions on the sociopragmatic level of speech [1] a number of phonetic and other paralinguistic differences between polite vs. informal speech registers have been found for situations talking either to a senior person or to a peer. We adapt the view that the concept of polite speech covers here typically a transmission of a social message of deference, submission and distance, where politeness (polite behavior) is considered to be a vehicle to ensure stability of social hierarchies. In the model of Brown & Levinson [2] this is categorized as negative politeness strategy. Such speaking style is different from the speech of, e.g., shop assistants or other service personel, that seeks an exaggeration of displayed interest, approval, or sympathy with the interlocutor, which is identified as positive politeness [2]. And in fact, Ito [3] and Ofuka [4] found for Japanese female speakers that high pitch register and higher degrees of breathiness correlate with higher degrees of perceived politeness. Higher pitch range in connection with perceived politeness are reported for British English and Dutch by Chen et al. [5] but also for a number of Spanish Varieties in Latin America (cf.[6]). However, for Korean females a pitch lowering was found [7] in polite speech, which was confirmed in a independent production study and also found for male speakers [8]. In a recent study it was shown that English and Korean listeners can detect the intended politeness purely based on prosodic expressions (in Korean), even if only presented with minimal material like a sentence [9]. Accuracy would increase for native and non-native listeners as they become familiarized with the voices. Similarly a reduction of pitch range in Catalan [6] has been found to prompt a perception of politeness. These findings do not meet the assumptions of high pitch as known from the predictions by Brown & Levinson [2] or the frequency code hyphothesis of Ohala [10, 11], but suggest a more differentiated hypothesis with regard to the phonetic profile of politeness. Additionally it brings into play other biologically motivated sound symbolisms in human and non-human vocal behavior, namely the effort code hypothesis by Gussenhoven [12], since here other voice quality parameters apart from pitch are used. Hence this paper seeks to contribute more rigorously to the question of cross-linguistic validity of such findings and seeks to replicate the previous study [8] with a different cultural and linguistic background. In order to address also the cultural pragmatic aspect, we conducted the same experiment in two linguistic expressions of cultures (Austria and Germany) [13, 14] of the same language (German). If a more universal usage of a "frequency code" holds we would – given the results for Korean [8, 9, 15] or Catalan [6] – expect similarily for polite speech: lower F0, lower F0 range, lower intensity, lower intensity range, more tensed VQ with parameters indicating this (i.e. acoustically, lower harmonics-to-noise ratio (HNR), or electroglottographically, lower open quotient). Further we would expect for polite speech more hesitation phenomena (onset hesitation, lengthening, hesitation 'markers') and more signals of interruption (audible breath intakes etc., cf.[8]).

2. Data and Methods

2.1. Participants and Procedure

The current data set comprises recordings of 13 speakers (11 female/2 male) from East Central German and of 18 speakers (8 female/10 male) from Western Austria. Recording sessions were carried out in the phonetics lab of the MPI EVA Leipzig and in the sound attentuated room of the Media Center of University of Innsbruck. For best replication of the previous study, we were aiming at students of the humanities as participants which were not freshmen any more but still in the age of 20 to 30. The latter ensured that our participants already had commuciation experience in hierarchies and asymmetrical relations, but were not too advanced in their carrier. To ensure consistency of procedure, participants were first ask to read aloud a written text in order to get acquainted with the setup. After clarifying the general procedure the participants were prompted with different situations via a screen. The 16 situations described demanded either a mailbox task, where a participant would be asked to leave a message for her friend or her boss on a voice

mail box. Or the task was a discourse completion task (DCT), where a participant would be asked to start a conversation, like *'Remind your professor to upload the power point files after the lecture'*. Linguistically the scenario descriptions were culturally adapted for the local areas in Austria and Germany so that infrequently used phrases and non-typical names were avoided.

2.2. Recordings and Measures

The recordings of the acoustic (headset microphones AKG C420 III and audix HT5) and electroglottographic signals (Glottal enterprise EG2) were accomplished by means of a multitrack recorder sound device 788T in a PCM format of 44.1kHz 16bit. *Praat* [16] was used as annotation tool and for acoustic and EGG measures. Manual annotation and labeling was carried out in order to assess word rate, occurence of hesitation markers as well as audible and silent breath intakes. Syllable estimation [17], f0, f0 variance, f0 range, intensity (RMS), intensity range, HNR as well as noise-to-harmonics ratio (NHR), as well as jitter (local, RAP) and shimmer (local, APQ3) [16, Manual] measures were based on the periodic parts of the microphone signal. The amplitude difference H1-H2 had not been chosen, since especially for analyzing running speech these have been lately critized to be heavily vowel dependent [18]. The EGG-signal served for the open quotient, i.e. duration of open phase by total period. These quotients were based in part on the DEGG-method where the first derivative is used for estimation of the instants of closing and opening (cf. [19]).

2.3. Notes on Statistics

We used *R* [20] with the packages *lme4* [21] and *glmmADB* [22] to perform a linear mixed effects analysis of the relationship between the individual parameters and the supposed attitude (polite vs. informal). In a first step we set attitude and gender as fixed effects into the model, in a second step we added group (Austrian, German) to the fixed effects. As random effects, we had random intercepts for speakers and tasks, as well as random slopes for the effect of attitude by speaker and by task. Visual inspection of residual plots served as check for deviations from homoscedasticity or normality, including overdispersion by particular random effects (task, speaker). P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. Likewise models with interaction (e.g. attitude*gender) were tested against such models without interaction (e.g. attitude+gender). For the count data (pauses etc.) we were using with negative binomial and zero.inflation=TRUE.

3. Results

3.1. Polite vs. Informal: Overall

3.1.1. General Observations

All participants reacted eagerly to solve the experimental tasks quickly and freely. Specifically the mail box task proved to be appropriate, especially in initiating the experiment and gaining near real-life utterances for most of the time. The formal vs. informal distinction was realized by all participants. It was not only that participants would choose the correct grammatical forms (e.g. du 2SG 'you; thou' vs. Sie 'you' 3SG) for addressing the imaginary interlocutor, but they would also use more passive or subjunctive constructions. With regard to pronunciation we almost always observed a shift away from close-to-standard patterns into to more areal or colloquial patterns in informal situations. This was more clearly obvious for the Austrian group, since the (spoken) standard for all German speaking countries is derived from more Northern German pronunciation habits, hence there is a greater phonetic distance between the two varieties and therefore the two situational styles/registers.

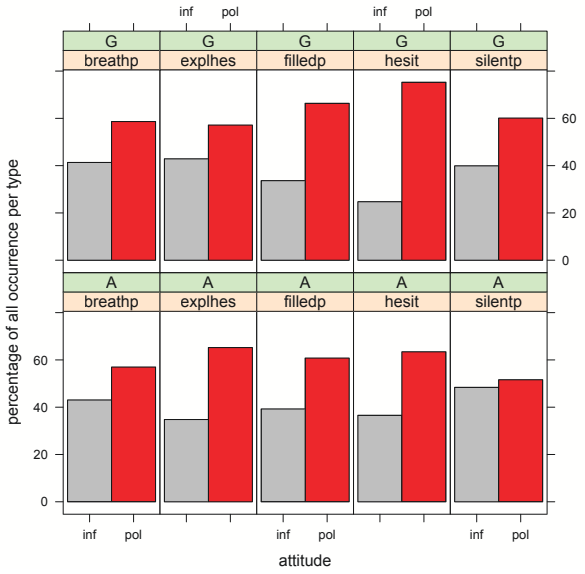


Figure 1: *occurrence of breath, silent and filled pauses as well as explicit and other (lengthening) hesitation markings, grouped by country (*Austria; Germany*)*

3.1.2. Hesitation Phenomena and Rates

One of the most striking phenomena is the occurence of hesitation phenomena throughout the corpus. In order to group the different appearances of these phenomena in a meaningful way we defined five types. First, we separated pauses into *silent pauses* and (audible) *breath pauses* (in and out). Then we would treate conventionalized hesitation markers – also known as non-lexical conversational sounds (cf. [23]) or fillers – like "ä","äh","ähm" ('hm, mh , em' etc.) as *filled pauses*. In contrary such coventionalized explicit phrases that contain lexical items like "ich denk", "glaub ich" ('I guess, I think') or conjunctions like "u:::nd", "dann:: ", "weil::" ('and, well, because' etc.) – functioning as discourse markers – were grouped as *explicit hesitations*. Although one can observe all kinds of combinations of the conventionalized markers, phrases and conjuctions plus a phonetic lengthening (like "u:::nd", "dann:: ", "weil::" or "und ähm", "weil ähm", "dass ähm") we nonetheless treated these lengthening behavior as one *hesitation* category. Silent pauses would also comprise instances of breath holding, i.e. a pause with oral or glottal closure, with a more or less sudden release. These cases were also treated under the same category, *hesitation*.

Overall we find higher counts and percentages for all types of hesitation marking and pauses in the formal (polite) condition. Breathing pauses (inhales) show a significant main effect (attitude=polite, Est. +0.363 ±2.99 SE, p=0.0028) with no (gender) interactions. However one needs also to take into account that the answers in the polite condition have a longer duration. The relation of effective duration and total duration of a task reflects the role of pauses per task similar to the rela-

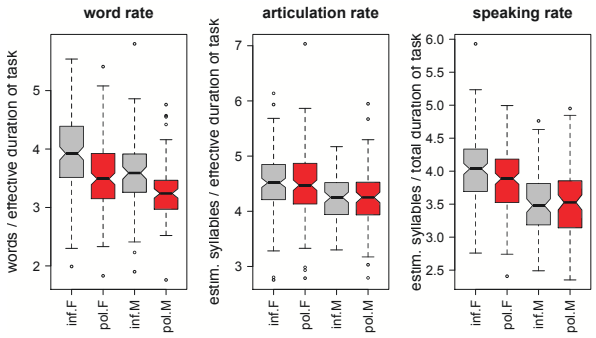


Figure 2: *Overall rates for all female (F) and male (M) speakers in polite (pol) and informal (inf) condition*

tion of speaking rate vs. articulation rate (-0.025s ± 0.01154, p=0.037). Closely related are the results of word rates (p=0.03), since here a higher number of pauses corroborates the lower rate of words (per task duration). Moreover the lower 'wordiness' may even be more pronounced since here all non-lexical markers have not been excluded yet.

3.1.3. Pitch, Intensity and Voice Quality Measures

Average f0 values display only a minimal overall main effect of -0.41st and ±0.217, i.e. lower f0 in the polite condition. Only half which comes out significant for interaction (attitude*gender) for the German subgroup (-0.44st ±0.273, p=0.012). Further only mean intensity differences reach significance in interaction (-1.24dB ±0.460, p=0.02746). The only acoustic voice quality parameter showing significant effects reflects the variance of HNR within a task: overall HNR SD is lower for polite speech with interaction effect (0.28dB ±0.116, p=0.019), in the Austrian (0.17dB ±0.109, p=0.023) and the German group (here only as main effect, 0.14dB ±0.098, p=0.015). This points towards a more non-homogenous use of voice quality on the breathiness-tensedness dimension. Taking into account correction for multiple testing we note that this result is corroborated by the results for open quotients (OQ; Fig. 3), where we see lower values in polite situations (mean OQ: 0.013 ±0.0114, p=0.0364; OQ SD: 0.0105 ±0.0041, p=0.043; here only reported for interaction with country since only female speakers were compared). These findings suggests a higher tensedness (or lower breathiness) for polite speech and thus far parallel the results for Korean [8].

3.2. Group specific behavior

3.2.1. Overall gender specific behavior

Gender specific main effects are found for parameters that are expected and well known to differ between genders (sex). These concern mainly pitch and voice quality (jitter, shimmer), where f0 range seems to be strikingly higher for male speakers in both conditions. As often observed, we find higher articulation rates for female speakers (Fig. 2; with interaction, 0.31words/s ±0.160, p=0.00014). More interesting in this regard is that word rates show a signficant effect with interaction for lower rates in polite speech with male speakers (0.013words/s ±0.080, p=0.039).

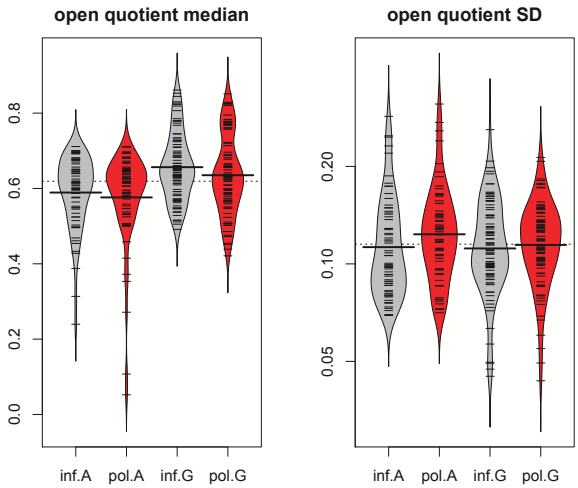


Figure 3: *open quotient measures for female participants from Austria (A; N=8) and Germany (G; N=11)*

3.2.2. German in Austria vs. German in Germany

Apart from a general trend for open quotients, which appear to be very small but significantly higher for female speakers from Germany (mean OQ:-0.0024 ±0.01806, p=0.039) than for female speakers from Austria, meaning that there comes into play a difference between these groups, perhaps pointing towards a general use of breathiness in the speaker population from Germany. Other measures where we find group (country) effects but also main effects for attitude are mean intensity values (-6.46dB ±1.542, p=0.00011) and HNR SD values (0.79dB ±0.3106, p=0.0053). With regard to the nature of these measures such differences may be easily explained by the influence of conditions (e.g. mic or booth) in the two recording sites.

3.3. Task specific behavior

The individual task categories (eight scenarios) that we were testing for phonetic aspects comprise a number of different speech acts, i.e. types of messages that have to be conveyed by the paticipant. These required urgently requesting an action (task 11, 12), requesting something critical (task 2,6), requesting something essential (task 1), apologizing for being late, making a compliment (task 5), and correcting a (minor/major) mistake (task 3,4). Indeed we observe, more often especially for task 5 (*compliment*) but also for task 6 (*minor request*), behavior that deviates from that of other tasks. For mean HNR measures in task 5, showing the lowest values in the dataset, a three way interaction of attitude, gender and task becomes marginally significant (-1.42dB ±0.737, p=0.0545). Moreover, mean HNR measures show for task 2 and 12, lower values for males and females whereas the other tasks display the contrary trend, e.g. males in task 3 (1.48dB ±0.521, p=0.0045). Strikingly the two groups behave almost symmetrically (s. Fig. 4) and we are not yet able to discard a particular task as behaving consistently different.

3.4. Speaker specific behavior

While looking at strategies of individual speakers we report in Table 1 individual slopes for a selection of parameters that show partially an opposing behavior (cf. [24]). For example

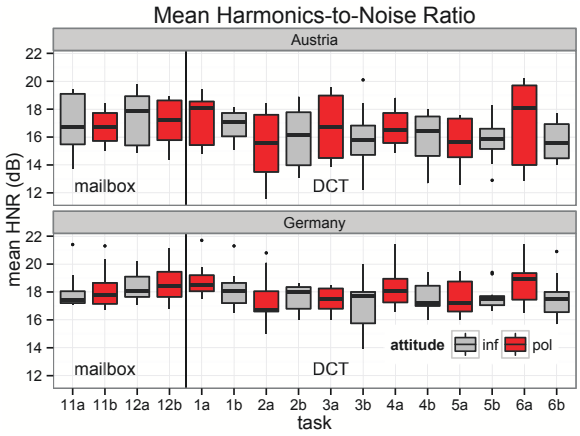


Figure 4: *taskwise HNR behavior (females only)*

speaker GM01 illustrates the complementary use of high intensity range difference but countertrend f0 range. At the lower end we find speaker AM02 with the opposite behavior. Simlarily HNR seems to be more strongly employed (e.g. by GM01). By contrast, all speakers exhibit the same trend for the pause quotient (effective task duration by total task duration).

4. Discussion

On the one hand the current results for speakers of two German varieties show tendencies for the use of hesitations and pauses as well as for the use of pitch, pitch range, intensity, intensity range, and HNR, that match those results for Korean in [8]. On the other hand, the two speech registers appear to be less distinct, given the lower number of significant differences, which can be assumed to mirror the fact that for German politeness or formality register is linguistically but also culturally less entrenched. Certainly we must stay aware of the reality of the experimental nature of the investigated speaking behavior. The general assumption here is that participants will simulate from the same source of acquired culturally transmitted and evolutionary conditioned behavioral pattern as they would in a non-experimental situation of a similar situation 'in situ'. The latter would then supposedly evoke higher arousal levels and lead to stronger difference between the attitudes (formal/informal). We cannot exclude at this moment that individual speakers may also fail to express deference and submission widely in favour of other expressions, such as friendliness or fear, while still being formal. This could in case of friendliness be interpreted as pursuing a positive politeness strategy, or in case of fear, as an exaggeration of the urgency message. Both may remain disrespectful and therefore potentially impolite if not addresssed to peers or if failing formality even with addressing a peer – a distinction that remains to be tested yet.

5. Conclusions

Since languages vary in their entrenchment of pragmatically relevant registers (such as different degrees of politeness) into the grammar, moreover the paralinguistic expression may vary between speech communities. Small scale samples, as ours, may not give enough power to estimate the actual group differences that interact with gender specific effects. Within communities we observe speakers using the multidimensional space of possible parameters (pitch, loudness, voice quality, fluency; cf.[25]) in its own particular way but also in the degree of expression. Already previous results for Korean [8] contradict the findings for Japanese and English [3, 4, 5, 7] where politeness was associated with higher pitch (and more breathiness). Although with our German speaking participants we find more non-homogenous behavior of subjects, our findings are still very similar, since we find clear trends for (less breathy) voice quality and higher rates of pausing/hesitations in polite speech. But we can also confirm the previous observations of Shin [7], who found no pitch differences for German. Given that we were using here still a parsimonious approach, using 'holistic' rather than detailed measures – an in-depth analysis of the actual moments of expression for pitch [6], intensity or breathiness/tensedness will certainly gain more insigths. In order to achieve ecological validity we will also turn these production data into a perception experiment.

Table 1: *Random slopes per speaker from polite condition sorted by f0 range values; pausequot (effective task duration / total task duration)*

speaker	f0 range	Int range	pause quotient	mean HNR
GM01	-0.08	2.65	-0.024	0.65
AM03	-0.01	1.18	-0.026	0.33
AM06	0.02	0.69	-0.021	0.36
AF09	0.76	1.92	-0.025	0.48
GF11	1.12	0.17	-0.026	0.49
AM08	1.18	0.40	-0.027	0.04
AF06	1.30	1.34	-0.023	0.19
AM07	1.37	0.81	-0.026	0.40
AF08	1.51	1.98	-0.023	0.49
GF05	1.62	2.08	-0.026	0.53
AM05	1.89	1.05	-0.030	0.31
GF03	1.91	1.89	-0.026	0.58
GF08	1.93	1.08	-0.027	0.30
GF04	2.04	2.25	-0.024	0.23
AF04	2.04	0.26	-0.027	-0.01
AF01	2.04	1.08	-0.027	-0.09
AF03	2.05	-0.31	-0.026	0.30
GF09	2.14	0.28	-0.023	0.27
AF02	2.17	1.14	-0.023	0.52
GF10	2.31	0.13	-0.027	0.22
AF07	2.55	0.68	-0.025	0.08
GF06	2.55	1.06	-0.023	0.53
AM10	2.67	1.34	-0.019	0.30
GF07	2.69	1.44	-0.027	0.38
GF02	3.03	1.29	-0.025	0.67
AM02	3.21	-0.08	-0.027	0.17
GF01	3.23	0.27	-0.026	0.33
GM02	3.49	1.14	-0.026	0.58
AM01	3.60	1.69	-0.027	0.34
AM09	3.61	0.42	-0.023	0.38
AM04	3.67	1.36	-0.025	0.24

6. Acknowledgements

We would like to thank all our participants and express our gratitude to Leonardo Lancia (MPI EVA), to Anton Tremetzberger (Uni Innsbruck) for assistance during the recording sessions, to Yvonne Kathrein for helping with the 'Austrification' of our stimuli and to Bodo Winter (UC Merced) for his technical advice. Finally we would like to thank the MPG, specifically Prof. Bernard Comrie for his support.

7. References

[1] M. Tatham and K. Morton, *Expression in speech: Analysis and synthesis*. Oxford: Oxford University Press, 2004.

[2] P. Brown and S. Levinson, *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press, 1987.

[3] M. Ito, "Politeness and voice quality–the alternative method to measure aspiration noise," in *Proceedings of the Second International Conference on Speech Prosody, Nara, Japan*, 2004, pp. 213–216.

[4] E. Ofuka, J. D. McKeown, M. G. Waterman, and P. J. Roach, "Prosodic cues for rated politeness in Japanese speech," *Speech Communication*, vol. 32, no. 3, pp. 199–217, 2000.

[5] A. Chen, C. Gussenhoven, and T. Rietveld, "Language-specificity in the perception of paralinguistic intonational meaning," *Language and Speech*, vol. 47, no. 4, pp. 311–349, 2004.

[6] M. Nadeu and P. Prieto, "Pitch range, gestural information, and perceived politeness in Catalan," *Journal of Pragmatics*, vol. 43, no. 3, pp. 841–854, 2011.

[7] S. Shin, "Grammaticalization of politeness: A contrastive study of german, english and korean," PhD Thesis, University of California, Berkeley, 2005.

[8] B. Winter and S. Grawunder, "The phonetic profile of Korean formal and informal speech registers," *Journal of Phonetics*, vol. 40, no. 6, pp. 808–815, 2012.

[9] B. Winter, L. Brown, K. Idemaru, and S. Grawunder, "Perceiving politeness from speech acoustics alone: A cross-linguistic study on Korean and English," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 4072–4072, 2013. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/134/5/10.1121/1.4830871

[10] J. J. Ohala, "An ethological perspective on common cross-language utilization of f0 of voice," *Phonetica*, vol. 41, pp. 1–16, 1984.

[11] ——, "The frequency code underlies the sound symbolic use of voice pitch," in *Sound symbolism*, L. Hinton, J. Nichols, and J. J. Ohala, Eds. Cambridge: Cambridge University Press, 1994, pp. 325–347.

[12] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004.

[13] S. Haumann, U. Koch, and K. Sornig, "Politeness in Austria: Politeness and Impoliteness," in *Politeness in Europe*, L. Hickey and M. Stewart, Eds. Multilingual Matters Clevedon, 2005, pp. 82–99.

[14] J. House, "Politeness in Germany: politeness in Germany," in *Politeness in Europe*, L. Hickey and M. Stewart, Eds. Multilingual Matters Clevedon, 2005, pp. 13–28.

[15] L. Brown, B. Winter, K. Idemaru, and S. Grawunder, "Phonetics and politeness: Perceiving Korean Honorific and Non-Honorific Speech trough Phonetic Cues," *Journal of Pragmatics*, 2014, accepted.

[16] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program]*, Version 5.3.10, retrieved 12 March 2012 from http://www.praat.org/, 2012.

[17] N. H. de Jong and T. Wempe, "Automatic measurement of speech rate in spoken dutch," *ACLC Working Papers*, vol. 2, no. 2, pp. 49–58, 2007.

[18] A. P. Simpson, "The first and second harmonics should not be used to measure breathiness in male and female voices." *Journal of Phonetics*, vol. 40, pp. 477–490, 2012.

[19] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *The Journal of the Acoustical Society of America*, vol. 115, no. 3, pp. 1321–32, 2004.

[20] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[21] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0*, http://CRAN.R-project.org/package=lme4, 2012.

[22] D. A. Fournier, H. J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. N. Maunder, A. Nielsen, and J. Sibert, "Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models," *Optimization Methods and Software*, vol. 27, no. 2, pp. 233–249, 2012.

[23] N. Ward, "Non-lexical conversational sounds in american english," *Pragmatics & Cognition*, vol. 14, no. 1, pp. 129–182, 2006.

[24] K. Drager and J. Hay, "Exploiting random intercepts: Two case studies in sociophonetics," *Language Variation and Change*, vol. 24, no. 1, p. 59, 2012.

[25] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," *15 th International Congress of Phonetic Sciences*, pp. 2417–2420, 2003.

Speech rate in the expression of anger: a study with spontaneous speech material

Miguel Oliveira, Jr.[1], *Ayane Nazarela Santos de Almeida*[1], *René Alain Santana de Almeida*[1], *Ebson Wilkerson Silva*[1]

[1] Faculdade de Letras, Universidade Federal de Alagoas, Brazil

miguel@fale.ufal.br, ayanesantos@hotmail.com, renealain@hotmail.com, ebswillk@gmail.com

Abstract

The study of the acoustic expression of emotion is, in general, the analysis of whether prosodic variables such as intonation (F0), speech rate, pauses, rhythm, intensity and duration, are reliable clues for the characterization of the emotional states of the speaker. The present paper aims to verify whether an association exists in Brazilian Portuguese between the basic emotion of "anger" and the prosodic variable "speech rate", as the literature often suggests there is for other languages. The corpus consisted of fragments of spontaneous speech recorded from a radio program. The fragments were selected on the basis of a perceptual test. For the production analysis, only excerpts that were identified by more than 75% of the participants of the perceptual test as associated to the categories "anger" and "neutral" were selected. The results demonstrated that, for the data that were used for the analysis, there is a general reduction in speech rate when utterances are associated with the emotion of "anger", if compared to utterances spoken in a "neutral" mode by the same speaker, contrary to what literature often indicates for other languages.

Index Terms: speech rate, anger, spontaneous speech

1. Introduction

1.1. Speech and emotion

There are several ways in which human emotions can be expressed at the time of communication: through the facial expression of the speaker, the verbal content of the utterance and the acoustic features of the speech, observed through the behavior of prosodic parameters such as intensity, duration, fundamental frequency (F0) and voice quality.

All these parameters collaborate to speech intelligibility and promote the understanding of the expression of emotions in human communication. [1], however, consider that the prosodic characteristics alone may warrant the recognition of emotions expressed by humans.

According to [2], real-world applications, notably those based on human–computer interaction, depend on coming to terms with the ways people express emotion. In order to contribute to the improvement of recognition and synthesis speech systems, [3] found to be of vital importance to speech technology to determine the vocal changes produced by emotional factors in various languages and cultures. Thus, [3] has analyzed the emotion and speech recognition in nine countries from three different continents and found that segmental and suprasegmental aspects, in each of the languages that were analysed, contribute significantly to the production and the perception of the emotional categories.

The classification and the concept of emotion are, however, quite controversial in the literature [4]. [5] asserts that the definitions of emotion are as varied as the researches that address it. The classification of emotions is particularly controversial because of its spontaneous, involuntary and unpredictable feature.

Despite the lack of general agreement in the literature, there are common basic aspects that several papers share on the subject, such as the fact that emotions are almost always directed to the object, that they are activated by internal or external stimuli and that they consist of momentary states of people [5]. It also seems to be a consensus among the researchers that emotion can be broken down into three basic categories: sadness, happiness and anger ([5], [6]).

Each of the emotions, as classified by the literature, is associated with specific prosodic features that individualizes them. [5], [7] and [8], in studies of the Dutch, French and German languages respectively, sought to associate emotional patterns in these languages with acoustic parameters such as pitch levels, vocal range, global average intensity and total utterance duration. In all of these studies, the authors observed that sadness is generally characterized by a lower F0, a lower voice and a slower rate, whereas anger and joy have a higher F0, a louder voice and a faster speech rate.

[9], on the other hand, points out that not always the speakers of a language expresses emotions in the same way with the same activation levels. Then, each of the basic emotions can generate different levels of activation in the speaker, depending on its manifestation. For example, [5] observed that irritation causes a low level of activation in speaker, while fury affects the speaker more intensely.

[10] asserts that there may be very subtle differences in the way basic emotions are manifested for individual speakers. The author, however, does not make reference to different emotions, but to subtypes of the same emotion. In that way, [10] considers "fury" (hot anger) and "irritation" (cold anger), for example, as being two different manifestations of the same emotion: anger.

According to [2], there are inconsistencies for several emotions and emotion-related states that have been studied. These, however, may reflect inconsistent procedure or different interpretation of emotion categories, or variations in terms of the type of data that was used in the analysis: real or simulated. Others, though, seem likely to reflect real differences in the vocal expression of emotion, from speaker to speaker, from culture to culture, and across genders and situations. [2] points out that comparisons between languages and cultures are limited, but they suggest substantial differences.

There are still very few studies on prosodic characteristics of emotions and emotion-ralated events in Brazilian Portuguese. [9], for example, investigated how the expression of three primary emotions affects the characteristic melodic contours of four speech acts in Brazilian Portuguese spoken in Rio de Janeiro. In order to test that, the author used an interaction between four speech acts (assertion, question, request an order) and four emotions (three primary emotions: sadness, happiness and anger, plus the "neutral" form). She

intended to answer whether these two categories – speech acts and emotion – are really independent prosodic dimensions or whether the interaction between them causes substantial changes in intonation patterns that's been found for speech acts alone.

The author found that speech acts and emotional patterns appear to be independent categories in terms of production. With regard to the perception, however, there is some overlap between the two categories. She concludes that intonation and voice quality should be seen as complementary categories, both necessary for the recognition of emotional states of the speaker. It must be pointed out, however, that [9] analyzed utterances produced by two actors from the southeastern area of Brazil. Most of the few studies addressing the prosody of emotions in Brazilian Portuguese is based on the southern and southeastern dialects, and uses monitored utterances produced by actors.

[11] presented a method for speech expressiveness, which combines a dimensional analysis of speech expression, a Principal Component Analysis technique, as well as a multiple regression analysis. The author used as the corpus for his study recordings of a radio show called Programa do Chupim, aired by Rádio Metropolitana de São Paulo, based on a southeastern area of Brazil. He concluded that if utterances are analyzed chronologically, they reveal clear expression changes, from an automatic acoustic analysis, what implies that acoustic parameters are sufficient for the detection of emotion in speech.

[12] analyzed the prosody and expressivity of speech through emotions expressed in a poem spoken by a professional actor from southeastern Brazil. [13], investigated the melodic contour associated to the emotion of anger in theatrical speeches, based on samples of three professional actresses from southeastern Brazil, collected in laboratory condition. [14] also analysed monitored speech with the goal of improving synthesized speech produced by a TTS (text to speech) system, by adding to it emotional information from an acoustic perspective. The model used by them was based on a southeastern dialect of Brazilian Portuguese.

It is, however, unclear to what extent monitored speech truly reflects an emotion that is characterized as involuntary and unpredictable. This is, actually, one of the most important limitations in studies of emotional speech [15]. In this direction, [16] emphasized that high levels of activation are not often found in the case of elicited emotions from control conditions.

1.2. Speech rate

One of the very first references to empirical research on speech rate dates back to [17], who summarizes his findings as follows: "it takes about twice as long to read (aloud, as fast as possible) words which have no connexion as letters which make words… When a passage is read aloud at a normal rate, about the same time is taken for each word as when words having no connexion are read as fast as possible." This study dealt with several languages, such as English, French, German, Italian, Latin and Greek, as he was also concerned with the different rate employed by the same speaker while speaking a foreign language. According to [17], the rate at which someone speaks a foreign language is determined by the familiarity that the person has with the language: the more familiar a person is with it, the faster the speech.

[18] was perhaps the first to introduce objective methodology for the study of rate in speech. Using a rather peculiar method of measurement that relied on soot marking from flames, he compared polysyllabic words with monosyllabic words, using the syllable per second unit of measure – a unit most widely employed today. In his experiment, he demonstrated that, for a given passage containing the same amount of syllables, polysyllabic words are read faster than monosyllabic words. He claims that this is due to the amount of meaning that is carried out in passages containing monosyllabic syllables: the larger the amount of meaning that is conveyed in a message, the longer the speaker will take to utter it.

Meaning is also considered to be a determining factor for the establishment of speech rate in [19]. It compared the repetition of nonsense syllables with the production of syllables articulated within words and found that nonsense syllables are often produced at a slower rate than those syllables that are part of real words. He also noted that people tend to be affected by curiosity when reading a passage of unknown content, which would result in a deliberate acceleration of the speech as a result of this curiosity.

[20] introduced developmental considerations to the study of speech rate, by investigating the speech of kindergarten children. They found that while boys tend to speak less than girls, they do so in a faster rate. [21] considered other external factors. According to these authors, emotions such as anger, fear, and indifference are closely related to fast speech rate, whereas contempt and grief are associated to slow rate.

[22] confirmed the hypothesis that meaning has a decisive influence on speech rate, by demonstrating that the more meaning there is invested in an utterance, the slower the articulation rate is. According to him, this has also to do with the emotional state of the speaker: psychological tension – a result of the demanding task of interpreting new meanings – would be the most immediate reason for speech rate variation.

On the basis of the overview presented above, it may be concluded that there is a historical tradition in the studies on speech rate to relate its observed variation to both semantic and emotional aspects associated to different speech activities. The study on speech rate, as an acoustic correlate of emotion, is still incipient for Brazilian Portuguese, as the literature review presented above indicated. Most of the few studies carried on so far focus on intonational characteristics in speech expressiveness.

One of the few studies for Brazilian Portuguese that addresses the acoustic parameter of speech rate as a correlate of emotion is that of [9]. According to her findings, there isn't a straightforward relation between duration and different basic emotions: duration may vary as a function of individual speakers. (See also [23]).

1.3. Research question

The main objective of the present study is to investigate whether speech rate is a reliable acoustic correlate, in Brazilian Portuguese, of a basic human emotion: anger. The rationale for this study lies in the scarcity of such investigation, despite the fact that the literature often relates the acoustic parameter of speech rate to human emotion ([5], [7], [8], [9]).

Contrary to most of the very few studies on the prosody of emotion in Brazilian Portuguese, the present investigation uses spontaneous material, representative of a dialect that has been neglected so far in this kind of investigation: that spoken in the northeastern area of Brazil.

2. Methods

The data used in this study consist of eighteen small fragments of recordings of spontaneous speech, extracted from a radio program called "A Hora do Mução". It is a popular practical joke program, in which the radio broadcaster Rodrigo Vieira Emerenciano embodies the character Mução. The radio broadcaster calls people in the northeastern area of Brazil, from suggestions made by friends and relatives of the victim, in order to annoy them by constantly making reference to a physical characteristic or a nickname of the victm. One of the advantages of this kind of interaction, in radio program, is the possibility of obtaining high levels of arousal of affective responses, due to the critical events introduced by the program presenter.

In the recordings, there is a clear difference between the speech of the called parties in the first half of the phone calls and their second half, when the reason for the calls is known (i.e. the practical joke is in action). In the second half of the calls, the speech of the called party is clearly linked to the expression of anger.

The recordings of this radio program are, therefore, an excelent material for the type of analysis proposed here because they make it possible to compare, for each individual speaker, prosodic features associated with neutral speech to those associated with the expression of anger.

Furthermore, contrary to what happens with most of the data used in research about the relationship between prosody and emotion, the recordings used in the present investigation reflect the speech associated with a spontaneously occurred emotion, what makes it ecologically valid. It must also be pointed out that telephone speech is ideal for the study of vocal signs of emotion, because contrary to face to face interactions, it doesn't present any visual information that could distract the analysis ([2], [24]).

The excerpts of the recordings were selected from a perceptual test conducted with forty graduate students in Arts at Federal University of Alagoas. The data that was used in this perception test consisted of fragments of recordings available on CD "Pegadinhas do Mução" [25], which were selected based on the following criteria: (i) sound quality (intelligible and noiseless signal), (ii) sex the speaker (all excerpts were extracted from recordings featuring men only, in order to avoid the inclusion of a variable that could confound the interpretation of the results), and (iii) content of lexical information (the extract were selected by taking into account their contents: for obvious reasons, they shouldn't offer any evident clues of anger – or any other emotion, for that matter). Regarding this last criterion, however, it must be pointed out that previous studies have demonstrated that prosody processing is segregated from linguistic semantic processing, probably because both depend on partially dissociated neural mechanisms ([26], [27]).

We understand that the corpus is reduced, however, the difficulty in obtaining data from spontaneous speech justifies this limitation. Moreover, we could not use excerpts of the recordings containing insults and in the case of emotion "anger", this occurs very often.

Transcriptions of the selected excerpts from recordings of three practical jokes, made with three different men, all speakers of northeastern dialects, along with their corresponding audio files were presented in slide-show, at random, to the participants of the perceptual test. For each excerpts, the participant had to identify an emotion related to it. The options were the three basic emotions: "anger", "happiness" and "sadness", as well as "neutral". The participants were instructed to listen to each stimulus as often as they wished before answering the form.

For the production analysis, only excerpts that were identified by more than 75% of the participants of the perceptual test as associated to the categories "anger" and "neutral" were selected.

There is a variety of units of measurement that are employed in the research on speech rate. The units range from sounds per unit of time ([22], [28], [29], [30]), to words per unit of time ([31], [32], [33], [34], [35]), to syllables per unit of time ([34], [36], [37]) and finally to beats per unit of time [35]. This plethora of units of measurements employed in the literature not only reflects a serious methodological flaw – as discussed in [37], but also makes the essential task of comparing results among various studies impossible.

The best-suited unit of measurement for speech rate, according to many authors, is the syllables per unit of time ([38], [39]). [36], for example, defines speech rate as the "rate of syllable succession." This is the unit adopted by [34], [38], [40], [41], [42], [43], [44] and [45], to name a few. Even though, as [39] points out, this unit of measurement has also the disadvantage of not taking into consideration the processes that may result in syllable omission, that are often found in rapid speech, such as assimilation and segmental deletion, what would obviously not be covered in this unit of measurement.

Therefore, the present study opted for a measure that is mostly used in the temporal research of speech for the sake of comparability. It does recognize the pitfalls related to this choice, but assumes that they are not so serious as to invalidate the analysis. Speech rate will be interpreted in this study using the measurement of syllables per second.

3. Results

Figure 1 below shows average speech rates of the speech samples that were selected for analysis, based on the perceptual test. It is broken down by emotion ("neutral" and "anger") and speaker (a total of three).

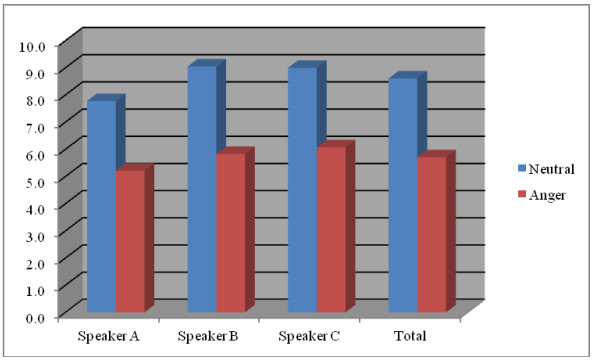


Figure 1: *Average speech rate (in syllables per second) of "neutral" and "anger" utterances, broken down by speakers.*

We used the mean for comparison between utterances associated with each emotion because we found that the difference between it and the median was insignificant, which allows us to state that there is no significant distortion in the mean rates found in analyzed excerpts and, therefore, can be considered a reliable measure for this data.

The results shown in Figure 1 clearly indicate that speakers employed a slower rate of speech in utterances associated to the emotion of anger, if compared to those considered to be neutral, both individually as well as a group. Speech rate of the excerpts labeled as "neutral" was, in average, 3 syllables per second faster than those that were labeled as associated to the emotion of anger.

A series of paired-samples two tailed t-tests were conducted to assess the significance of the differences in speech rate between "neutral" and "anger" expressions for each speaker and for the total sample. Table 1 below shows the results of these tests:

Table 1. *Results of t-tests, broken down by speaker and by total sample.*

Sample	Results
Speaker A	[$t(5) = 7.02$, $p < 0.0009$]
Speaker B	[$t(5) = 11.08$, $p < 0.0001$]
Speaker C	[$t(5) = 6.77$, $p < 0.0011$]
Total sample	[$t(17) = 13.86$, $p < 1.08E-10$]

Table 1 evidences that for all speakers, speech rates differ significantly between the two conditions: "neutral" and "anger". It also indicates that the differences are significant when subjects are regarded as a group. Results in Table 1 show a low probability of the results being a fortuity, as indicated by the values of p.

4. Discussion

This study set out to investigate whether speech rate is a reliable acoustic correlate, in Brazilian Portuguese, of a specific basic emotion: "anger". It followed the tradition of studies that explore how speakers use prosody to encode discrete emotions, such as happiness, anger, disgust, etc.

Emotional prosody in Brazilian Portuguese has been studied very scarcely so far. The few papers that address the issue deal almost exclusively with simulated emotion and are based on data representative of dialects spoken in the southeastern area o Brazil. In order to contribute with the still incipient research on emotional prosody in Brazilian Portuguese, the present investigation used spontaneous recordings representative of dialects spoken in northeastern Brazil.

According to the literature, the emotion of "anger" is characterized acoustically by fast speech rate, high voice intensity, high F0/pitch level, much F0/pitch variability, rising F0/pitch contour, fast voice onsets, and microstructural irregularity ([5], [7], [8], [9]).

What the general results from the analysis of speech rate reported here suggest is that there are possible dissonances between what the literature establishes as a prosodic pattern associated with the emotion of "anger" and what has been identified as a pattern for the same emotion in Brazilian Portuguese, based on spontaneous speech material. The results demonstrate that there is a general reduction in speech rate when utterances are associated with the basic emotion of "anger", if compared to utterances spoken in a "neutral" mode by the same speaker. Statistical analysis indicate that this difference is not a fortuity or due to individual characteristics of the speakers, because the samples of each speaker were analyzed individually, as well as a group, and very similar characteristics with regards to this speech rate's variable in the expression of anger was found.

It is known that emotions are manifested in a continuous way, in varying dimensions, depending on the level of activation / stimulation [9]. The emotion of "anger", for example, can be expressed on a scale ranging from a mild "irritation" to a "wrath".

In this paper, however, we didn't consider the varied levels of activation, but rather the basic emotion "anger". We are of course aware that, depending on where in the continuum a given emotion is located, its corresponding acoustic characteristic may vary [46]. Notwithstanding, faster speech rate is commonly associated to any level of activation in the scale of "anger" [10].

The difference in terms of speech rate as a correlate of the emotion of "anger" reported here, as compared to the results reported elsewhere, may be interpreted as a result of the specificities of data that were used for the analysis: spontaneous material, derived from telephone speech, uttered by male speakers of scarcely studied dialects of Brazilian Portuguese. As [2] points out, variations in this kind of study may reflect differences in the vocal expression of emotion, from speaker to speaker, from culture to culture, and across genders and situations. In order to find out whether this is systematically the case, further investigation needs to be done, with a larger corpus contemplating all possible variables that are potentially important for any attempt to generalize.

5. Conclusion

According to [47], communication of emotions is crucial to social relationships and survival. It is, thus, essential to understand all the aspects related to it, including its acoustic properties. It has been suggested, however, that signs of emotion in speech is not consistent across individuals and occasions, what calls for systematic descriptions of their properties for proper comparisons between languages and cultures.

The analysis presented here allowed us to infer that in spontaneous speech data representing the variety of Portuguese spoken in northeastern Brazil, there is not an association between the basic emotion of "anger" and an increase in speech rate, as the literature often indicates for other languages. Instead, many excerpts identified as being representative of the emotion "anger" by the participants of the perception test were enunciated with a speech rate slower than that in the excerpts perceived as "neutral".

Contrary to most studies on the prosody of emotion, the present study opted to use spontaneous material for the analysis, because, as [10], we believe that if research continues to be almost exclusively concerned with the simulation of emotion, nothing much will be gained in terms of getting to understand to what extent we are operating within a closed system of association or translation rules that may be only indirectly tied to the underlying biology of emotion.

The present paper stands as a contribution to the acoustic characterization of emotional patterns in Brazilian Portuguese. We plan for future investigation to enlarge the corpus, as to reflect other variables, and to include other prosodic parameters in the analysis, such as variation in fundamental frequency, pause and intensity. Perceptual tests with stylized samples of the corpus are also planned, as we understand that it is fundamental to describe what prosodic parameters are relevant from a perceptual perspective.

6. References

[1] Schröder, M. “Experimental study of affect bursts”. Speech communication special issue speech and emotion, 40: 99-116, 2003.

[2] Douglas-Cowie, E. et al. “Emotional speech: Towards a new generation of databases”. Speech Communication, 40(1): 33-60, 2003.

[3] Scherer, K. “A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology”. Proceedings of 6th ICSLP, 2: 379-382, Beijing, 2000.

[4] Cowie, R. and Cornelius, R. R. “Describing the emotional states that are expressed in speech”. Speech Communication, 40(1): 5-32, 2003.

[5] Paeschke, A. “Prosodische Analyse emotionaler Sprechweise”. Berlin: Logos Verlag, 2003.

[6] Sawamura, K., Deng, J., Akagi, M., Erickson, D., Li, A., Sakuraba, K., Minematsu, N. and Hirose, K. “Common factors in emotion perception among different cultures”. Proceedings of 16th ICPhS, Saarbrücken, 2113-2116, 2007.

[7] Leon, P. “Précis de phonostylistique parole et expressivité”. Paris: Natan, 1993.

[8] Bezooyen, R. “Characteristics and recognizability of vocal expression of emotion”. Dordrecht: Foris, 1984.

[9] Pereira, M. C. C. “A expressão das emoções em atos de fala no português do Brasil: produção e percepção”. Dissertação de Mestrado em Letras Vernáculas, Universidade Federal do Rio de Janeiro, 2009. Online: http://www.letras.ufrj.br/posverna/mestrado/PereiraMCC.pdf, accessed on 29 Jul 2013.

[10] Scherer, K. “Vocal affect expression: a review and a model for future research”. Psychological Bulletin, 99: 143-165, 1986.

[11] Barbosa, P. A. “Detecting changes in speech expressiveness in participants of a radio program”. Proceedings of Interspeech, 2155-2158, 2009.

[12] Santos, I. “Expressividade da Fala:o desvelar da locução de um poema a partir da Análise Acústica e da filosofia de Spinoza”. Dissertação de Mestrado em Linguística Aplicada e Estudos da Linguagem, Pontificia Universidade Católica de São Paulo, 2010. Online: http://www.sapientia.pucsp.br/tde_busca/arquivo.php?codArquivo=12132, accessed on 13 Aug 2013.

[13] Vassoler, A. M. O. and Martins, M. V. M. “A entoação em falas teatrais: uma análise da raiva e da fala neutra”. Estudos Linguísticos, 42: 9-18, 2013.

[14] Reis, B. F. and Martins, V. V. “Síntese prosódica da fala em português do Brasil”. X SBAI, 1185-1188, 2011.

[15] Swerts, M. and Hirschberg, J. “Prosodic predictors of upcoming positive or negative content in spoken messages”. J. Acoust. Soc. Am., 128(3): 1337, 2010.

[16] Scherer, K. R. “Vocal communication of emotion: A review of research paradigms”. Speech Communication, 40: 227-256, 2003.

[17] Cattell, J. M. “The time it takes to see and name objects”. Mind, 11: 63-65, 1886.

[18] Beer, M. “Die Abhängigkeit der Lesezeit von psychologischen un sprachlichen Faktoren”. Zeitschrift für Psychologie, 56: 264-298, 1910.

[19] Fröschels, E. “Untersuchungen über das Sprechtempo”. Monatsschrift für Ohrenheilkunde und Laryngo-Rhinologie, 54: 867-871, 1920.

[20] Olson, W. C. and Koetzle, V. S. “Amount and rate of talking of young children”. Journal of Experimental Education, 5: 175-179, 1936.

[21] Fairbanks, G. and Hoaglin, L. W. “An experimental study of the durational characteristics of the voice during the expression of emotion”. Speech Monographs, 7: 85-90, 1940.

[22] Essen, O. V. “Sprech tempo als Ausdruck psychischen Geschehens”. Zeitschrift für Phonetik, 3: 317-341, 1949.

[23] Moraes, J.; Rilliard, A. “Prosody and Emotion in Brazilian Portuguese”, in M. Armstrong, N. Henriksen, and M. Vanrell [Eds], Interdisciplinary approaches to intonational grammar in Ibero-Romance intonation, Issues in Hispanic and Lusophone Linguistics, to appear.

[24] Paulmann, S. and Pell, M. D. “Is there an advantage for recognizing multi-modal emotional stimuli?” Motivation and Emotion, 35: 192-201, 2011.

[25] Emerenciano, R. V. “Pegadinhas do Mução”. Rio de Janeiro: Polysom. 1 disco compacto: digital, stereo, 2010.

[26] Friederici, A. D. and Alter, K. “Lateralization of auditory language functions: A dynamic dual pathway model”. Brain & Language, 89: 267-276, 2004.

[27] Ross, E. D. and Monnot, M. “Neurology of affective prosody and its functional–anatomic organization in right hemisphere”. Brain & Language, 104: 51-74, 2008.

[28] Heinitz, W. “Die Bewertung der Dauer in phonetischen Aufnahmen”. Vox 153, 1921.

[29] Hegedüs, L. “Sprechtempoanalysen im Ungarischen”. Zeitschift für Phonetik, 10: 8-20, 1957.

[30] Fónagy, I. and Magdics, K. “Speed of utterance in phrases of different lengths”. Language and Speech, 4: 179-192, 1960.

[31] Brubaker, R. S. “Rate and pause characteristics of oral reading”. Journal of Psycholinguistic Research, 1(2), 1972.

[32] Barik, H. C. “Cross-linguistic study of temporal characteristics of different speech materials”. Language and Speech, 20: 116-126, 1977.

[33] Grosjean, F. and Deschamps, A. “Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation”. Phonetica, 14: 141-148, 1975.

[34] Grosjean, F. and Deschamps, A. “Analyse des variables temporelles du français spontané”. Phonetica, 26: 126-156, 1972.

[35] Scollon, R. “Tempo, Density, and Silence: Rhythms in Ordinary Talk”. Center for Cross-Cultural Studies. Fairbanks, University of Alaska, 1981.

[36] Abercrombie, D. “Elements of General Phonetics”. Edinburgh: Edinburgh University Press, 1967.

[37] Meinhold, G. “Allgemeine Probleme der Sprechgeschwindigkeit”. Zeitschrift für Phonetik, 25: 492-505, 1972.

[38] Uhmann, S. “Contextualizing relevance: on some forms and functions of speech rate changes in everyday conversation”, in P. Auer and A. D. Luzio, The Contextualization of Language. Amsterdam, Benjamins: 297-336, 1992.

[39] O'Connell, D. C. and Kowal, T. D. “Cross-linguistic pause and rate phenomena in adults and adolescents”. Journal of Psycholinguistic Research, 1: 155-164, 1972.

[40] Blaauw, E. “On the perceptual classification of spontaneous and read speech”. Research Institute for Language and Speech, Utrecht University, 1995.

[41] Fon, J. “Speech rate as a reflection of variance and invariance in conceptual planning in storytelling”. Proceeding of the ICPhS, 1999.

[42] Grosz, B. and Hirschberg, J. “Some intonational characteristics of discourse structure”. Proceeding of the International Conference on Spoken Language Processing, Banff, 1992.

[43] Goldman-Eisler, F. “The rate of changes in the rate of articulation”. Language and Speech, 4: 171-174, 1961.

[44] Van Donzel, M. “Prosodic Aspects of Information Structure in Discourse”. Faculteit der Geesteswetenschappen, Amsterdam, University van Amsterdam: 195, 1999.

[45] Wood, S. “Speech tempo”. Working Papers of the Phonetics Laboratory, Lund University, 9: 99-147, 1975.

[46] Scherer, K. “Vocal measurement of emotion”. In Plutchik, R.; Kellerman, H. (Org.). Emotion: Theory, research, and experience. The measurement of emotion, 4: 233-260, New York: Academic Press, 1989.

[47] Ekman, P. “An argument for basic emotions”. Cognition & Emotion, 6: 169-200, 1992.

Audiovisual Perception of Expressions of Mandarin Chinese social affects by French L2 Learners

Yan Lu[1], Véronique Aubergé[2], Nicolas Audibert[3], Albert Rilliard[4]

[1] GIPSA Lab, CNRS, Stendhal University, Grenoble France
[2] LIG Lab, CNRS, Grenoble France
[3] Laboratoire de Phonétique et Phonologie, Univ. Sorbonne-Nouvelle/CNRS, Paris, France
[4] LIMSI-CNRS, Orsay, France
yan.lu@gipsa-lab.grenoble-inp.fr, Veronique.Auberge@imag.fr,
nicolas.audibert@univ-paris3.fr, albert.rilliard@limsi.fr

Abstract

This study focuses on confusions made by French L2 learners vs. native subjects in the perception of 11 audiovisual Mandarin Chinese social affects. Two groups of French L2 learners of Mandarin Chinese were selected according to their Chinese level : 9 beginners (A1) vs. 10 intermediate learners (A2). Subjects evaluated the 11 social affects in audio, visual and audiovisual condition. Comparison of confusions between learners of level A1 vs. A2 indicates few significant differences, mostly in audiovisual condition and without a clear gain for one group over the other. The comparison of French L2 learners pooled together vs. native speakers reference sheds light on major confusions to be targeted by specific methods and exercises. Cross-modality comparisons suggest a limited contribution of informations conveyed by acoustic prosody in the identification of audiovisual social affects by L2 learners.

Index Terms: social affects, attitudes, audio-visual perception, L2 prosody, Mandarin Chinese

1. Introduction

The face to face interaction functions of prosody are the main vector of the "socio-affective glue" that builds the communication channel [1] and are expressed following different cognitive processing levels [2]: from involuntary controlled expressions (emotions) to the voluntary control of the social affects of the speaker [3,4]: intentions, attitudes, social cues etc). During the face-to-face communication, people express their affects within the audio-visual speech prosody [5], and as social affects are constructed socially for and by the language, and prosodic realization of one specific social affect in a specific language may be ambiguous or unknown in the learner's language [6], the cross-cultural approach looks more convenient to spotlight the cultural specifications of the expression of social affects. Meanwhile, the social affects are generally learned in childhood within the language community in question and they can also be learned by the learners of a foreign language or a second language if they are different from the social affects in their mother language. But it is preliminarily necessary for foreign learners to recognize first the social affects expressed in the target language. Some studies have shown that the foreign language learners did not perceive the attitude expressed in the target language in the same way with the native people, and the attitudinal prosody needs to be taught in the L2 class [7, 8].

For these reasons, we intend in this study to examine in a cross-cultural context the perception of the audio-visual expressions of Chinese social affective prosody by both native subjects and French learners of Mandarin Chinese. The largest part of the analysis focuses on differences between native Chinese listeners and French learners for the perception of the same social affect, investigating the potential influence of the learners' language skill on their perception of these social affects. The relative contribution of the acoustic vs. visual modalities is also compared across groups of subjects. Finally, consequences for L2 teaching of differences in the perception of social affects between learners and native speakers will be discussed.

2. Method

2.1. Selection of social affects

A large audio-visual Chinese corpus (acted speech) of 19 social affects (each expressed on a set of utterances varying within length, syntax and tone location) was initially validated in an acoustic only perception experiment [9].On the basis of an acoustic cross-perception experiment by French naive listeners [10], 11 social affects were selected among the 19. These social affects were observed problematic for both native and foreign listeners during the previous perception experiments because of their "attractivity" (cumulated percentage of confusions from other attitudes to each one) and showed great difference in perception behavior between native subjects and foreign ones. Hence, they are supposed to be also problematic and difficult for French learners of Chinese in their face-to-face communication with native speakers. The 11 social affects selected are composed of the attitudes, intention or opinion of the speaker about what he says; the characteristics of the social relation implied in the interaction (e.g. "politeness") and the socio-cultural context of interaction (e.g. "infant-directed speech"). Table 1 presents the 11 Chinese social affects and their abbreviation.

Table 1. *Summary of the 11 social affects selected*

Social affects and abbreviation	
declaration (DECL)	obviousness (OBVI)
question (QUES)	neutral surprise (NEU-S)
irritation (IRRI)	politeness (POLI)
doubt (DOUB)	authority (AUTH)
contempt(CONT)	infant-directed speech (IDS)
disappointment (DISA)	

2.2. Audiovisual corpus

The audiovisual speech corpus is based on a 4-syllable long sentence, which is constructed to bear a literally neutral meaning but could be expressed with all social affects studied.

This sentence was performed with 11 social affects by one native Chinese female speaker, who speaks an unmarked standard Mandarin Chinese. The audio part of the corpus has been validated in [9], where all of 11 attitudes have been recognized over chance level. The sentence used in this experiment was better recognized for all social affects and considered the most representative of a prototypical expression of the targeted social affect. This sentence is "四天三夜" (si4 tian1 san1 ye4, "four days and three nights" in English). Thus, 11 short videos were used in the present experiment.

2.3. Subjects

A first group consisted of 30 Chinese native listeners (12 males and 18 females, mean age = 33.3) as the reference group of "optimal" performances of Chinese perception. Two groups of L2 learners are composed according to their results to a test of placement in Chinese language according to CECRL (the Common European Framework of Reference for Languages), taken at the beginning of the term. One group was composed of 9 French learners of Mandarin Chinese whose acquired Chinese level is A1: beginners, having taken less than 100 hours of teaching (1 male and 8 females, mean age = 20.7). Another group was composed of 10 French learners of level A2 – having taken less than 200 hours of teaching (3 males and 7 females, mean age = 22.3). All 19 French subjects study the Mandarin Chinese as foreign language in the LANSAD (Languages for the specialists of other disciplines) Department of University Stendhal-Grenoble 3.

Subjects in each group were divided randomly into two sub-groups of equal size with a different presentation order: audio only → video only → audiovisual in one sub-group, and video only →audio only →audiovisual condition in the other.

2.4. Perceptual evaluation protocol

Before the test, subjects were briefly presented the setting of the experiment and a description of each attitude with examples of situations in which it can happen. They took the test in a quiet room with closed headphone, using a graphical user interface developed with LiveCode® for stimuli presentation and answers collection. Each stimulus was presented once, in a different random order for each subject.

The Chinese audiovisual corpus was presented to subjects in three different conditions:

- Audio only (AU): in this condition, images were hidden. Subjects were instructed to listen to what the speaker said before answering.
- Video only (VI): in this condition, the sound was turned off. Subjects were instructed to carefully watch the facial and body movements of the speaker.
- Audiovisual (AV): in this condition, both modalities were presented simultaneously. Subjects were instructed to watch the facial and body movements of the speaker and at the same time to listen to her voice

For each stimulus presented, subjects were asked to judge which attitude was expressed by the speaker, by choosing among the eleven labels proposed.

3. Results

Subjects' answers were pooled into a confusion matrix for each group of listeners x presentation condition. In each presentation condition, confusion matrices cells values were compared against chance level, and between groups using chi-squared tests for comparison of proportions.

3.1. Native subjects performance

As expected from previous studies on this corpus with Chinese and completely naive French listeners, the native listeners performed better than French learners for all modalities of presentation of social affects. In audiovisual condition, they recognized significantly over chance level (9%) all attitudes except "contempt" (recognition rate: 33%). "Declaration" and "irritation" were significantly recognized over chance in all three conditions.

3.2. A1 vs. A2 learners

A first analysis was performed comparing confusions of L2 learners with level A1 vs. L2 learners with level A2, revealing few inter-group differences. The largest part of significant differences was found in audiovisual presentation condition. Compared to group A1, the group A2 confused significantly less question with obviousness and doubt with neutral surprise, outperforming both native subjects and A1 learners in their identification of the expression of doubt. Surprisingly, a better performance of group A1 was found in the identification of authority. However, a comparison with native subjects reveals that the confusion pattern observed in group A2 for expressions of authority is similar to native subjects confusions, authority being largely confused with obviousness.

In audio-only condition, the expression of doubt was significantly less confused with disappointment by learners of the group A2 compared to group A1. In visual-only condition, learners in group A2 confused significantly less contempt with irritation than group A1, but they confused significantly more doubt with irritation.

3.3. Native subjects vs. French L2 learners

Results reported supra do not picture a clear advantage in performances of the group of A2 learners vs. A1. In order to get more insight into major differences between native Mandarin Chinese subjects and French L2 learners, answers in groups A1 and A2 are pooled altogether in a 'L2 learners' group (19 subjects) in the following analysis.

Table 2, 3 and 4 summarize confusions by the 30 native Mandarin Chinese listeners and French L2 learners respectively in audio-only, visual-only and audiovisual presentation condition, all learners answers pooled. Each cell in the confusion matrices has two values, native subjects' performance (top) and L2 learners' performance (bottom). Values significantly different from chance level are flagged by stars on the right part. Stars on the left part of a cell indicate a significant difference between the group of native speakers and the group of French L2 learners.

For instance, the cell corresponding to neutral surprise (NEU-S, 3rd column, 3rd line) in the matrix diagonal in audio-only condition (Table 2) indicates that native Mandarin Chinese subjects identified neutral surprise at 60%, which is significantly higher than chance ($p<.01$), while L2 learners did not identify it better than chance (16%, $p>.05$). The left part of this cell also reports that the ratio of correct identification of audiovisual neutral surprise is significantly higher ($p<.01$) in the native subjects group than in the L2 learners group.

Table 2. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in audio-only condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

Audio	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	50% *	0%	0%	0%	0%	10%	33%	0%	7%	0%	0%
	32%	0%	0%	0%	0%	21%	21%	0%	26%	0%	0%
NEU-S	0%	** 60% **	10%	13%	0%	0%	10%	0%	0%	7%	0%
	0%	16%	21%	5%	0%	0%	32%	5%	0%	16%	5%
QUES	7%	13%	57% **	13%	3%	0%	* 0%	3%	0%	0%	3%
	0%	26%	32%	16%	0%	0%	16%	0%	0%	0%	11%
DOUB	3%	10%	23%	27%	0%	3%	10%	13%	0%	3%	7%
	11%	26%	16%	11%	0%	0%	0%	11%	0%	11%	16%
IDS	13%	0%	7%	7%	53% *	7%	7%	3%	0%	0%	3%
	16%	11%	0%	0%	68% **	0%	0%	0%	0%	0%	5%
POLI	43% *	0%	0%	0%	3%	** 30%	10%	0%	13%	0%	0%
	53% *	0%	0%	5%	16%	0%	5%	5%	0%	5%	11%
OBVI	27%	0%	3%	0%	0%	3%	40%	10%	13%	0%	3%
	32%	0%	5%	11%	0%	0%	16%	16%	21%	0%	0%
CONT	3%	17%	* 7%	30%	0%	0%	7%	27%	3%	3%	3%
	11%	5%	32%	16%	5%	0%	5%	11%	0%	11%	5%
AUTH	13%	0%	0%	0%	0%	0%	** 37%	3%	33%	10%	3%
	26%	0%	0%	0%	0%	5%	0%	11%	47% *	11%	0%
IRRI	10%	3%	3%	3%	0%	0%	13%	7%	13%	43% *	3%
	5%	11%	0%	5%	0%	0%	11%	5%	26%	32%	5%
DISA	* 37%	0%	0%	0%	3%	7%	10%	7%	7%	0%	30%
	5%	5%	0%	0%	5%	11%	5%	11%	0%	5%	53% *

Table 3. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in visual-only condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

Visual	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	50% *	0%	3%	3%	3%	10%	20%	0%	7%	0%	3%
	58% **	5%	0%	5%	0%	0%	5%	5%	11%	0%	11%
NEU-S	7%	23%	30%	13%	7%	10%	7%	0%	0%	0%	3%
	5%	11%	32%	5%	0%	0%	16%	0%	11%	5%	16%
QUES	33%	0%	10%	0%	10%	20%	17%	0%	10%	0%	0%
	53% *	5%	5%	0%	0%	21%	16%	0%	0%	0%	0%
DOUB	0%	3%	23%	50% *	0%	3%	0%	0%	0%	7%	13%
	0%	5%	11%	53% *	0%	0%	0%	0%	0%	21%	11%
IDS	10%	0%	7%	0%	27%	40%	13%	0%	3%	0%	0%
	32%	5%	5%	0%	21%	21%	11%	0%	5%	0%	0%
POLI	23%	7%	0%	0%	10%	50% *	10%	0%	0%	0%	0%
	21%	11%	0%	0%	5%	37%	21%	0%	0%	5%	0%
OBVI	* 3%	0%	3%	3%	0%	13%	37%	13%	13%	0%	13%
	26%	5%	0%	0%	0%	11%	47% *	0%	0%	0%	11%
CONT	10%	3%	7%	10%	0%	7%	10%	* 23%	0%	* 3%	27%
	11%	0%	0%	26%	5%	0%	0%	0%	0%	26%	32%
AUTH	20%	0%	3%	0%	3%	0%	20%	10%	33%	3%	7%
	11%	0%	0%	5%	0%	0%	32%	11%	37%	5%	0%
IRRI	0%	0%	3%	10%	0%	0%	0%	20%	0%	43% *	23%
	0%	0%	0%	11%	0%	0%	0%	26%	0%	37%	26%
DISA	7%	0%	7%	0%	0%	0%	0%	23%	0%	0%	63% **
	5%	11%	0%	** 21%	0%	0%	5%	11%	0%	0%	47% *

- For audio only modality: native subjects recognized significantly better neutral surprise and politeness than French learners; the native subjects confused more authority with obviousness and disappointment with declaration than French learners; on the other hand, French learners confused more contempt with question and question with obviousness than native subjects.
- For video only modality: native subjects recognized better contempt than French learners (A1: 0%, A2: 0%); French learners identified mistakenly more obviousness with declaration, contempt with irritation, and disappointment with doubt than native subjects.
- For audiovisual modality: native subjects identified better neutral surprise than French learners, who, on the contrary, recognized better infant-directed speech; native subjects confused more declaration with obviousness, doubt with question and contempt with doubt than French L2 learners. However for L2 learners, neutral surprise was more confused with irritation, question with obviousness and contempt with disappointment.

Table 4. *Confusion matrix for 11 Chinese social affects of 30 Chinese native listeners (top position in cells) vs. 19 French listeners (bottom position in cells), in audio-visual condition. Lines: presented attitudes; columns: recognized attitudes. Stars indicate significant differences (chi-squared test for proportions). Left part of cells: native Mandarin Chinese speakers vs. French learners performance; Right part: comparison with chance level for both groups of listeners: *: p<.05; **: p<.01.*

AV	DECL	NEU-S	QUES	DOUB	IDS	POLI	OBVI	CONT	AUTH	IRRI	DISA
DECL	53% *	0%	0%	0%	0%	10%	* 33%	0%	3%	0%	0%
	79% **	0%	0%	0%	0%	11%	* 5%	0%	5%	0%	0%
NEU-S	7%	* 67% **	10%	17%	0%	0%	0%	0%	0%	** 0%	0%
	0%	* 32%	26%	0%	0%	0%	11%	5%	5%	** 21%	0%
QUES	10%	13%	47% *	23%	3%	3%	* 0%	0%	0%	0%	0%
	16%	32%	21%	11%	0%	5%	* 16%	0%	0%	0%	0%
DOUB	0%	10%	* 30%	47% *	0%	0%	0%	7%	0%	7%	0%
	0%	16%	* 5%	63% **	0%	0%	0%	5%	0%	5%	5%
IDS	10%	0%	0%	3%	* 57% **	17%	10%	3%	0%	0%	0%
	11%	0%	5%	0%	* 84% **	0%	0%	0%	0%	0%	0%
POLI	33%	0%	0%	0%	7%	50% *	10%	0%	0%	0%	0%
	16%	0%	0%	5%	11%	63% **	5%	0%	0%	0%	0%
OBVI	23%	0%	0%	0%	0%	3%	53% *	10%	10%	0%	0%
	21%	0%	0%	0%	0%	0%	63% **	5%	11%	0%	0%
CONT	3%	0%	13%	* 20%	0%	0%	7%	33%	0%	3%	* 20%
	5%	0%	11%	* 0%	0%	0%	5%	21%	0%	11%	* 47% *
AUTH	3%	0%	0%	0%	0%	0%	37%	0%	60% **	0%	0%
	0%	0%	0%	0%	0%	0%	16%	5%	74% **	5%	0%
IRRI	3%	0%	0%	0%	0%	0%	0%	27%	0%	63% **	7%
	0%	0%	0%	11%	0%	0%	0%	42%	5%	37%	5%
DISA	7%	0%	0%	0%	0%	0%	0%	7%	0%	0%	87% **
	0%	5%	0%	0%	5%	0%	0%	11%	0%	0%	79% **

Identification rates and confusions were also compared between presentation conditions using chi-square tests, for each group of subjects. For the sake of concision, those results are not reported extensively in this paper. Cross-condition comparisons indicate different multimodal strategies for the identification of social affects between groups. While native subjects tend to rely more on acoustic cues (with neutral surprise, question and infant-directed speech significantly better recognized in audio condition, and only disappointment better recognized in visual condition), this tendency is only partly reproduced in L2 learners performance (with question and infant-directed speech better recognized in audio condition, and doubt, politeness and obviousness better recognized in visual condition). Most differences between native subjects and L2 learners are found in the comparison of audiovisual vs. audio-only condition: while native subjects show a significant gain in audiovisual condition for authority and disappointment, L2 learners significantly benefit from the audiovisual information for declaration, doubt, politeness and obviousness.

4. Discussion and conclusion

This paper investigated the perceptual behavior of 19 French learners of Mandarin Chinese vs. 30 native listeners for 11 Mandarin Chinese social affects, presented in three different conditions: audio-only, video-only and audiovisual conditions. Meanwhile it also examined the correlation between the listeners' language skill and their perceptual behavior.

According to the results of analysis, the perception of all subject groups for the audiovisual modality shows the best scores for almost all attitudes [7]. Though differences were found between groups of learners with different level, the French learners in A2 level showed no clear advantage over the learners in A1 level. However, significant differences were found between native Chinese speakers and French L2 learners as a whole: the native listeners recognized better "neutral surprise" in both audio only and audiovisual modalities, and it was more confused with "question" in audiovisual modality by French learners; "politeness" was better recognized by native subjects in audio only modality; although "contempt" was relatively better recognized by native subjects in video only modality, in fact, the recognition rate of native subjects was not satisfactory (23%); "infant-directed speech" was unexpectedly better recognized by French learners than by native subjects. Limited differences in audio-visual condition between native Mandarin Chinese subjects and French L2 learners suggest that in a face-to-face communication context, L2 learners might compensate their relative inability to identify the acoustic prosodic correlates of social affects by relying more extensively on visual cues.

Meanwhile, French learners showed more difficulties in recognizing "politeness" in audio only condition, "question" and "contempt" in video only condition. "Politeness" was mostly confused with "declaration", that may be because of their similar prosodic characteristic. The facial expression of "question" was considered similar with that of "declaration", because both of them are basic communicative functions expressed by utterance modalities and are neutral in terms of affective sate. Acoustically, "question" was more confused with "neutral surprise" by French learners. These observations suggest that L2 teaching of Mandarin Chinese for French learners could benefit from integrating specific exercises on social affects, particularly concerning their acoustic realization with a focus on "neutral surprise" and "politeness".

5. Acknowledgements

This study was supported jointly by the French National IDEFI Innovalangues and the Major Program for the National Social Science Fund of China (13&ZD189).

6. References

[1] Aubergé V., Sasa Y., Robert T., Bonnefond N., Meillon B. (2013) “Emoz: a wizard of Oz for emerging the socio-affective glue with a non humanoid companion robot”. In proceedings of WASSS 2013, Grenoble, France.

[2] Aubergé, V., “A Gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP”, Speech Prosody Proc., Aix-en-Provence, France, 151-155, 2002.

[3] Fónagy I. (1983). La vive voix. Essais de psycho-phonétique, Paris, Payot.

[4] Léon, P., “Précis de phonotylistique, parole et expressivité”, Nathan, Paris, 1993.

[5] Barkhuysen, P., Krahmer, E. and Swerts, M., “Cross-modal perception of emotional speech”, ICPhS Proc, Saarbruecken, Germany, 2133-2136, 2007.

[6] Shochi, T., Aubergé, V., and Rilliard, A., “How prosodic attitudes can be false friends: Japanese vs. French social affects”, Speech Prosody Proc, Dresden, 692-696, 2006.

[7] De Meo, A. and Pettorino, M., “L’acquisizione della competenza prosodica in Italiano L2 da parte di studenti sinofoni”, in E. Bonvino and S. Rastelli [Eds], La didattica dell’Italiano a studenti cinesi e il progetto Marco Polo, Pavia University Press, 67-78, 2011.

[8] Shochi, T., Gagnié, G., Rillard, A., Erickson, D. and Aubergé, V., “Learning effect of French prosodic social affects for Japanese learners of French language”, Speech Prosody Proc, Chicago, IL, USA, paper 155, 2010

[9] Lu, Y., Aubergé, V. and Rilliard, A., “Do you hear my attitude? Prosodic perception of social affects in Mandarin”, Speech Prosody 2012 Proc., 685-688, Shanghai, China, 2012.

[10] Lu, Y., Aubergé, V. and Rilliard, A., “Tonal Influences on the prosodic Cross-linguistic Perception of Mandarin Social Affects by French and Vietnamese listeners”, the Third International Symposium of Tonal Aspect of Language Proc., Nanjing, China, 2012.

Coordination between gesture and prosody in two versions of the "Great Gatsby: 1974, 2013"

Nuzha Moritz[1], *Christophe Damour*[1]

[1] Départment Des Langues Etrangères Appliquées, Université de Strasbourg - France
[1] Déparement Des Arts du Spectacle, Université de Strasbourg - France
moritz@unistra.fr, christophe_damour@yahoo.fr

Abstract

The cross-disciplinary study (phonetics and film study) aims at highlighting the coordination between posture and prosody in two versions of "The Great Gatsby". The central aim of the study is to understand how prosodic variations are related to gesture in different acting schools. Formal and functional analysis of gesture and their relation to prosody, shows striking contrast between the acting styles.

Index Terms: prosody, gestures, postures, acting schools.

1. Introduction

In this paper a methodology and preliminary results of coordination between gesture and prosodic variations of two actors' productions are presented. The goal of this study is to assess the expression of emotion through speech and gestures according to two well known acting schools.

In human communication, gestures have been the subject of investigations in fields like sociology, psychology, 'natural history' tradition of interaction studies; linguistics etc. [1] describes gestures as an integral part of speech, "a close examination of the coordination of gesture with speech suggests, these two forms of expression are integrated, produced together under guidance of a single aim". Their semantic and pragmatic functions, parallel those of speech [2]. In the same way as speech, gestures are categorized in functions and forms. A number of function categories have been proposed in different domains. We have deemed adequate to apply a semantic and expressive approach instead of a structural or syntactic one, as the study deals with emotions. Our theoretical framework will be based on the semantic and expressive functional category of gestures presented by [2]: *Iconic* (the gesture gives an image of the shape or action, e.g. hands apart showing something huge), *metaphorics* (the gesture gives an image of an abstract concept, e.g. when speaking about arts), *deictics* (or pointing gestures, showing something), *emblems* (lexicalised gesture) and *effectives* (conveying emotions). Gestures have different forms, "…all gestures can be categorized according to their temporal and spatial characteristics" [3]. They can be described either as *dynamic* or *static* [3]. Forms of gesture differ from one culture to another (a head nod to say "yes" is different in Bulgaria than in most of the other European countries, they use the "no" head movement instead); they are combined most of the time with vocal cues. For example a high pitch and intensity plus hands gesticulation can represent anger. We will investigate if there is a correlation between forms of gestures and acoustic cues in the two versions of the "Great Gatsby".

2. Prosodic analysis

In the field of emotions, different prosodic cues have been used to describe emotions and to find relevant information [4], [5]. In previous studies combination of several parameters, have been studied: duration, amplitude, fundamental frequency, intensity, pitch, speech/syllable rate etc. But fundamental frequency contours are considered as the most relevant cue in perceiving emotions. In our experiment, analysis relies on fundamental frequency contours (minimum, median, maximum) as well as combined to gestures.

3. Acting schools

Emotions and attitudes in films are imparted mainly through the instructions of the film's director but also through the actor's acting styles and personality. The two versions of the "Great Gatsby", hence fore (GG1 for Jack Clayton, 1974 film and GG2 for Baz Luhrmann, 2013), reveal a visible difference in the aesthetic, acting styles and the use of voice techniques. The direction of GG1 and GG2 corresponds to different acting schools: the "American Academy of Dramatic Arts" and the Actors Studio/Method Acting respectively [6, 7, 8]. The former was founded in 1884 in New York, where working on voice was a predominant aspect of the actors' training. Actors indicate their emotions mainly through the use of voice variations, while body movements and gestures were considered as secondary. It is considered as an "intellectual" school of acting based on the text/script with a rather static attitude. The latter is a more recent school founded in 1947 and is mainly based on psycho-physiological exercises where actors are trained on how to show feelings through motion. In Method Acting School, postures and gestures are predominant; emotions are shown through the body which totally embodies the character [8, 19, 10].

4. Methods and Materials

4.1. Materials

The material analyzed is drawn from "The Great Gatsby" 1974 (GG1) scene 10 (from 1:36:07 to 1:41:09) and 2013 version (GG2) chapter 9 (from 1:30:33 to 1:37:36). The scene shows the confrontation between Tom and Jay Gatsby in the presence of Tom's wife and two friends. They were all lounging in armchairs as it was a very hot day, everybody was sweating and one could feel the tense atmosphere., This scene has been used in this study as a source for examples of different gestures and their correlation with prosody (mainly f0 variations). We have decided that it would be more interesting to concentrate on the husbands' reactions as they reveal a larger number of body movements.

4.2. Measurements

The method of analysis employed in this study is as follows: small segments of the scene were scanned in the aim of picking out postures and gestures with the corresponding speech units. Gestures and the associated pitch variation levels were annotated. A screen shot for each movement was segmented and plotted out on a chart for both versions. The movements were then labelled using the set of terms proposed by [2] for describing the functional and formal categories. In this study, we concentrated on the movements of the hands, arms head and trunk. Facial expressions were not included. The data was then labelled using Praat software [9], and an adapted human annotation scheme of lINTSINT [10]. Four tiers were labelled: orthographic, prosodic for inflection points of F0 contour. The tag assigned to each inflection point is relative to its predecessor and successor along the contour. The tag set is: M (medium), T (top), B (bottom) (for the speaker's voice range), H (higher), L (lower), U (up-step), D (down-step) and S (same). Corresponding F0 values to each inflection point is indicated in the third tier. The fourth tier shows the correlative gestures [11].

5. Results and discussion

Overall results show striking differences both in terms of acting style, gestures and emotions. In GG1 the nervous emotional tension in the confrontation scene is conveyed by the film-making techniques and the final cut rather than by the actors. Jay Gatsby (Robert Redford) was slouching in an arm-chair with a listening position, almost during the entire scene with limited gestures and using a monotonous calm voice. This static "peacock" attitude, contrasts with the husband's one, who was voluble and mobile, sitting down, standing up, pacing up and down the room, his hands in his trouser pockets, walking behind Jay Gatsby trying not to show his anger. Bruce Dern (husband) started hostility and provocation by asking: *"I'd like to know what kind of row you're trying to cause."* He was standing up facing Jay Gatsby; the utterance was accompanied by a slight trunk bend and a low pitch (110Hz) with no voice variation (27 Hz). In GG2, when Joel Edgerton asks J. Gatsby: *"What kind of a row you trying to cause in my house anyway"*, he was standing away from J. Gatsby, hands crossed, very little pitch variation (69Hz) but a F0 increase on *"are you*" which correspond to his (T at 151Hz). In the second utterance the husband was annoyed by his wife's question asking him to have some self control. In GG1 and GG2, they both replied *"self control!"* as if surprised. A difference in attitude and pitch variation between the two is obvious (see figure 1 and 2). A higher pitch level (in GG2, and larger variation (165Hz) was observed with a down step at the end of the utterance, although he had a static attitude, standing up and hands crossed. Whereas in GG1, the husband speaks with calm low voice but there is an audible step-up in the terminal pitch at "control" (205Hz) and a head nod at the same time. In a last attempt to conquer his wife's heart again, he supplicated her in a whispery charming voice "*And not that day that I picked you up in my arms and carried you...*" Here the action of *"carry"* is said with a slight rise at the beginning of the utterance, then almost the same pitch but "carry" is showed by *an* iconic gesture and a rather long pause (1.51 ms) before the end. This gives account of his despair, as if he was weighing his words and expecting a positive reply after the effect of his seductive voice. The husband in GG2 version used a very low pitch with a range of (B at 78Hz and T at 126 Hz), hands always crossed as gesture. The scene ends up by the husband provoking J. Gatsby; "*certainly not for a swindler…*" in GG1, with a left head tilt, a (B at 87Hz) and no increase in F0. J. Edgerton's reaction in GG2 *"certainly not for a common swindler like you*", his attitude was still static, but one can perceive the intensity in his voice when he called J. Gatsby *"a common swindler», but* with a small increase in F0. In this confrontation scene, we have noticed that the acting style and the use of gestures and voice variations are completely different in the two versions (see table 1). In GG1, there is a real contrast between J. Gatsby (Robert Redford's) static listening position and Bruce Dern who was mobile and demonstrating his speech with several gestures (n° 15: stop gesture, n°18 deictic/pointing…) [11], thus giving more strength to his speech. The expected fist fight and "blast" did not take place in this version. This might explain the small pitch variations in his voice (see table 2). Jay Gatsby's attitude in GG2 is totally different; Di Caprio delivered a "neurotic" version in this scene, showing clearly his irritability and highly nervous state through micro- gestures and postures, for instance: shaking his foot, leaning his head, walking forward to and backward from the window, helping himself to a drink several times, shouting etc. In the confrontation scene, Joel Edgerton, embodies the role of a deceived husband, ready to fight, although he stood stiff and edgy far away from J. Gatsby with his hands crossed for most of scene. But emotions were perceived more in his acting style when comparing pitch contours. (See table 2). As he does not show many gestures in this scene, he acts using his voice, with perceived pitch variations or even with sarcastic laughter (twice) "*you must be crazy"*. This static attitude or posture corresponds to the classical school of acting and contrasts with Di Caprio's Actors Studio/Method and its dynamic approach. In GG1, Bruce Dern's acting style and emotions, are conveyed more through gestures (with clear trajectories), than through the use of vocal cues. He embodied the role according to the Modern Actors method school.

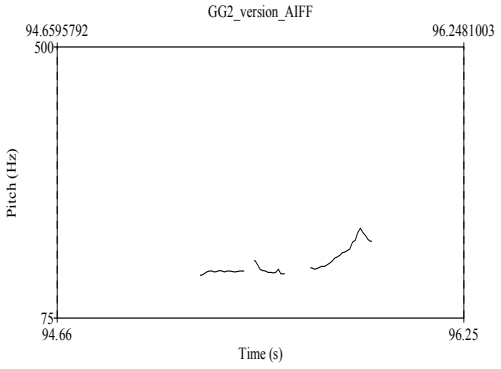


Figure 1: *"Self-control!" GG1*

The coordination between gesture and prosody is different in the two versions. In GG1, Bruce Den's gestures are combined with slight changes in pitch; however F0 values do not show systematic marked cues. His gestures sometimes correspond to a Top level like in GG1 *"self control*", to a rise *"not that day"* or even to a Bottom value *"like you"*. In GG2 the use of voice is predominant, with pitch variations reaching the Top range of the speaker but combined with static gestures.

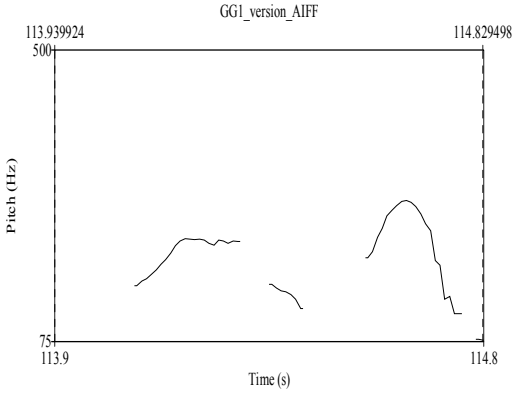


Figure 2: *"Self control!" GG2*

GG1/ Gesture	Dynamic/ Static	GG2	Dynamic/ Static
1. Trunk forward	+	Hands crossed	-
2. Head nod	+	Hands crossed	-
3. Trunk forward	+	Hands crossed	-
4. Hands lifted *Iconic* "carry"	+	Hands crossed	-

Table 1. Spatiotemporal components of gestures (+ for dynamic, - for static)

Utterances/ GG1	F0 values Hz Minimum/ Maximum	Utterances/ GG2	F0 values Minimum Maximum
I'd like	115	…are you	151 (T)
…you're	88	…cause	82 (B)
Self	147	… Control	280 (T)
control	215	… control	215 (B)
…and not	138	… Bowel	127 (T)
... my arms	108	… your	78 (B)
...common	87 (B)	Certainly	149 (T)
..a ring	127 (T)	Like you	77 (B)

Table 2. F0 variations (maximum: Top, minimum: Bottom)

6. Conclusion

As can be seen from the results the coordination between postures and voice variations in acting engenders the film's atmosphere in general and brings a significant difference in the actor's role. Pitch variations associated to postures allowed Di Caprio according to the "Acting Method" to personify a "neurotic" character in the confrontation scene, thus bringing out a more energetic acting style and a greater impact on the audience. The outcome of Joel Edgerton's performance (from the same "Acting method") is surprisingly "classical", using more voice variations with limited and static postures to embody the role of a deceived husband. Alternation enabled him to convey his emotions passing from provocation, supplication, seduction, and denigration which led to the final blast of anger. The tension in the confrontation scene in GG1 is realized by film-making techniques such as: framing, cutting, silence… a "Koulechov effect" where editing the different film plans created an atmosphere of tension, which was not really conveyed by the actors' performance. However the coordination between postures and pitch variations increased the dynamics to the interactions in the confrontation scene. The coordination between gestures and prosody seems to be clear in differentiating the two acting schools, although no recurrent acoustic cues were found in this study. In this new field of research; implementation of more methodological tools will broaden the scope of the coordination between prosody and gestures in film studies.

7. References

[1] Kendon, A., "Some relationships between body motion and speech." In A. W. Seigman and B. Pope, eds., Studies in Dyadic Communication. Elmsford, NY: Pergamon Press, 177-210, 1972

[2] McNeill, D.," Hand and Mind: What Gestures Reveal about Thought". Chicago, London: University of Chicago Press, 1992

[3] Gibbon, D. "Gesture Theory is linguistics: On Modelling Multimodality as Prosody", 23rd pacific Asian conference On languages, Information and computation, 9-18, 2009

[4] Campbell, N? "Perception of affect in speech— Toward an Automatic processing of paralinguistic information in Spoken conversation" in Proc. *ICSLP*, Jeju, Korea, 881-884, 2004

[5] Roach, P. "techniques for the phonetic description for Emotional speech", school of Linguistics and applied Language studies, university of Reading, U.K., 2000

[6] Naremore, J., "Acting in the cinema", Berkeley, California University Press, 1988.

[7] Damour, C., « L'Actors Studio : une révolution stylistique ? » In Gaffez, F., Amiens International Film Festival, 2009

[8] Cieutat, M.. and Viviani C., "Pacino/ De Niro, « Regards croisés", Paris, Nouveau Monde, First Edition, 2000.

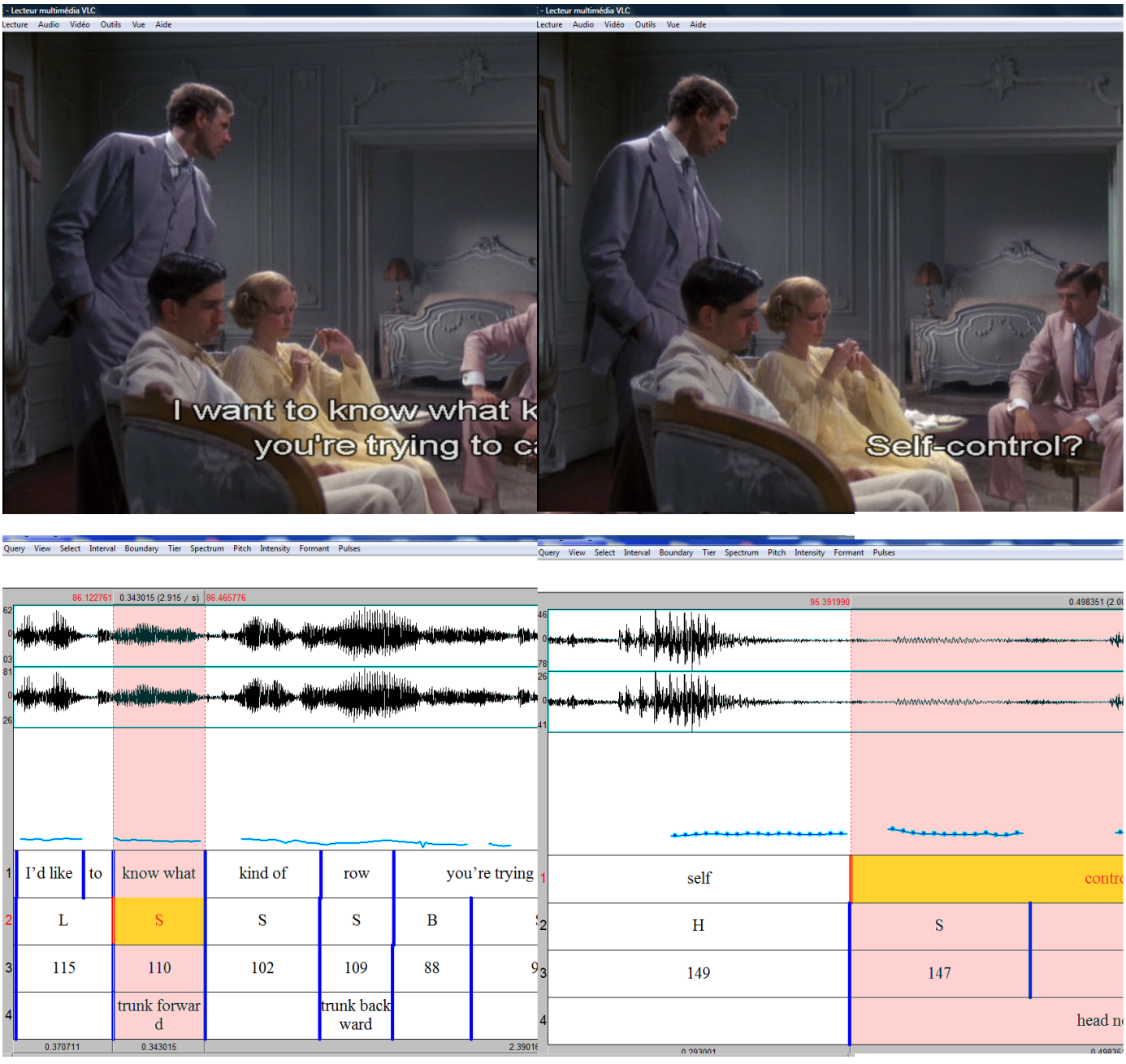
[9] Boersma, P. and Weenink, D., "Praat: Doing phonetics by Computer". http://www.praat.org, 2008

[10] Hirst, D. and DI Cristo, A, « Intonation System » Cambridge University Press, 1998

Gatsby (2013).VOB

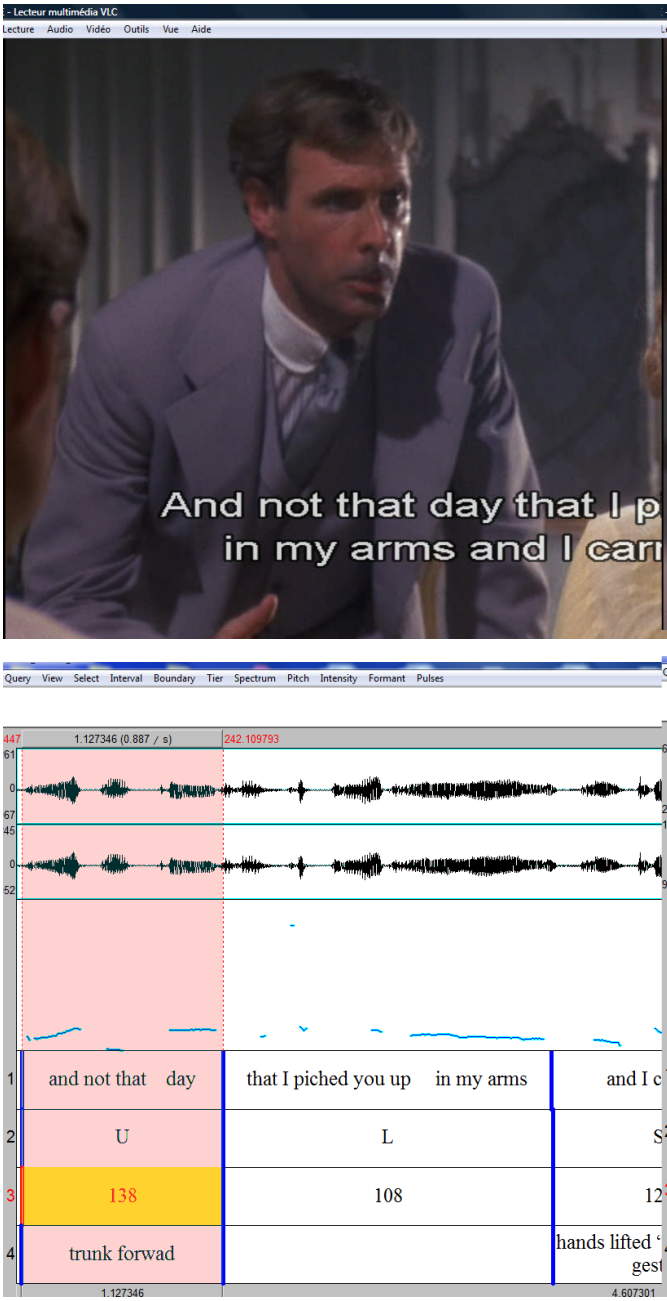
Gatsby (1974).VOB

[11] **“Great Gatsby” 1974**

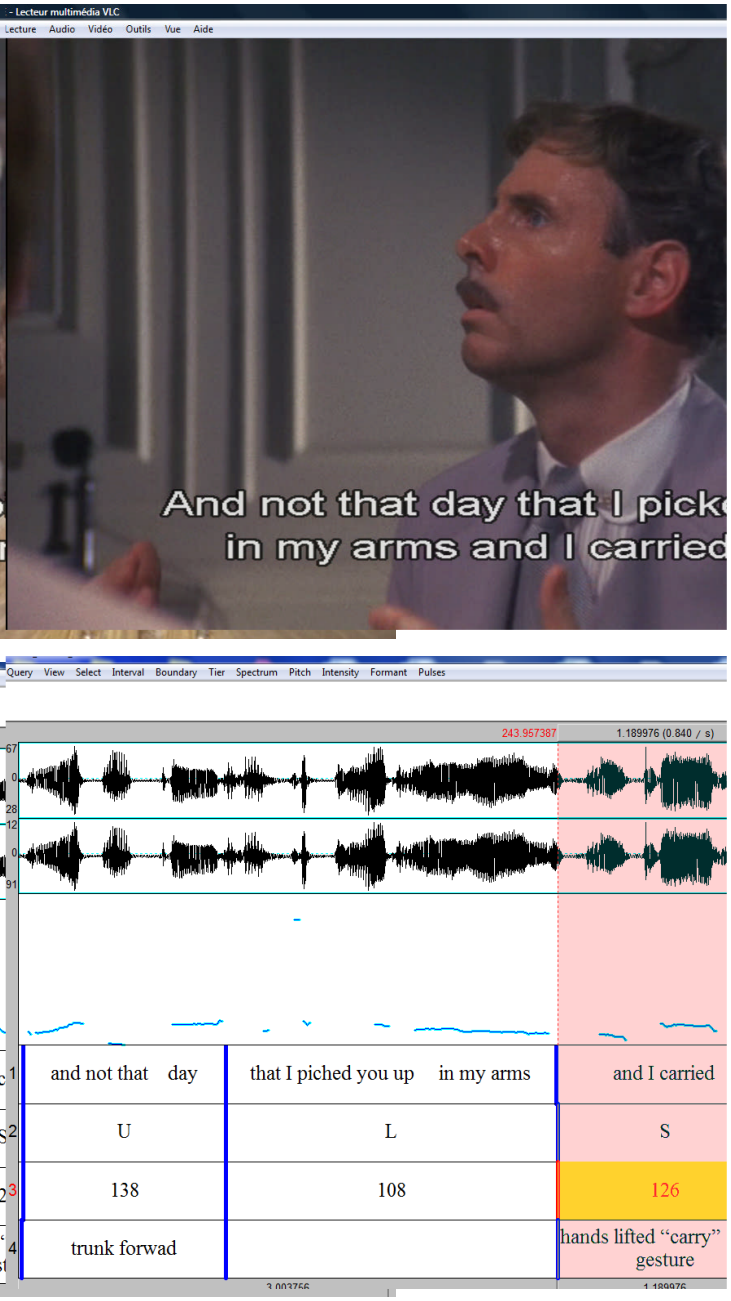


1. Husband: **trunk forward**

2. Husband: **head nod**

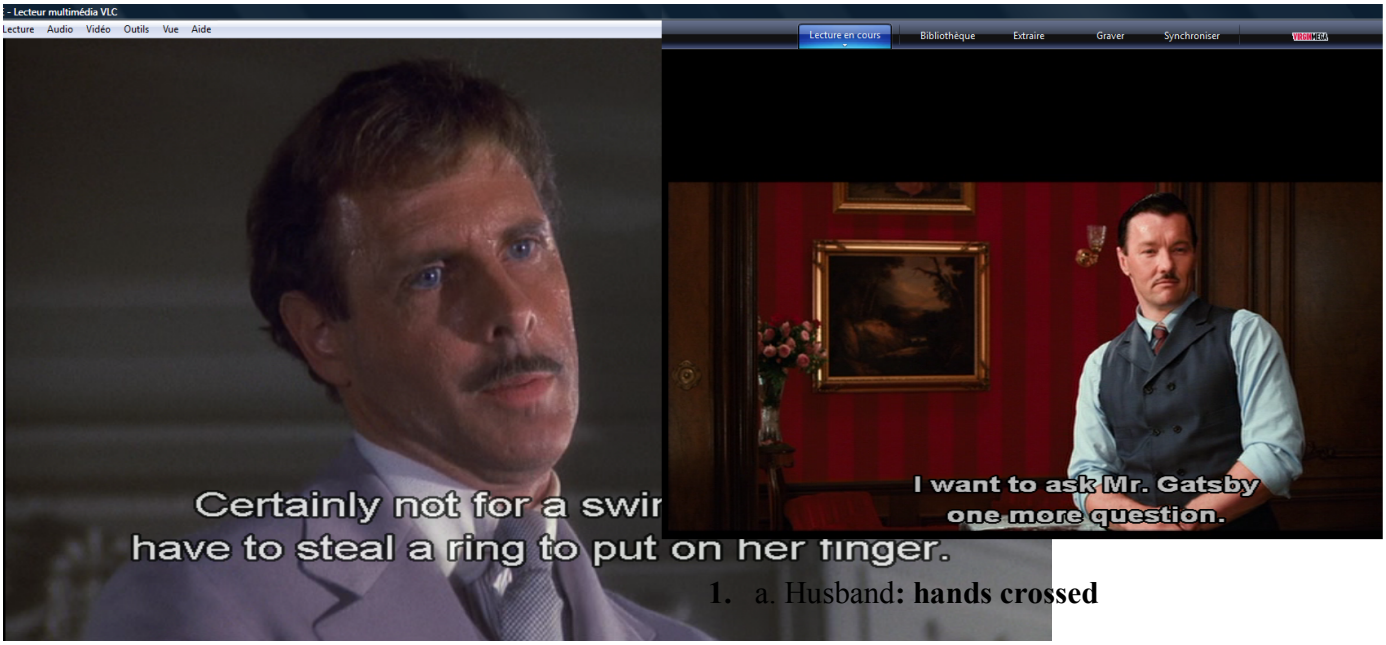


3. **a**. Husband: **bends trunk**, staring at his wife

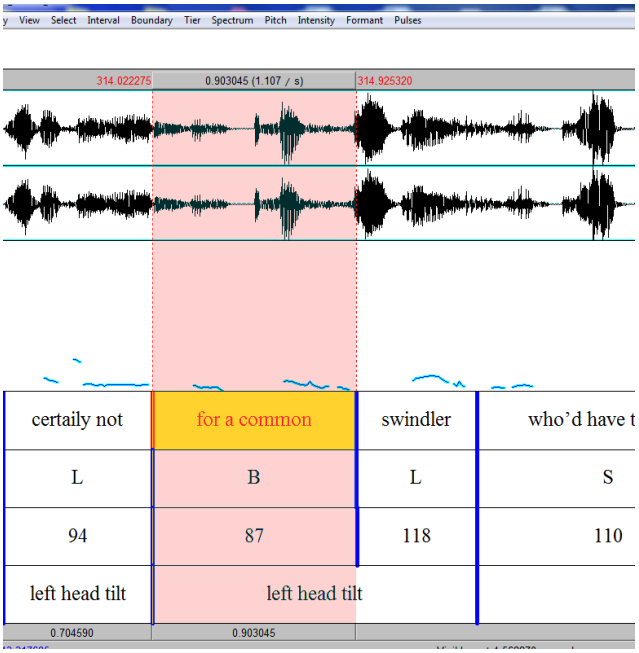


3**. b**. Husband: **hands lifted up** showing the **"carry" gesture**

“Great Gatsby” 2013

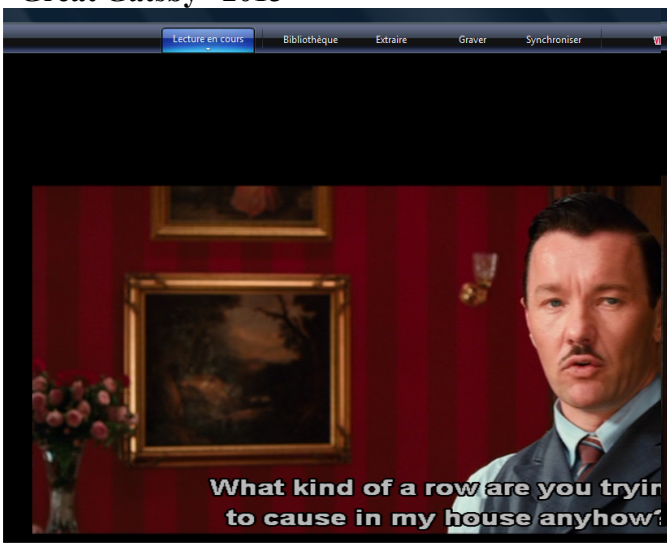


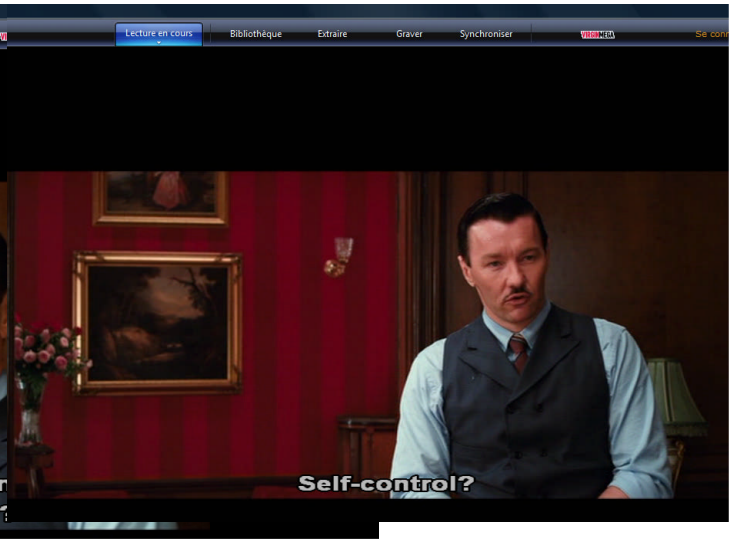
1. a. Husband**: hands crossed**

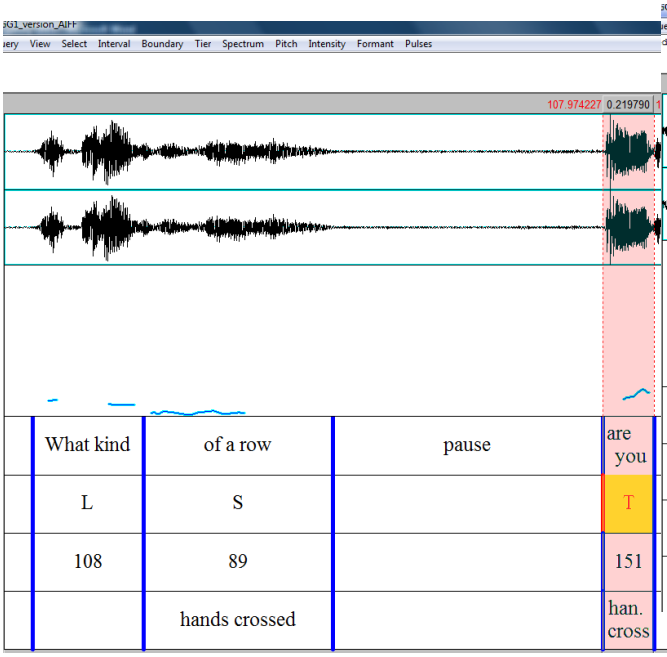


4. Husband: **left head tilt**

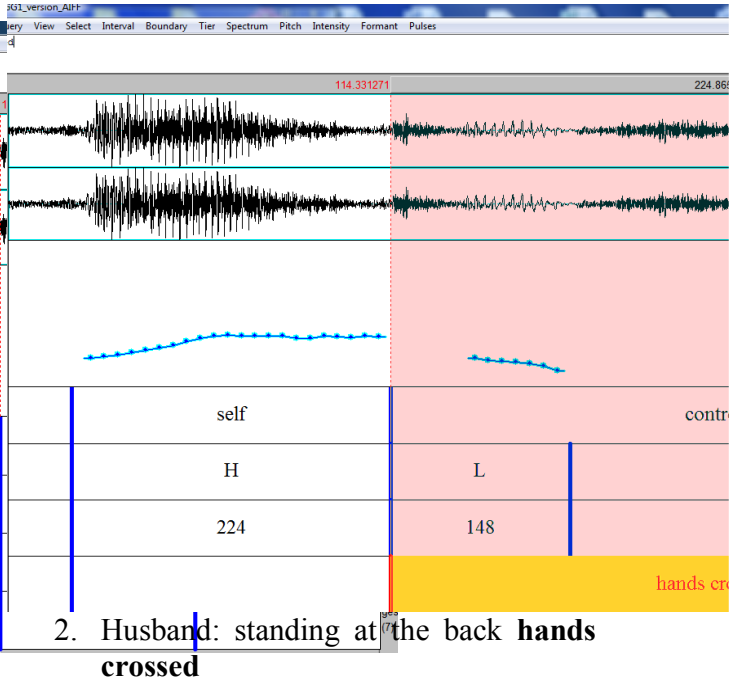
“Great Gatsby” 2013



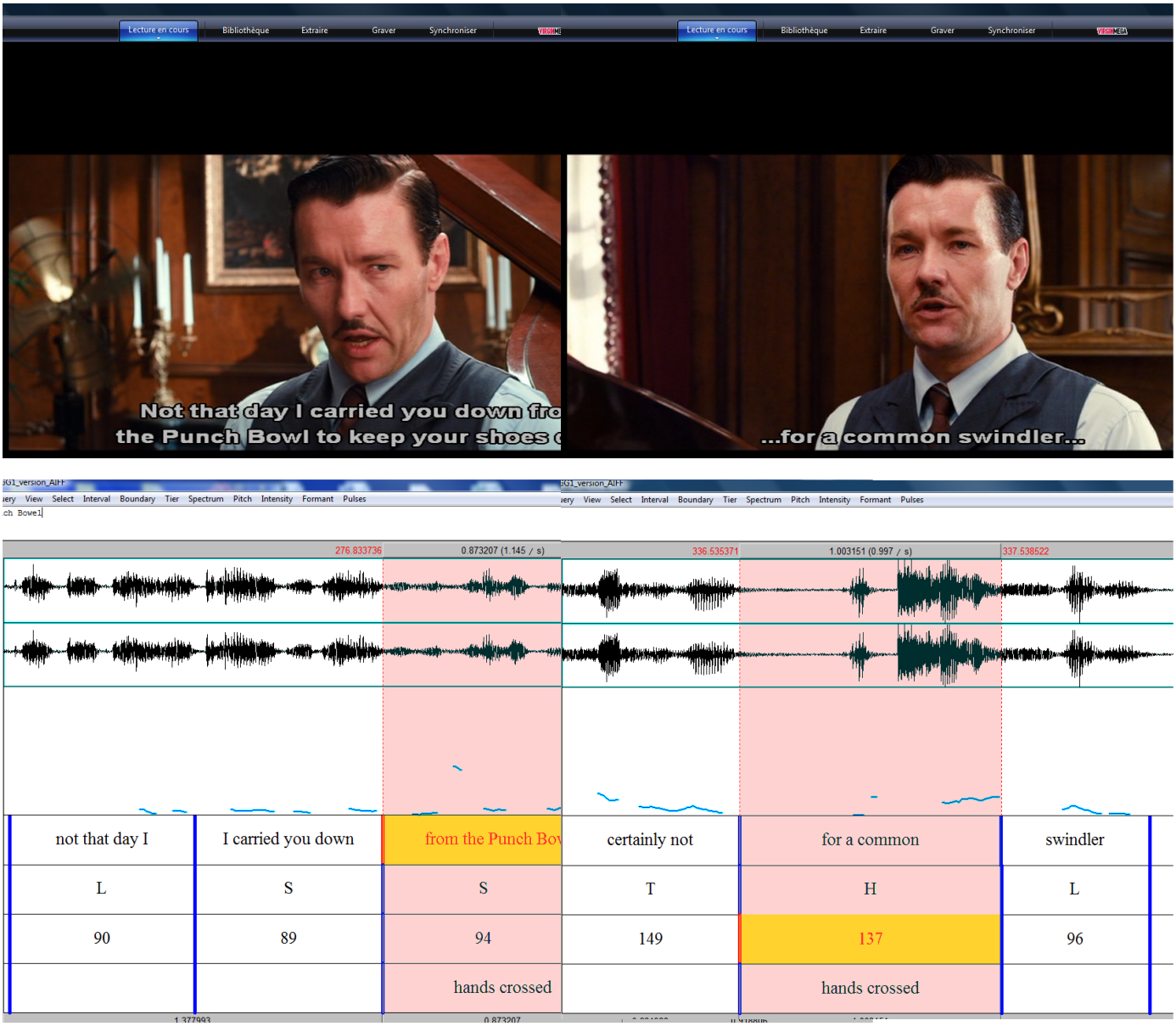




1. **b**. Husband: standing at the back **hands crossed!**



2. Husband: standing at the back **hands crossed**



3. Husband: moves one step forward, (**hands still crossed)**

4. Husband: Standing straight up (**hands still crossed**)

7 Tuesday 3

Explorations in the prosodic characteristics of synchronous speech, with specific reference to the roles of words and stresses

Fred Cummins[1], *Judit Varga*[2]

[1]UCD School of Computer Science and Informatics, University College Dublin
[2]School of Psychology, Trinity College Dublin
fred.cummins@ucd.ie,vargajudit89@gmail.com

Abstract

We examine the prosodic characteristics of read speech produced alone or in synchrony with a co-speaker in English. Previous work has demonstrated a marked difference between these two speaking conditions in Mandarin, but not English. We employ word lists that are either simple sequences of trochees, or complex lists with regular stress alternation but irregular word boundaries. Inter-onset intervals are examined and no major differences between solo and synchronous interval sequences are found. Viewed from the perspective of two generative models, however, there is weak evidence for some small difference in the dependence of interval duration on serial position.

Index Terms: synchronous speech, stress timing, word lists, joint speech

1. Introduction

Synchronous speech is a laboratory variant of the more general phenomenon of joint speech, where speakers utter the same words in unison [1, 2]. Familiar ethological examples include collective prayer, and the chant of protesters. In a synchronous speaking task, novel texts are typically employed, and subjects do not have difficulty in reading these while keeping in time with one another. The speech so produced is perhaps best characterized as unmarked. To conform to the task demands, speakers must shear their speech of unpredictable temporal variation, stemming, for example, from dramatic expression, or idisyncratic phrasing. This results in speech in which linguistic contrasts are preserved, but many sources of variability that serve to make phonetic analysis both rich and complex are otherwise absent. Synchronization among speakers has since been employed by several researchers as a means for eliciting speech suited to phonetic analysis [3, 4, 5, 6]. Most such work has been done on English, although O'Dell and colleagues (2010) have employed the same methods in studying Finnish speech rhythm and Kim and Nam (2008) examined Mandarin Chinese.

Underlying the use of synchronous speaking to elicit phonetic data is the strong assumption that there is no additional source of alteration to the speech introduced by the device of having speakers synchronize. It is known that synchronous speech tends to be relatively slow in rate, though within the range of rates adopted by speakers when speaking alone [1]. At issue is rather whether there is any form of systematic prosodic alteration to synchronous speech other than a relatively slow speaking rate. To date, this has been tested only informally, by listening to synchronous speech and noting that indeed, it sounds like unremarkable English speech.

In a recent study comparing English and in Mandarin Chinese, we observed that Chinese synchronous speech appeared to exhibit an exaggerated syllable timing compared with speech produced by one person at a time (hereafter, "solo speech") [7, Sample recordings available in Supplementary Materials online]. Sentences were produced almost as if they were lists of unconnected words, and this was evidenced by a slight difference in PVI calculated based on syllable onsets. We hypothesize that there is an interaction between the means that best satisfy the demands of synchronization, and the phonological structures of the language, such that the relatively simple syllable and word forms of Chinese lend themselves to regular, temporally predictable production in a way that the more complex phonological structures of English do not. Additional evidence for this hypothesis was found in that synchronization among Chinese speakers was more resistent to perturbation induced by having slightly mis-matched texts than their English counterparts, for whom such intervention frequently led to complete cessation of speaking [7]. A follow up study including a wider variety of languages is currently underway.

We here return to the question of whether English speech produced synchronously is, in fact, unaltered, compared to solo speech. Several consideration reveal this question to be more complex that it appears at first blush. Solo speech, and joint speech, are each produced with great amounts of variability due to context, purpose, and the identity and concerns of the speakers. Neither variety admits of reduction to a simple unmarked form that can stand for all others. Synchronous speech, more narrowly circumscribed, may, indeed appear as a largely invariant speaking style, as novel texts are used that are divorced from any ongoing behavioral context, and the additional constraint of remaining in synchrony prevents the overlay of any overly dramatic or expressive phrasing. With what should it then be compared? What kind of (solo) speech might serve as a gold standard? Put like this, the question is clearly unanswerable. However if we limit the domain of possible texts radically, we may make some meaningful comparisons between the two styles. This is the approach adopted here, where we use simple word lists of 8 unconnected words.

By using simple word lists, we make use of texts that admit of very little variability when read either alone or together. This serves to constrain the potential variability of the solo speech. But using invariant word lists represents a drastic simplification. In order to then re-admit some potential temporal complexity, we contrast simple word lists (8 trochees) with complex lists in which the relative sequence of stresses and word onsets is non-coincident, such that either word onsets or stress onsets may form the basis of regular timing, but not both. If synchronization is facilitated by enhancing an underlying regularity (as we suspect in Chinese) then we might observe a more

regular sequence of either word onsets or stressed syllable onsets in synchronous productions of complex lists compared to solo productions.

The use of word lists allows us to also evaluate the temporal characteristics of the speech in terms of models of sequential production. Two highly influential models are the Wing and Kristofferson model of interval timing [8] and the hierarchical timing model of Rosenbaum [9]. The Wing & Kristofferson model has found frequent application in teasing apart sources of variability in tapping studies [10, 11, and many others]. It assumes that overt behavior is shaped by a distinct timing process (the central clock) which influences the movement of effectors. Both clock and peripheral physiology are assumed to be potential sources of variability in observed behavior, but the assumptions of the model permit decomposition of that variability into distinct clock and peripheral sources, which may be subject to distinct pathologies, or independent perturbation. The model rests on the assumption that the two sources are independent, and this assumption is warranted only if the lag one autocorrelation of an interval sequence lies between zero and -0.5 (for full justification, see Wing and Kristofferson, 1973). We test this below.

Rosenbaum's model looks for evidence of hierarchical structure in short regular sequences that would suggest constraints on timing that are not strictly sequential (as in Wing & Kristofferson's model). Hierarchical execution of movement plans involves the depth-first traversal of a tree, and would lead to a dependence of interval duration on serial position within sequence. For short sequences of 8 taps, this leads to alternating short-long patterning from binary grouping at the lowest level, with additional short-long alternation at higher levels of composite units of two or four taps. As our word lists consist of short sequences of 8 accents or stresses, we ought to look for any sign of non-sequential, hierarchical influences that would be indicated by a non-monotonic dependence of interval duration on serial position. Such non-sequential effects would be compatible with some form of hierarchical production model, while a specific pattern in which interval lengths are ordered as $\{4\} > \{2,6\} > \{1,3,5,7\}$ would fit the specific form of the model adduced on the basis of sequences of 8 taps in the 1983 paper.

2. Methods

25 dyads (50 speakers) took part in the study. Speakers were either relative strangers (12 dyads) or were highly familiar couples (13 dyads). All familiar dyads were of mixed sex, while among the strangers, 7 were mixed, 2 were male-male and 3 were female-female. Ages ranged from 21 to 56, and all were native speakers of Hiberno-English. Subjects were recruited on the campuses of two Dublin universities, and ethical approval was provided by the School of Psychology at Trinity College Dublin.

Subjects were recorded as part of a larger data gathering exercise. Relevant to the present study, they each read 4 word lists alone ("solo") and 4 word lists together ("sync"). Each group of 4 comprised 2 simple trochaic lists and 2 complex lists (see below). Solo readings were done before synchronous in all cases. Readings were done using head mounted microphones, and recordings were made to parallel audio channels.

Half the wordlists were simple, in which case they consisted of 8 trochees, selected so that stressed syllable onsets were of simple CV form, with C $\in$ {b,d,g}. Sample simple list: *banter, body, dagger, guinness, batty, dancer, bingo, gutter*. The other half were complex. In complex lists, there was a regular alternation of stressed and unstressed syllables (half began with a stressed, half with an unstressed element), but word boundaries were selected to be irregular, due to varying numbers of syllables per word. Sample complex list with a weak initial syllable: *deny, debunking, boot, divide, deduction, bike, barbaric, ban*, with a strong initial syllable: *bad, debugging, body, boot, debacle, banter, bog, degrading*. Both word onset and stressed syllable onsets were constrained to be of CV form with C $\in$ {b,d,g,v,n,m,l}. Each list in each condition was unique and was spoken exactly once. A total of 11 lists were discarded due to speech errors or mispronunciation.

Word onsets and stressed syllable onsets were located in time using a P-centre estimation algorithm, first presented in Cummins and Port (1998). This identifies the time of an onset as the halfway point through a local rise in the amplitude envelope of the bandpass filtered signal (with cut offs at 500 and 2000 Hz). This algorithm works well when syllable onsets are suitably constrained, as here, and in a very few cases where no appropriate local rise was found, manual measurement was made (less than 1% of data points). This provided a sequence of 8 word onsets and 8 stressed syllable onsets for each list. For simple lists, these coincide, providing a single set of 8 onsets, or 7 successive intervals. For complex lists, these provide 2 alternate ways of looking at the list, as 7 intervals demarcated either by word onsets or by stressed syllable onsets.

3. Results

We have three kinds of interval series: trochaic, complex calculated from word onsets, and complex calculated from stressed syllable onsets. We first examine the distribution of interval durations in the solo and synchronous conditions, for each kind of series. Fig. 1 provides an overview of the distribution of interval durations as a function of serial position. The first thing to note is that there is no obvious macroscopic difference in the central values or serial characteristics of the synchronous and the solo data. As has been documented before, the variability across speakers seems to be reduced in the synchronous case [13]. Fig. 2 shows the distribution of the interval durations for each series in each condition. It is clear that the solo distributions are right skewed, while the synchronous distributions are more compact and more nearly symmetric. The reduction in variance is confirmed by one-sided F-tests which verify reduced variability in the synchronous condition for the trochees (F(643,643)=2.7, $p < .001$), the word onsets in complex lists (F(643,587)=1.9, $p < .001$) and stress syllable onsets in complex lists (F(615,587)=2.8, $p < .001$).

The complex lists were designed so that word onsets and stress syllable onsets did not coincide for every word. In order to see whether subjects imposed regularity on the word or stress onsets, we calculated the normalized Pairwise Variability Index for successive interval durations within each series. This provides a measure of variation among successive intervals, and is minimized in isochronous series, and maximized in alternating long-short series. The nPIV is calculated as

$$\mathrm{nPVI} = 100 \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \quad (1)$$

Fig. 3 shows the nPVI scores for each series type and condition. In both solo and synchronous conditions, the intervals formed by successive word onsets in the complex lists are con-

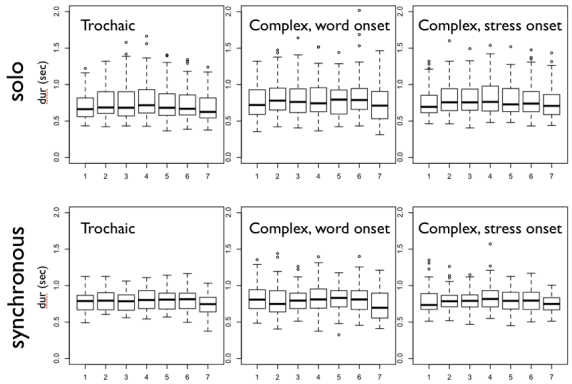


Figure 1: *Interval duration as a function of serial position.*

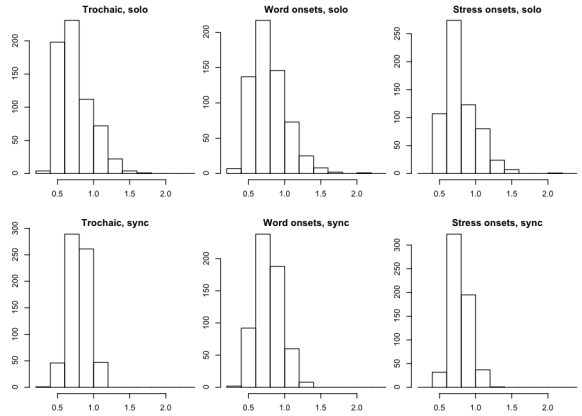


Figure 2: *Histograms of interval durations.*

siderably less regular (higher nPVI) than those formed by the stress syllable onsets. There is no evidence here of any increase in regularity (lower nPVI scores) in the synchronous condition compared with the solo.

We now turn to analyses that examine the assumptions and predictions of the two generative models, the hierarchical production model of Rosenbaum [9], and the clock model of Wing and Kristofferson [8]. The hierarchical production model makes the general prediction that serial position will influence duration, and, for sequences of 8 events, or 7 intervals, it predicts the relative durations based on tree-traversal distance of a hypothetical underlying metrical tree.

In order to investigate the effect of serial position on interval duration in the solo lists, a repeated-measures ANOVA was carried for each of the three series with familiarity (two levels) as a between subjects factor, and serial position and repetition as within subject factors (7 and 2 levels, respectively). For the trochees, there was a main effect of serial position ($F(6,264)=6.6$, $p < .001$) and a main effect of repetition ($F(1,44)=4.2$, $p < .05$). For the word onsets in complex lists there was only a main effect of serial position ($F(6,252)=3.8$, $p < .01$), and similarly for the stressed syllable onsets in complex lists there was only a main effect of serial position ($F(6,252)=3.7$, $p < .01$). A similar analysis for the synchronous lists is complicated by the fact that the utterances of two speakers speaking in unison can not be considered at all indepen-

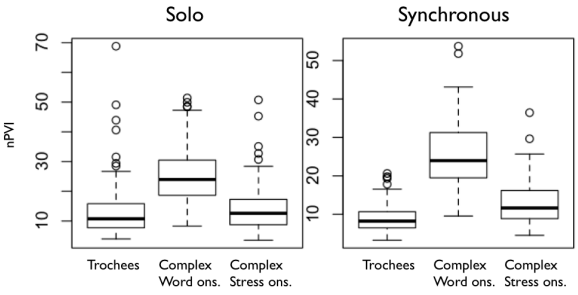


Figure 3: *nPVI scores for each series type and condition.*

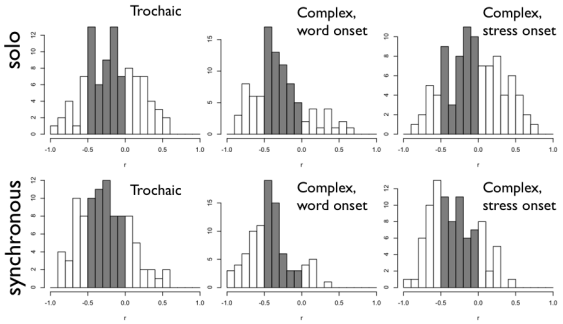


Figure 4: *Histograms of lag 1 autocorrelations.*

dent of one another. We therefore conducted a similar analysis for the first and second speaker separately, and report only those findings that were consistent across both speakers. There was a main effect of serial position only for the trochaic series ($p < .001$), but not for either the word onsets or the stressed syllable onsets.

There are thus effects of serial position in the solo data, but not in the synchronous data, except for the maximally regular trochaic series. For the solo data, this is in accordance with the general premise of the model of Rosenbaum et al., but does not test the predictions of the specific form of that model described in Rosenbaum et al. (1983), which predicts that interval 4 should be the longest, followed by intervals 2 and 6, with intervals 1, 3, 5 and 7 somewhat shorter. This structure is not evident in the aggregate data, and there is not enough data per subject or dyad to test this on a by-subject/by-dyad basis. However, simple calculation of the index of the longest interval in each series does not support the Rosenbaum predictions, as Interval 4 is the longest interval in no more than 30% of series in the trochaic solo series, and less in all other cases.

Turning now to the Wing and Kristofferson model, it is predicated upon an assumption of the separability of variance in timing due to two independent sources: a central clock or timekeeper, and a peripheral effector system. This hypothetical independence is possible only if the lag 1 autocorrelations within each series fall in the range [-0.5, 0]. Lag 1 autocorrelations outside that range violate the assumptions of the model. In the many studies that have employed this model, the proportion of the data violating this constraint, and hence the validity of the model, has varied greatly form case to case.

Fig. 4 shows histograms of the lag 1 autocorrelations for each series. Only those series falling within the grey bins ac-

cord with the basic assumptions of the Wing and Kristofferson model. The proportion of series that violate those assumptions ranges from 41% to 55%, and violations are found both above and below the bounds of [-0.5, 0]. The serial production model of Wing and Kristofferson is thus not appropriate for interpreting these data. Violations are as common in the solo data as they are in the synchronous case.

4. Discussion

The principal question addressed in this small study is whether systematic differences between speech produced alone or in synchrony with a co-speaker can be found in English. Such differences have been found in Chinese, where synchronous speech has been found to be more list-like, with regularization of inter-syllable onset intervals [7]. In order to constrain the kind of prosodic change that might be observed, lists were employed, both simple trochaic sequences, and complex sequences in which either word onsets or stress syllable onsets might for the basis for any hypothetical regularization during synchronous production.

One difference we expected to find, and did, is that interval timing is less variable between subjects in the synchronous condition. This difference does not imply any systematic change to the prosodic features of any given utterance, however.

At first blush, there appears to be little in the way of variation in the prosodic features of synchronous utterances compared to solo utterances. In both cases, trochees are produced with a low pairwise variability, and the complex lists are produced with similar regularity in the sequence of stressed syllable onsets (but not word onsets). This is itself valuable empirical evidence that stress feet, in the sense of Abercrombie,[1] and not lexical units, constitute temporally extended constituents that may, under specific conditions, form the basis of isochronous sequences in English speech. This is in line with observations from the speech cycling paradigm in English [12], and contrasts with observations made under similar constraints in Korean [14] and Japanese [15].

We then examined the appropriateness of two generative models of sequential interval production that might be called to task in accounting for these data. The simpler model of simple sequential production based on the hypothesis of a modality-independent timer is provided by Wing and Kristofferson (1973). The presuppositions of that model were found to be very inappropriate for the present data set, with almost half of all observations showing non-local serial dependencies that violate the assumptions of the model. This was true for solo and synchronous data in equal measure.

When we viewed the data through the lens of the hierarchical production model of Rosenbaum, however, we observed some possible differences between solo and synchronous interval sequences. A liberal application of this model, that captures the notion of hierarchical production, but remains agnostic about the precise form of control or implementation underlying such production, predicts only that there will be position dependent differences among interval durations that are not a simple function of the immediately preceding neighbor (as in Wing and Kristofferson). This was true of both simple and complex lists in the solo condition, but only true of the simpler trochaic lists in the synchronous condition. The more rigorous prediction that falls out of positing a specific binary tree underlying production, that accounted very well for tapping data in the original study, did not well describe the present data.

Given the failure of the more specific model to capture for form of the non-local dependency of interval duration on serial position, we must be cautious in drawing strong conclusions. The difference betwen solo and synchronous data we observed is slight, and not yet well characterized. Prosodic differences between solo and synchronous speech in English are slight, if they exist at all, and at this stage, it appears that synchronous speech may still be a valuable way of eliminating variability from English speech while leaving linguistic contrasts unaffected. On the basis of our experinece with Mandarin, however, it is clear that this assumption does not generalize to all other languages, and that even in the case of English, further investigation is warranted.

[1] The Abercrombian stress foot is the interval from one stressed syllable onset to the next, including any and all intervening unstressed syllables. It assumes a simple binary stress distinction that may or may not be an appropriate characterization of English stress, and that certainly does not generalize to all languages.

5. References

[1] F. Cummins, "On synchronous speech," *Acoustic Research Letters Online*, vol. 3, no. 1, pp. 7–11, 2002. [Online]. Available: http://ojps.aip.org/ARLO

[2] ——, "Practice and performance in speech produced synchronously," *Journal of Phonetics*, vol. 31, no. 2, pp. 139–148, 2003.

[3] J. Krivokapić, "Prosodic planning: Effects of phrasal length and complexity on pause duration," *Journal of Phonetics*, vol. 35, no. 2, pp. 162–179, 2007.

[4] M. Kim and H. Nam, "Synchronous speech and speech rate," *Journal of the Acoustical Society of America*, vol. 125, no. 5, p. 3736, 2008.

[5] M. A. Poore and S. Hargus-Ferguson, "Methodological variables in choral reading." *Clinical Linguistics and Phonetics*, vol. 22, no. 1, pp. 13–24, January 2008.

[6] M. L. O'Dell, T. Nieminen, and L. Mustanoja, "Assessing rhythmic differences with synchronous speech," in *Speech Prosody 2010 Conference Proceedings*, vol. 100141, 2010, pp. 1–4.

[7] F. Cummins, C. Li, and B. Wang, "Coupling among speakers during synchronous speaking in English and Mandarin," *Journal of Phonetics*, vol. 41, no. 6, pp. 432–441, November 2013.

[8] A. M. Wing and A. B. Kristofferson, "Response delays and the timing of discrete motor responses," *Perception and Psychophysics*, vol. 14, no. 1, pp. 5–12, 1973.

[9] D. A. Rosenbaum, S. B. Kenny, and M. A. Derr, "Hierarchical control of rapid movement sequences," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, no. 1, pp. 86–102, 1983.

[10] J. Gibbon, R. M. Church, and W. H. Meck, "Scalar timing in memory," *Annals of the New York Academy of sciences*, vol. 423, no. 1, pp. 52–77, 1984.

[11] S. M. Rao, D. L. Harrington, K. Y. Haaland, J. A. Bobholz, R. W. Cox, and J. R. Binder, "Distributed neural systems underlying the timing of movements," *The Journal of Neuroscience*, vol. 17, no. 14, pp. 5528–5535, 1997.

[12] F. Cummins and R. F. Port, "Rhythmic constraints on stress timing in English," *Journal of Phonetics*, vol. 26, no. 2, pp. 145–171, 1998.

[13] F. Cummins, "Synchronization among speakers reduces macroscopic temporal variability," in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 2004, pp. 304–309.

[14] Y. Chung and A. Arvaniti, "Speech rhythm in Korean: Experiments in speech cycling," in *Proceedings of Meetings on Acoustics*, vol. 19, 2013, p. 060216.

[15] K. Tajima and R. F. Port, "Speech rhythm in English and Japanese," *Phonetic interpretation: Papers in laboratory phonology VI*, pp. 317–334, 2003.

Effects of native dialect on Mandarin listeners' use of prosodic cues to English stress

Zhen Qin, Annie Tremblay

Department of Linguistics, University of Kansas, United States

qinzhenquentin2@ku.edu, atrembla@ku.edu

Abstract

This study investigates the effect of native dialect on the use of prosodic cues to English stress by Standard Mandarin (SM) listeners, Taiwanese Mandarin (TM) listeners, and English listeners. Both SM and TM use fundamental frequency (F0) to realize lexical tones, but only SM uses duration together with F0 to realize lexically contrastive full-full vs. full-reduced stress patterns. Native English listeners and second language learners of English who spoke SM or TM as native language and were at similar proficiencies in English completed a sequence-recall task. English disyllabic non-words that differed in stress placement were resynthesized to contain only F0 cues, only duration cues, or converging F0 and duration cues. The results showed that SM-speaking learners used duration more than TM-speaking learners to recall English non-words. Native dialect is suggested to be considered in second language speech processing models.

Index Terms: lexical stress, prosodic cues, native dialect.

1. Introduction

Languages differ in the extent to which prosodic cues contribute relevant information for lexical access. Second-language (L2) learners weigh prosodic cues to stress as a function of how these cues are used in the native language (L1) (for discussion, see [1]). The present study investigates how the L1 dialect influences Mandarin-speaking L2 learners' use of prosodic cues in the processing of English stress.

In English, although prosodic cues such as fundamental frequency (F0), intensity, and duration signal stressed syllables in accented words [2,3,4], listeners make limited use of these cues, as English stress is also signaled by segmental information (e.g., full vs. reduced vowels; for discussion, see [1]). Although English listeners can use prosodic cues to English stress [5,6], their word recognition is not inhibited by errors in stress placement unless these errors also result in segmental (i.e., vowel quality) changes in the word [7,8,9,10,11].

Prosodic cues have a greater functional load in Mandarin than in English, in part because Mandarin is a tonal language in which lexical identity is signaled primarily by F0 cues (e.g., [mā] 'mother,' [má] 'hemp,' [mǎ] 'horse,' and [mà] 'scold') [12,13]. While all dialects of Mandarin share this tonal system, some dialects differ in whether or not they also have lexical stress. For example, Standard Mandarin (SM) has a stress pattern that contrasts full-full and full-reduced disyllabic words (e.g., respectively, ***dongxi*** 'west and east' vs. ***dong****xi* 'stuff') [14,15]. Whereas words with the full-full pattern are stressed on both syllables, those with the full-reduced pattern are stressed only on the first syllable (word-initial/trochaic stress). The reduced vowel in the latter carries a neutral tone rather than a lexical tone and it is shorter than the full vowel [14,15]. These two stress patterns are thus signaled with duration and F0, with these two cues perhaps contributing to lexical access in SM [16,17]. By contrast, Taiwan Mandarin (TM) does not have this stress distinction: Words have the full-full pattern, and the second syllable of disyllabic words is not reduced (e.g., ***dongxi*** 'west and east' or 'stuff') [13]. Given the absence of the full-reduced stress pattern in TM, F0 may play a more important role than duration for lexical access in TM [18].

Research on Mandarin listeners' use of prosodic cues when processing English words has yielded inconsistent findings. Some speech perception studies found that, like native English speakers, Mandarin-speaking L2 learners of English from mainland China but tested in the US relied more on segmental cues than on prosodic cues to English stress and did not differ from native English speakers in their use of F0 and duration cues to English stress [19]. Other research, however, suggests that Mandarin-speaking L2 learners of English from mainland China and tested in China could use only F0 cues to English tress [20]. Finally, and perhaps more surprisingly, it has been shown that Mandarin-speaking L2 learners of English from Taiwan but tested in the US could use duration alone in the perception of English stress [21].

These inconsistent findings may stem, at least in part, from the different proficiency of the Mandarin-speaking L2 learners of English that were tested: The L2 learners who lived in the US [19,21] are likely to be more proficient than those who lived in China [20]. Once proficiency is controlled for, it is unclear whether the L1 dialect would influence L2 learners' processing of English stress. Whereas the above studies examined L2 learners who spoke different Mandarin dialects [19,20 vs. 21], they did not investigate whether the particular properties of these dialects—specifically, whether or not the L1 dialect has lexical stress—would affect L2 learners' encoding of English stress. L1 effects on L2 learners' use of prosodic cues to stress are well documented [22, 23, 24]; much less is known about the effect of L1 dialect on non-native listeners' use of such cues in speech processing.

To fill this gap, the present study investigates the use of F0 and duration cues to English stress by English listeners and L2 learners of English who speak SM or TM as L1 dialects and are at a similar proficiency in English. For F0 cues, given that SM and TM have similar tonal systems, we predict that the SM and TM groups will not differ from each other in their use of F0 to encode stress in English words. For duration cues, since SM has a lexical stress contrast signaled in part by duration cues, we predict that SM listeners will perform better than TM listeners in the use of duration cues to English stress. Finally, because SM does not have a reduced-full stress pattern, we predict that SM listeners will differ from TM listeners only on words that are stressed on the initial syllable—not on words that are stressed on the final syllable.

2. Methods

We conducted a sequence recall task adapted from [22,23]. This paradigm requires listeners to encode phonetically variable stimuli that differ in stress placement and hold them in short-term memory for a brief amount of time. Given the

memory load it imposes, this paradigm taps into L2 learners' abstract phonological knowledge and is thus ideally suited for investigating the effect of L1 dialect on the use of prosodic cues in the encoding of L2 stress.

2.1. Participants

Participants included 15 L2 learners of English who spoke SM (mean age: 24.2; SD: 4.0) or TM (mean age: 25.7; SD: 3.4) as L1. Their results were compared to those of 15 native English listeners (mean age: 21.3; SD: 3.9). The two groups of L2 learners did not differ in their age of acquisition of English (p>.1) or in their proficiency in English (p>.1), assessed with a cloze test [25]. No participant had any speech or hearing impairments.

2.2. Stimuli

The stimuli were minimal pairs of English non-words that differed in stress placement (e.g., ['fʌði] vs. [fʌ'ði]). The non-words had a $C_1V_1C_2V_2$ structure. To exclude segmental cues of stress such as vowel reduction, /ɪ/, /ʊ/ and /ʌ/ were used in V_1 position and [i] was used in V_2 position, as they are not reduced to schwas in their respective positions in English. For C_1 and C_2, only fricatives were used to avoid providing segmental cues to stress. In total, four segmental non-words (/sɪvi/, /zʊθi/, /fʌði/, and /hʌfi/) were used. All stimuli were produced by a female native speaker of American English (Midwest). She read each non-word four times in the carrier sentence "Say _ again." Two tokens were selected from the four repetitions of each word type. The stimuli included a total 16 tokens (4 segmental words × 2 stress patterns × 2 tokens). Any given sequence (described in the next section) never contained the same token twice.

As shown in Table 1, the non-words with initial vs. final stress differed significantly in F0 (p<.01) and duration (p<.01) ratios of the first syllable (σ1) to the second syllable (σ2).

Table 1. *Acoustic characteristics of non-words with initial stress and non-words with final stress.*

	F0						Duration					
	Word-initial			Word-final			Word-initial			Word-final		
	σ1	σ2	ratio	σ1	σ2	ratio	σ1	σ2	ratio	σ1	σ2	ratio
Mean	226	162	1.4	174	192	0.9	163	293	2.2	140	328	3.0
SD	22	15	0.1	7	12	0.1	76	26	0.9	73	32	1.4

This study investigates the use of F0 and duration cues to the perception of English stress. The stimuli were resynthesized such that stress would be conveyed by both F0 and duration ("F0+duration cues" condition), by F0 alone ("F0 cues" condition), or by duration alone ("duration cues" condition). The non-words were first normalized for intensity at 70 dB. The F0 and duration cues were then manipulated in PSOLA [26] such that the non-words would have the average F0 and/or duration of the natural stimuli reported in Table 1. Thus, in the conditions where F0 conveyed stress placement, the F0 of σ1 and σ2 would be that corresponding to the averages in Table 1, and in the conditions where duration conveyed stress placement, the duration of σ1 and σ2 would be that corresponding to the averages in Table 1. In the condition where F0 was the only cue to stress placement, the duration of each syllable corresponded to the average duration across stress patterns but not across syllables (i.e., 156 ms for σ1 and 310 ms for σ2). This was done in order to preserve word-final lengthening in the words. In the condition where duration was the only cue to stress placement, the F0 of each syllable corresponded to the average F0 across both syllables and stress patterns (i.e., 189 Hz).

2.3. Procedures

The sequence-recall task had two main phases: a familiarization (or association) phase and a testing phase.

In the familiarization phase, participants were trained to associate 1 and 2 (on a keyboard) with, respectively, words with initial stress and words with final stress. This was done using real words (i.e., *trusty* vs. *trustee*). Participants received feedback on whether or not their responses were correct. After 18 trials, an association test was conducted, in which participants were required to correctly identify the stress pattern of the tokens they heard (as 1, word-initial stress, or 2, word-final stress). Only the participants who received an accuracy score equal to or higher than 95% were invited to complete the testing phase.

In the testing phase, participants were asked to recall sequences of four tokens by pressing 1 and 2 in the correct order. Each test trial began with the auditory presentation of a sequence that included two stress-initial tokens and two stress-final tokens with the same segments (e.g., [fʌ'ði] ['fʌði] ['fʌði] [fʌ'ði]). Each sequence consisted of four different tokens from the same cue condition (F0+duration cues, F0 cues, or duration cues). To shorten the duration of the experiment, three of the six possibilities of number ordering (i.e. [1122], [2211], [1212], [1221], [2121], [2112]) were used for two segmental non-words, and the other three possibilities were used for the other two segmental non-words. The order of sequences and of tokens within each sequence was randomized across participants. The experiment included 36 experimental sequences (4 segmental words × 3 conditions × 3 orderings).

The non-words in the sequences were separated by an interstimulus interval of 50 ms (see [22,23]). The last interstimulus interval in the sequence was followed by the prompt "OK" to prevent participants from using echoic memory to recall the sequences. The inter-trial interval was 1,500 ms. Participants completed a practice session of 12 real-word sequences (e.g., *trustee trusty trustee trusty*) to ensure that they would understand the paradigm before the actual experiment.

2.4. Data analysis

Logit mixed-effects models were conducted on the participants' sequence-encoding and word-encoding accuracies. The sequence-encoding accuracy was computed to examine the overall effect of prosodic cue on the processing of English stress; the word-encoding accuracy was computed to compare the processing of English stress for words with initial stress vs. words with final stress. The models were fitted in R, using the lmer( ) function from the lme4 package for mixed-effects models (for discussion, see [27]).

Model 1 analyzed the sequence-encoding accuracy of the three groups in the first three conditions, with cue condition (F0+duration cues vs. F0 cues or duration cues), L1 (English vs. SM or TM), and the interaction between the two as fixed variables. In this model, the F0+duration cues condition and the English group were used as baselines. Model 2 analyzed only the L2 learners' corresponding sequence-encoding

accuracy, with L1 dialect (SM vs. TM) instead of L1 as fixed variable. The baseline for L1 dialect was SM.

Models 3-4 analyzed the participants' word-encoding accuracy in the first three conditions but separately for words with initial vs. final stress, with cue condition (F0+duration cues vs. F0 cues or duration cues), L1 (English vs. SM or TM), and the interaction between the two as fixed variables. The baselines were the same as in Model 1. Models 5-6 analyzed only the L2 learners' word-encoding accuracy in the first three conditions separately for words with initial vs. final stress, but with L1 dialect (SM vs. TM) instead of L1 as fixed variable. The baselines were the same as in Model 2. In all the models, participant and item were crossed random variables.

If our predictions are correct, we should find significant interactions between the effect of duration cues (as compared to the F0+duration baseline) and the effects of L1 and L1 dialect in the sequence-encoding results, with TM listeners showing a larger difference between the duration cues condition and the baseline condition than the English and SM listeners. In the word-encoding results, we should find similar results only for non-words with initial stress.

3. Results

3.1. Sequence-encoding accuracy

The accuracy of the sequence encoding in the three cue conditions is shown for the three groups in Figure 1.

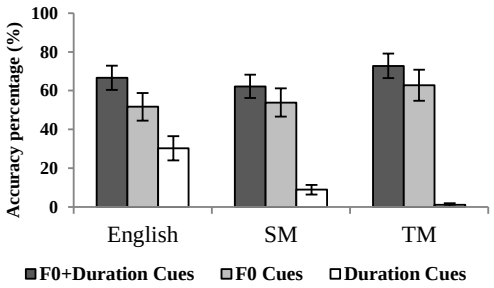


Figure 1. *All participants' mean sequence-encoding accuracy (and standard errors).*

A first logit mixed-effects model was conducted on the participants' sequence-encoding accuracy. The results, reported in Table 3, showed significant effects of cue for F0 and for duration, as well as significant cue (duration) × L1 interactions for both the SM and TM groups. This indicates that both the SM and TM groups showed a larger difference between the F0+duration cues condition and the duration cues condition than the English group did.

Table 2. *Model 1: Logit mixed-effects model on all participants' sequence-encoding accuracy (df=1612).*

Variable	Est.	SE	*z*	*p*
Cue (F0)	–0.7	0.24	–7.1	<.01
Cue (Duration)	–1.8	0.25	–3.2	<.001
L1 (SM)	0.2	0.85	–0.5	>.1
L1 (TM)	0.5	0.47	1.0	>.1
Cue (F0) × L1 (SM)	0.4	0.33	1.1	>.1
Cue (Duration) × L1 (SM)	–1.5	0.42	–3.6	<.001
Cue (F0) × L1 (TM)	0.2	0.35	0.5	>.1
Cue (Duration) × L1 (TM)	–4.8	0.86	–5.5	<.001

A second logit mixed-effect model was conducted on the L2 learners' sequence-encoding accuracy. The results, shown in Table 4, revealed a significant effect of cue for duration, as well as a significant cue (duration) × L1 dialect interaction. This indicates that TM listeners showed a larger difference between the F0+duration cues condition and the duration cues condition than the SM listeners did.

Table 3. *Model 2: Logit mixed-effects model on the L2 learners' sequence-encoding accuracy (df=1615).*

Variable	Est.	SE	*z*	*p*
Cue (F0)	0.4	0.23	–1.7	>.1
Cue (Duration)	–3.3	0.34	–9.6	<.001
L1 dialect	0.7	0.48	1.5	>.1
Cue (F0) × L1 dialect	–0.2	0.34	–0.6	>.1
Cue (Duration) × L1 dialect	–3.3	0.90	–3.7	<.001

To summarize, as we predicted, the SM and TM groups do not differ from each other in the use of F0 cues to encode sequences of English non-words. However, TM listeners differ from both English listeners and SM listeners in the use of duration to encode English non-word sequences.

3.2. Word-encoding accuracy

As mentioned above, word-encoding accuracy was also analyzed to compare listeners' performance on words stressed on the initial syllable and words stressed on the final syllable. Figure 2 displays the accuracy of the individual word encoding for the three cue conditions, with the top panel showing the accuracy for non-words with initial stress and the bottom panel showing the accuracy for non-words with final stress.

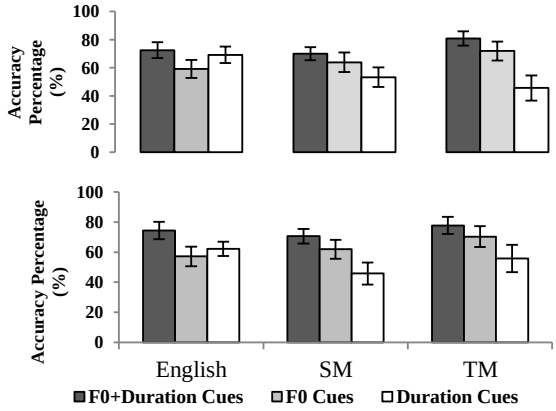


Figure 2. *All participants' word-encoding mean accuracy (and standard errors) for non-words with initial stress (top panel) and non-words with final stress (bottom panel).*

Third and fourth logit mixed-effects models were conducted on the participants' word-encoding accuracy separately for non-words with initial stress and non-words with final stress. The results are reported in Table 4. For non-words with initial stress, the model revealed significant effects of cue for F0, as well as significant cue (duration) × L1 interactions for both the SM and TM groups. This indicates that for non-words with initial stress, both the SM and TM groups showed a larger difference between the F0+duration cues condition and the duration cues condition than the English group did. For non-words with final stress, the model revealed significant effects of cue for F0 and duration, as well

as significant cue (duration) × L1 interactions for both the SM and TM groups. This again indicates that for non-words with final stress, both the SM and TM groups showed a larger difference between the F0+duration cues condition and the duration cues condition than the English group did.

Table 4. *Models 3-4: Logit mixed-effects model on all participants' word-encoding accuracy (df=1612).*

Word-initial stress				
Variable	Est.	SE	*z*	*p*
Cue (F0)	–0.7	0.17	–4.0	<.001
Cue (Duration)	–0.2	0.17	–1.0	>.1
L1 (SM)	–0.2	0.26	–0.7	>.1
L1 (TM)	0.4	0.27	1.6	>.1
Cue (F0) x L1 (SM)	0.4	0.23	1.6	>.1
Cue (Duration) x L1 (SM)	–0.6	0.23	–2.5	<.05
Cue (F0) x L1 (TM)	0.2	0.25	0.6	>.1
Cue (Duration) x L1 (TM)	–1.5	0.24	–6.2	<.001
Word-final stress				
Variable	Est.	SE	*z*	*p*
Cue (F0)	–0.9	0.17	–5.2	<.001
Cue (Duration)	–0.7	0.17	–3.8	<.001
L1 (SM)	–0.1	0.50	–0.2	>.1
L1 (TM)	–0.6	0.50	1.2	>.1
Cue (F0) x L1 (SM)	0.5	0.24	1.8	>.05
Cue (Duration) x L1 (SM)	–0.6	0.24	–2.4	<.05
Cue (F0) x L1 (TM)	0.4	0.26	1.4	>.1
Cue (Duration) x L1 (TM)	–0.8	0.26	–3.0	<.01

Fifth and sixth logit mixed-effects model were conducted on the L2 learners' word-encoding accuracy separately for non-words with initial stress and non-words with final stress. The results are reported in Table 5. For non-words with initial stress, the model revealed a significant effect of cue for duration, a significant effect of L1 dialect, and a significant cue (duration) × L1 dialect. This indicates that TM listeners showed a larger difference between the F0+duration cues condition and the duration cues condition than the SM listeners did for non-words with initial stress. For non-words with final stress, the model revealed significant effects of cue for duration and F0, but no interaction between either cue and L1 dialect. This means that TM listeners did not differ from SM listeners in their use of duration cues in non-words with final stress.

Table 5. *Models 5-6: Logit mixed-effects model on L2 learners' word-encoding accuracy (df=1615).*

Word-initial stress				
Variable	Est.	SE	*z*	*p*
Cue (F0)	–0.3	0.16	–1.8	>.05
Cue (Duration)	–0.7	0.16	–4.7	<.001
L1 dialect	0.6	0.23	2.7	<.01
Cue (F0) x L1 dialect	–0.2	0.24	–0.9	>.1
Cue (Duration) x L1 dialect	–0.9	0.23	–3.9	<.001
Word-final stress				
Variable	Est.	SE	*z*	*p*
Cue (F0)	–0.4	0.17	–2.6	<.01
Cue (Duration)	–1.2	0.17	–7.2	<.001
L1 dialect	–0.7	0.58	1.3	>.1
Cue (F0) x L1 dialect	–0.1	0.26	–0.3	>.1
Cue (Duration) x L1 dialect	0.2	0.26	–0.8	>.1

In summary, the SM and TM groups pattern like each other in the use of F0 to encode stress in individual English non-words, irrespective of whether stress is word-initial or word-final. However, as predicted, SM and TM listeners differ from each other in the use of duration cues to word-initial stress, a stress pattern that exists in SM, but not in the use of duration cues to word-final stress, a stress pattern that does not exist in SM.

4. Discussion and Conclusion

Our results showed that the SM- and TM-speaking L2 learners of English did not differ from each other in the use of F0 cues to process English stress. We attribute these results to the presence of lexical tones in Mandarin, which are signaled primarily by F0 cues.

However, TM listeners differed from both English and SM listeners in the use of duration cues to English stress, with TM listeners being relatively less accurate than the other two groups in the duration cues condition as compared to the F0+duration cues condition. These results were found in the participants' sequence-encoding accuracy and their word-encoding accuracy for non-words with initial stress. These findings are exactly as predicted: In SM but not in TM, duration is used to encode a stress distinction (i.e., full-full vs. full-reduced). This in turns allow SM listeners to use duration for encoding English stress. Importantly, SM does not have a reduced-full stress pattern, and it is precisely in this word-final stress condition that SM and TM patterned similarly.

The results also showed that SM listeners differed from English listeners in the use of duration cues to English stress. One possibility is that the greater occurrence of vowel reduction in English than in SM may lead English listeners to be more sensitive to duration cues than SM listeners. Alternatively, SM may have relatively few words that contrast in stress placement, potentially leading SM listeners to rely less on duration cues as compared to English listeners [13].

Our results are consistent not only with the proposal that non-native listeners weigh prosodic cues to stress as a function of the role of these cues in the L1 [1], but also with more general cue-weighting accounts of L2 speech perception [28,29]. Learning to process English stress depends in large part on the prosodic cues used to realize stress and on whether these cues are similarly used to access words in the L1. The present study also suggests that not only the L1, but also the L1 dialect, plays a crucial role in determining whether non-native listeners can use specific prosodic cues to encode English stress. L2 speech processing models should thus consider the influence of L1 dialect, which may impact how prosodic cues are weighted in L2 processing.

5. Acknowledgements

We would like to thank Yu-fu Chien for help with the data collection in Taiwan; Dr. Allard Jongman, Dr. Joan Sereno, Dr. Jie Zhang, as well as the members of LING 850, for their valuable comments on this research; and all participants.

6. References

[1] Cutler, A. Native listening: Language experience and the recognition of spoken words, MIT Press, 2012.

[2] Beckman, M. E. Stress and Non-Stress Accent, Foris, 1986.

[3] Fry, D. "Duration and intensity as physical correlates of linguistic stress", Journal of the Acoustic Society of America, 27:765-768, 1955.

[4] Lieberman, P. "Some acoustic correlates of word stress in American English", Journal of the Acoustic Society of America 32:451-454, 1961.

[5] Cooper, N., Cutler, A. and Wales, R. "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners", Language and Speech, 45:207-228, 2002.

[6] Cutler, A., Wales, R., Cooper, N. and Janssen, J. "Dutch listeners' use of suprasegmental cues to English stress", in J. Trouvain and W. J. Barry [Eds], Proceedings of the 16th International Congress for Phonetic Sciences, 1913-1916, Pirrot, 2007.

[7] Cutler, A. "Forbear is a homophone: Lexical prosody does not constrain lexical access", Language and Speech, 29:201-220, 1986.

[8] Braun, B., Lemhöfer, K. and Mani, N. "Perceiving unstressed vowels in foreign-accented English", Journal of the Acoustical Society of America, 129:376-387, 2011.

[9] Cutler, A. and Clifton, C. E. "The use of prosodic information in word recognition", in H. Bouma and D. G. Bouwhuis [Eds], Attention and performance X, 183-196, Erlbaum, 1984.

[10] Fear, B. D., Cutler, A. and Butterfield, S. "The strong/weak syllable distinction in English", Journal of the Acoustical Society of America, 97:1893-1904, 1995.

[11] Small, L. H., Simon, S. D. and Goldberg, J. S. "Lexical stress and lexical access: Homographs versus nonhomographs", Perception and Psychophysics, 44:272-280, 1988.

[12] Chao, Y.-R. A grammar of spoken Chinese. University of California Press, 1968.

[13] Duanmu, S. The phonology of standard Chinese. Oxford University Press, 2007.

[14] Chen, Y. and Xu, Y. "Production of weak elements in speech—evidence from F0 patterns of neutral tone in Standard Chinese", Phonetica, 63:47-75, 2006.

[15] Lin, M. and Yan, J. "Beijinghua qingsheng de shengxue xingzhi" [Acoustic characteristics of neutral tone in Beijing Mandarin], Fangyan, 3:166-178, 1980.

[16] Shen, X. S. "Relative duration as a perceptual cue to stress in Mandarin", Language and Speech, 36:415-433, 1993.

[17] Ye, Y. and Connine, C. M. "Processing spoken Chinese: The role of tone information", Language and Cognitive Processes Special Issue: Processing East Asian Languages, 14:609-630, 1999.

[18] Hallé, P. A., Chang, Y.-C. and Best, C. T. "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners", Journal of Phonetics, 32: 395-421, 2004.

[19] Zhang, Y. and Francis, A.L. "The weighting of vowel quality in native and non-native listeners' perception of English lexical stress", Journal of Phonetics, 38, 260-271, 2010.

[20] Wang, Q. "L2 stress perception: The reliance on different acoustic cues", in P. Barbosa, S. Madureira, and C. Reis [Eds], Proceedings of Speech Prosody 2008, 135-138, Campinas, Brazil, 2008.

[21] Lai, Y. Acoustical Realization and Perception of English Lexical Stress by Mandarin Learners, Unpublished PhD Dissertation. University of Kansas, 2008.

[22] Dupoux, E., Peperkamp, S. and Sebastián-Gallés, N. "A robust method to study stress 'deafness'", Journal of the Acoustical Society of America, 110:1606-1618, 2001.

[23] Dupoux, E., Sebastián-Gallés, N., Navarrete, E. and Peperkamp, S. "Persistent stress 'deafness': The case of French learners of Spanish", Cognition, 106: 682-706, 2008.

[24] Tremblay, A. "Is second language lexical access prosodically constrained? Processing of word stress by French Canadian second language learners of English", Applied Psycholinguistics, 29:553-584.

[25] Brown, J. D. "Relative merits of four methods for scoring cloze tests", Modern Language Journal, 64:311-317, 1980.

[26] Boersma, P. and Weenink, D. Praat: Doing phonetics by computer [computer program], version 5.3, retrieved from http://www.praat.org, December 2012.

[27] Baayen, R. H. Analyzing linguistic data: A practical introduction to statistics. Cambridge University Press, 2008.

[28] Francis, A.L. and Nusbaum, H.C. "Selective attention and the acquisition of new phonetic categories", Journal of Experimental Psychology: Human Perception and Performance, 28(2): 349-366, 2002.

[29] Holt, L.L. and Lotto, A.J. "Cue weighting in auditory categorization: Implications for first and second language acquisition", Journal of the Acoustical Society of America, 119: 3059-3071, 2006.

The interplay between prosodic phrasing and accentual prominence on articulatory lengthening in Italian

Caterina Petrone,[1] Mariapaola D'Imperio[1,2], Leonardo Lancia[3] & Susanne Fuchs[4]

[1] Laboratoire Parole et Langage, CNRS & Université Aix-Marseille, Aix-en-Provence, FR;
[2]Institut Universitaire de France (IUF), FR ;
[3] Max Planck Institute for Evolutionary Anthropology, Leipzig, GE;
[4]Zentrum für Allgemeine Sprachwissenschaft, Berlin, GE

(caterina.petrone; mariapaola.dimperio)@lpl-aix.fr, leonardo_lancia@eva.mpg.de, fuchs@zas.gwz-berlin.de

Abstract

The distribution of preboundary lengthening within the phrase-final word is controversial. In CV syllables immediately preceding a prosodic boundary, the acoustic duration of the syllable onset C is less involved than that of the following rime V in the lengthening phenomenon. Moreover, preboundary lengthening might be extended to the stressed/accented rime within the phrase final word. On the other hand, articulatory the constriction gesture for the onset consonant can be lengthened despite not being immediately adjacent to a boundary. In this preliminary study, we explore the effects of prosodic boundary and prominence in Italian, at both acoustic and articulatory level. Bilabial consonants in CV onset position were examined. The consonants were inserted in unstressed (word final) and stressed (penultimate vs. antepenultimate) syllables occurring close to prosodic boundaries of different levels. In final syllables, the acoustic duration of the onset consonant was not affected by the prosodic boundary manipulation whereas the closing gesture duration showed a pattern of lengthening which was stronger for higher level prosodic boundaries. In non-final syllables, no acoustic/articulatory effect was found for onset consonants but only on the stressed vowels in penultimate position. Structural, phonological and phonetic constraints might be at work in determining preboundary lengthening.

Index Terms: preboundary lengthening, articulatory movements, onset duration, Italian, prosodic hierarchy, pi-gesture.

1. Introduction

Acoustic studies in different languages have found that segments in a phrase-final word (i.e., close to a prosodic boundary edge) undergo final ("preboundary") lengthening [1]). Prosodic phrasing also influences articulatory gestures of segments close to the phrase edge, with gestures being slower and less overlapped before a major prosodic boundary [2,3]. [2] invokes the activation of a prosodic (or π-) gesture, whose function is to slow down the timing of the co-occurring articulatory gestures for boundary-adjacent segments. As [1] put it, this gesture indicates "the onset and offset of a time period during which the clock that controls the timing of other gestures ticks more slowly".

An important issue concerning preboundary lengthening is how its precise distribution is determined within the phrase-final word. Though the magnitude of the effect is larger on the syllable immediately adjacent to the boundary, a few acoustic studies has also showed that preboundary lengthening can also affect syllables which are away from the boundary (see [1] and references therein). According to [1], there are at least two possible approaches to account for such a phenomenon.

The *Structure-based* view predicts that the distribution of (acoustic) final lengthening is conditioned by linguistic structure. For instance, within the same syllable (e.g., a word-final CV), only the syllable rime (the vowel V) would be affected by lengthening. Moreover, the leftward extension of preboundary lengthening would be conditioned by the position of the stress within the phrase-final word. For instance, the *Word Rime hypothesis* [1] predicts that preboundary lengthening would start from the rime of the stressed syllable up to the rime of the word-final syllable, while the syllable onset consonant in both the stressed and final syllables would be unaffected by preboundary lengthening (cf. also [1] for a model of multiple targets of preboundary lengthening).

The *Content-based* view, on the other hand, predicts that the actual portion undergoing lengthening is structurally variable since it results from a lengthening gesture of fixed duration. In particular, [3] claimed that in English, the half-point of the π-gesture is temporally anchored to the boundary itself, and that its temporal activation is fixed. As a consequence, the π-gesture can overlap with earlier or later portions of the last word within a phrase depending on structural complexity (e.g., presence or absence of a coda consonant) and intrinsic gesture duration (e.g., presence of lax vs. tense vowels) of the phrase-final syllables. Note that, articulatory studies conducted within the Content-Based view, such as the ones supporting the π-gesture hypothesis, report preboundary lengthening of articulatory constriction movements for the onset consonant of final syllables [3] and of prominent syllables in non-final position [4], though the effects appear to be speaker dependent.

The magnitude of phrase-final effects is not only modulated by syntagmatic factors, but also by paradigmatic ones. Given that prosodic structure is hierarchically organized, phrase-final effects are predicted to be stronger before higher-level boundaries at both acoustic and articulatory level [5, 6]. In Italian -the language under investigation here- the number of phrasing levels is still controversial. Whereas both the intonation (IP) and the intermediate phrase (ip) are well attested, a third level of phrasing has been tentatively proposed, the Accentual Phrase (*AP*) (see also [7] for evidence for this level in French). Specifically, in Neapolitan Italian, [8] found that a tone is inserted at the right edge of the *AP*, which is differently specified in questions (H_{AP}) and in statements (L_{AP}). Also, [9] found that the duration of the accented syllable cumulatively increases with prosodic boundary strength, thus suggesting an interaction between prosodic phrasing and accentual prominence in Italian.

In the present study, we test whether preboundary lengthening affects the production of the consonant in syllable onset position in Italian. If the Structure-based view is correct, its duration in word final CV syllables will be constant since only the syllable rime (V) undergoes lengthening. Alternatively, the syllable onset will be sensitive to vicinity to a prosodic boundary and the effect will be greater for higher level boundaries. More specifically, the effect of prosodic boundary on both acoustic onset duration and articulatory closing movements will increase from the syllable to the *IP*-level. If the *π*-gesture is anchored to the phrase edge, we expect also greater amplitudes and lower velocity of preboundary closing movements. Moreover, in line with the *Word Rime hypothesis*, preboundary lengthening should start from the stressed/accented vowel, independent of its position within the word (i.e., penultimate vs. antepenultimate).

The target words were inserted in both statement and question utterances, since we know that Neapolitan Italian questions and statement temporal structure is different both at a global and local [10] level. As a consequence, we expect that the effects of prominence and prosodic hierarchy on the acoustic and articulatory parameters will also differ across sentence modality.

2. Corpus and methods

The target consonant was /m/ which was always in onset position in the syllable /ma/. The consonant was preceded and followed only by /a/ to guarantee relatively large articulatory movements. Moreover, the syllable /ma/ appeared either in final or non-final position within the word (e.g., ABRAma, TaMAra[1]). Six trisyllabic words were included. They were always proper names. When the position of /ma/ was word-final, the target syllable was unstressed (ABRAma, PAnama). When /ma/ was in non-final position, the target syllable could also bear a stress. To control for the effect of stress location, the stressed syllable could be either in penultimate (TaMAra) or antepenultimate (MArica) position. The stressed syllables were also accented. The six words were inserted in carrier Subject-Verb-Object (SVO) sentences.

Table 1. *Example of sentences by boundary type.*

Boundary type	Sentences
IP	*Le lettere da Malaga]$_{AP}$ e da Panama]$_{IP}$ per quanto ne so, stanno nel cassetto* "The letters from Malaga and from Panama, as far as I know, are in the drawer"
ip	*Le lettere da Malaga]$_{AP}$ e da Panama]$_{ip}$ stanno nel cassetto*
AP	*Le lettere da Panama]$_{AP}$ e da Malaga]$_{ip}$ stanno nel cassetto*
syll	*Le lettere da MaRIna]$_{AP}$ e da Marica]$_{ip}$ stanno nel cassetto* "The letters from Marina and from Marica are in the drawer"

The target syllable appeared in different prosodic position within the sentences, depending on the vicinity to the right boundary of four prosodic constituents (IP, ip, AP and syllable). While the IP-boundary was triggered by the insertion of a parenthetical clause [11], the long subject constituent was expected to be uttered as a single ip [12]. Moreover, since the AP edge seems to coincide with the end of the prosodic word [8], the syllable ("syll") condition only included the unstressed target syllables, which were in AP-internal position. That is, they were located in initial position within the word (da MaRIna; da MaRIsa, "*from Marina/Marisa"*). Sentence modality (Q vs. S) was also varied. Examples of target sentences are shown in Table 1.

Simultaneous acoustic and articulatory data were collected by means of Electromagnetic Articulography (AG500) at the ZAS laboratory (Berlin). As for acoustics, the duration of /m/ and the following /a/ and of the stressed vowel were manually labeled in PRAAT [13]. The articulation of the labial consonant was investigated by calculating lip aperture, i.e. the Euclidean distance between upper (UL) and lower lip (LL). For this purpose, articulatory movement tracking of two transducers placed at UL and LL were used. Kinematic data were labeled and measured by means of the MView software developed by Mark Tiede. Three points were defined: the onset of the closing movement from /a/ to /m/, the time point of the movement extremum and the end of the opening movement.

Three dependent variables were derived. In particular, for the closing movements, we measured: (1) closing gesture duration (i.e., time from onset of the closing movement to the lip closure); (2) time-to-peak velocity (i.e., time from onset of closing movement to velocity peak of the closing gesture); and (3) displacement (i.e., Euclidean distance between onset of the closing gesture and the offset).

One male, native speaker of Neapolitan Italian, was asked to read the sentences 3 times. No instructions were given as for the prosodic phrasing or accentuation of the sentences. Stimuli were randomly presented on paper sheets.

The statistical analyses included a series of mixed models, in which each acoustic and articulatory variable was analyzed as a function of stress location (penultimate vs. antepenultimate), sentence type (Q/S) and boundary type (IP/ip/AP/syll). Contrasts between successive levels of boundary type (IP vs. ip; ip vs. AP; AP vs. syll) were calculated. A maximal random effects structure with intercepts and slopes for Words was considered [14]. The inclusion of the random terms as well as of the interactions among fixed factors was warranted by likelihood ratio tests. The cutoff for significance is $|t| > 2$ [15].

3. Results

3.1.1. Acoustic Results

The duration of the onset consonant and that of the following vowel in word-final syllables are shown in Fig. 1. The consonant duration was on average 43 ms. The statistical analysis showed no difference with respect to boundary type, sentence modality and stress location. However, the vowel following the target consonant was significantly shorter in AP (65 ms) than in ip (80 ms) in Q [β = -0.014; t = -3.3; SE= 0.004]. In Q, only the contrast between IP (74 ms) and ip (62 ms) was significant [β = -0.011; t = -2.2; SE= 0.005]. In both modalities, no difference was found between the lower levels of constituency (AP vs. syll). As for the effect of sentence modality, final vowels in IP and ip conditions were shorter in

[1] Target syllables are underlined, stress position is indicated by capitals.

S than in Q but the amount (6 ms) is close to duration measurement error.

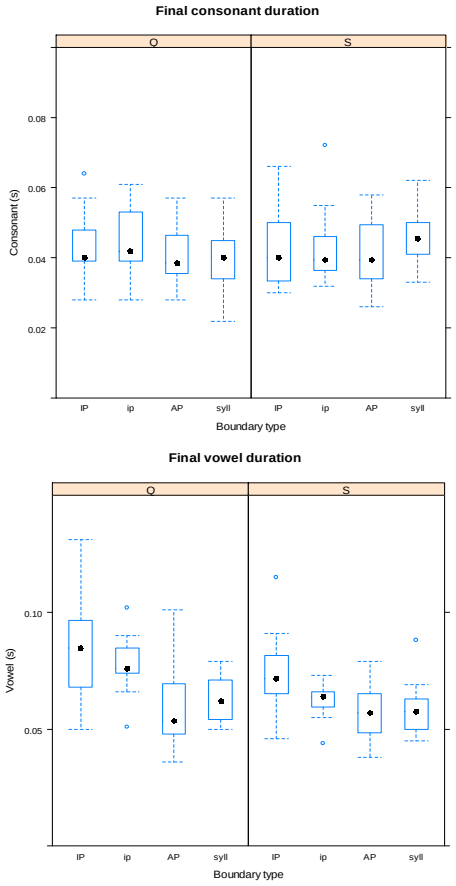


Figure 1: *Boxplots for onset consonant (top) and vowel (bottom) duration in the word final syllable against boundary type and split across sentence type. Data are collapsed across stress location.*

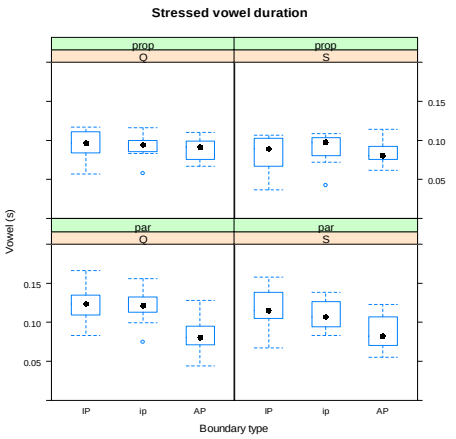
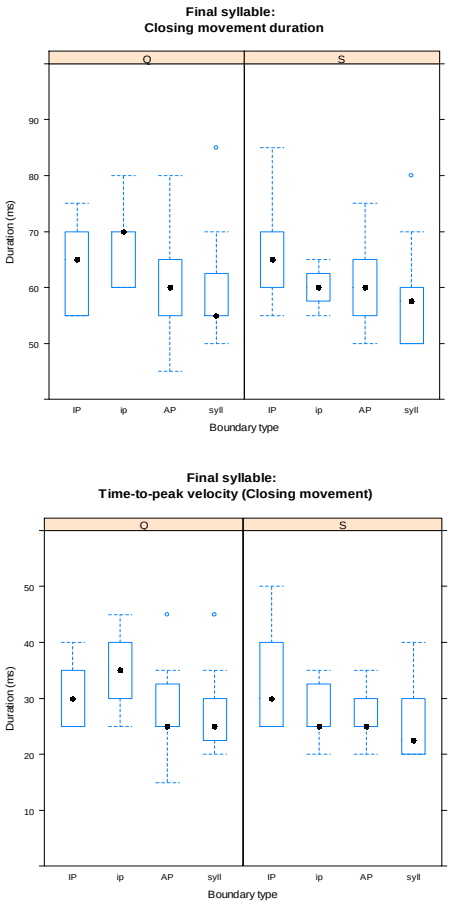


Figure 2: *Boxplots for stressed vowel duration against boundary type. Data are split across sentence type and stress location (prop= proparoxytons; par = paroxytons).*

In antepenultimate position, the stressed vowel duration was on average 90 ms (in Q) and 86 ms (in S). However, no significant difference was found across sentence modality and boundary type. In penultimate position, there was no difference in stressed vowel duration between IP and ip. The duration was shorter in AP than in ip [β = -0.02; t = -2.9; SE= 0.006] in both Q and S. The duration of the stressed vowel is shown in Fig. 2.

3.1.2. Articulatory Results

The closing movement of the onset consonant in word-final syllables was affected by prosodic boundary type (Fig. 3). The closing movement duration in Q was significantly shorter in AP than in ip [β = -8.4; t = -2.7; SE= 3]. Neither the contrast between IP vs. ip nor that between AP vs. syll was significant. In S, there was no significant effect of the successive contrasts. The effect of boundary type on time-to-peak velocity also varied with sentence type. In fact, while no significant contrast was found in S, a significant contrast between ip and AP was found in Q [β = 9.9; t = 2.2; SE= 4.4]. Displacement was smaller in AP than in ip [β = -3.5; t = -6.2; SE= 0.5] in both Q and S. The contrast between AP and syll was significant in both sentence modalities [β = -2.6; t = -5.2; SE= 0.4]. However, there was no difference between IP and ip. No effect of stress location was found on the three articulatory parameters. The closing movement duration, time-to-peak velocity and displacement of the onset consonant in the stressed syllables were not affected by boundary type and sentence modality.



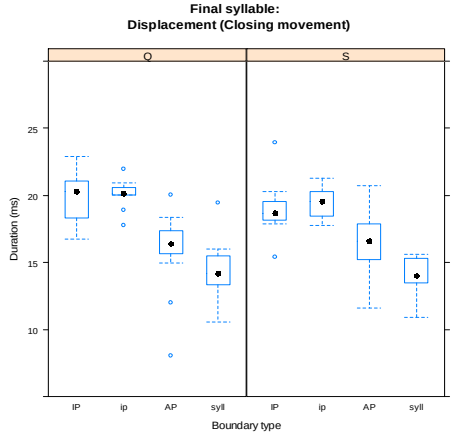


Figure 3: *Boxplots for the duration (top), time-to-peak velocity (middle) and displacement (bottom) of the closing movement of the onset consonant in the final syllable. Data are split across sentence type.*

4. Discussion

While the acoustic results did not show a significant lengthening of the onset consonant in the word final syllable, a clear preboundary lengthening was found for the word final vowel. This result is in line with the *Structure-based* view, which predicts that preboundary lengthening affects only specific linguistic elements. Similarly as English, in Italian the final syllable rime appears to be the sub-syllabic unit on which lengthening applies. The acoustic duration of the stressed vowel was longer before a major (ip/IP) than a minor (AP/syll) prosodic boundary (cf. [1] for similar results in English). This was true only when the stressed syllable was one syllable away from the boundary (in penultimate syllables). This means that preboundary lengthening cannot be determined solely by structural factors. If prominence (accent/stress) would have played a role in determining the domain of final lengthening (as predicted by the *Word Rime hypothesis*), an effect of boundary type should have been found on the stressed vowel independent of its position within the word. On the contrary, durational effects of boundaries are visible only on the segments closer to the edges of the phrase.

This could be in line with the Content View, for which the domain of lengthening depends on the activation of a gesture of fixed duration. It might be possible that, given the simple syllabic structure of the word final syllables (CV), the lengthening gesture could overlap with earlier portions of the word (the penultimate syllables) which would show in turn acoustic lengthening. However, because of its fixed duration, the gesture could not overlap with even earlier portions of the word (the antepenultimate syllables).

A straightforward interpretation of the results is though limited by the partial mismatch between the acoustic and kinematic data. The kinematic results showed a lengthening pattern for the closing movement of the preboundary labial consonant, as well as for time-to-peak velocity and displacement for the same segment, which is in line with the π-gesture hypothesis. This suggests that differently from acoustic lengthening, articulatory constriction movements for the onset consonant of final syllables are affected by prosodic constituency [3]. Since articulatory constrictions usually begin before the acoustic beginning of consonants, variations in the articulatory movements (e.g., closing movement duration) might be acoustically reflected in the lengthening of the preceding vowel rather than on the lengthening of the onset consonant of the word-final syllable.

Moreover, acoustically, preboundary lengthening may extend to penultimate syllables, whereas the articulatory lengthening appears to be limited to final syllables. In fact, the articulatory parameters for the onset consonants within the stressed syllables were unaffected by prosodic constituency. This suggests that the temporal scope of final lengthening is larger in the acoustic than in the articulatory domain. Taken together, the acoustic and articulatory data seem to support a Hybrid view of preboundary lengthening [1], which proposes that the scope of lengthening is determined by structural, phonological and phonetic properties of the phrase-final word. However, more data are needed to better understand the articulatory-acoustic mapping.

As for the number of prosodic levels in Italian, the acoustic and articulatory data showed mixed evidence of two or three levels of phrasing, depending on the parameter examined. In line with the standard view of prosodic hierarchy, this suggests the existence of only two/three categorically distinct levels of prosodic constituency. However, it has been proposed that Italian speakers can distinguish three or even four [8, 9] different categories. A possible explanation for such a discrepancy is that prosodic boundaries are produced in a gradient rather than a categorical manner [16]. As a consequence, values for segmental duration would gradually increase from the lowest to the highest level of the prosodic hierarchy instead of clustering around a limited number of values (one for each prosodic constituent).

Finally, the effects of boundary type on segmental duration differed across sentence modality, indicating that different cues other than the parameters examined (e.g., F0) may be at work to signal prosodic constituency in questions and statements.

5. Conclusion

As predicted by the π-gesture hypothesis, labial closing movements of preboundary consonants show a hierarchical effect of prosodic boundary type despite not being immediately adjacent to the juncture (being one segment away). The effect is both temporal (lengthening of the gesture) and spatial (larger amplitude), though the temporal effect is stronger. On the other hand, the effect does not extend to the closing movements of the stressed syllable, neither for penultimate nor for antepenultimate stress. More data is needed to determine if the findings can be generalized to other speakers and consonantal types and whether the effect is caused by lengthening of the preceding vowel.

6. References

[1] Turk , A. & S. Shattuck-Hufnagel (2007). Phrase-final lengthening in American English. *JPhon*, 35(4), 445- 472.

[2] Byrd, D. & E. Saltzman (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *JPhon*, 31(2), 149-180.

[3] Byrd, D., Krivopavic, J. & Lee, S. (2006). How far, how long: On the temporal scope of prosodic boundary effects. *JASA*, 120, 1589-1599.

[4] Byrd, D., & Riggs, D. (2008). Locality interactions with prominence in determining the scope of phrasal lengthening. *JIPA* 38, 187–202.

[5] Fougeron, C. and Keating, P. (1997), Articulatory strengthening at edges of prosodic domains. *JASA*, 101, 3728-3740.
[6] *Cho, T. &* and *Keating*, P. (2001). Articulatory strengthening at the onset of prosodic domains in Korean. *JPhon* 28:155-190.
[7] Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (ed.), *Intonation: Analysis, modelling and technology*. Dordrecht, 209-242.
[8] Petrone, C. & D'Imperio, M. (2008). Tonal structure and constituency in Neapolitan Italian: Evidence for the accentual phrase in statements and questions. In *Proceedings of Speech Prosody 2008*, 301-304.
[9] Petrone, C. (2008). *Le rôle de la variabilité phonétique dans la représentation des contours intonatifs et de leur sens*. PhD thesis, Université Aix-Marseille.
[10] Cangemi, F. & D'Imperio, M. (2013) Tempo and the perception of sentence modality, *Laboratory Phonology*, 4(1), 191-219.
[11] Nespor, M., & Vogel, I. (1986). Prosodic phonology. Dordrecht : Foris.
[12] Prieto, P., D'Imperio, M., Elordieta, G., Frota, S. & Viga´ rio, M. (2006). Evidence for soft preplanning in tonal production: Initial scaling in Romance. In Proceedings of Speech Prosody. Dresden: TUD Press Verlag der Wissenschaften GmbH, 803-806.
[13] Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.
[14] Barr, D. J., Levy R., Scheepers C. & Tily H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. JML, 68(3), 255-278.
[15] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. JML, 59(4), 390-412.
[16]. Krivokapić, J. (2007). The planning, production, and perception of prosodic structure. Ph.D. Thesis. University of Southern California.

Quasi-neutralization of stress contrasts in Spanish

Francisco Torreira [1], *Miquel Simonet* [2], *José I. Hualde* [3]

[1] Max Planck Institute, Nijmegen, The Netherlands

[2] University of Arizona, USA

[3] University of Illinois at Urbana-Champaign, USA

Francisco.Torreira@mpi.nl, simonet@email.arizona.edu, jihuade@illinois.edu

Abstract

We investigate the realization and discrimination of lexical stress contrasts in pitch-unaccented words in phrase-medial position in Spanish, a context in which intonational pitch accents are frequently absent. Results from production and perception experiments show that in this context durational and intensity cues to stress are produced by speakers and used by listeners above chance level. However, due to substantial amounts of phonetic overlap between stress categories in production, and of numerous errors in the identification of stress categories in perception, we suggest that, in the absence of intonational cues, Spanish speakers engaged in online language use must rely on contextual information in order to distinguish stress contrasts.

Index Terms: lexical stress, stress cues, phonological neutralization, Spanish.

1. Introduction

In Spanish, the correlates of lexical stress appear to be subtler than in other languages with lexically contrastive stress that have been investigated in this respect, such as English or Dutch, and also closely related languages such as Portuguese and Catalan. Spanish lacks systematic reduction of vowels in unstressed syllables [1], and durational differences between stressed and unstressed syllables in this language are relatively small compared to Portuguese [2, 3]. This results in lexical differences in stress placement often being difficult to perceive for learners of Spanish as a second language [4, 5].

The most robust cues to determine the position of lexical stress in a Spanish word are present when it carries an intonational pitch accent, in which case the pitch accent is associated to the lexically-stressed syllable, lending it phonetic prominence. When the word does not carry an intonational pitch accent, on the other hand, the perception of lexical stress may be jeopardized. Given the important role of pitch accents in disambiguating stress contrasts, the question arises whether lexical stress distinctions are maintained under such conditions, as has been reported for other stress languages such as Dutch and English [6, 7, among others]. To address this issue, [8] conducted a study where Spanish speakers were asked to produce parenthetical reporting clauses, which systematically exhibit a low flat pitch contour judged to lack pitch accents. Under this condition, minimal pairs involving paroxytone and oxytone verb forms (e.g. *determino* vs. *determinó*, stressed syllables in bold) were found to be distinguished in production by consistent differences in vowel duration, and less reliably by intensity.

Reporting clauses such as the ones used by [8] are rather rare in conversational Spanish. On the other hand, a very frequent context where words do not appear to carry pitch accents in Spanish is in phrase-medial position within broad-focus long intonational phrases (IPs). This is illustrated by the following examples, which are typically realized as single intonational phrases ending in rising continuation intonation:

a) [***Siem**pre que **mi**ro la **ho**ra*]$_{IP}$, …
'Every time I look up the time, …'

b) [***Siem**pre que mi**ró** la **ho**ra*]$_{IP}$, …
'Every time she looked up the time, …'

Our observations from the Nijmegen Corpus of Casual Spanish (NCCSP) [9] suggest that in such contexts stress contrasts might be lost, without duration and intensity cues compensating for the absence of a pitch accent. Audio examples of unaccented words in phrase-medial position extracted from the NCCSp can be accessed online at [10]. These examples contain the words ***de**jo* 'I leave' and *de**jó*** 'she left', which contrast only in the position of lexical stress.

In the present study, we investigate to what extent the lack of pitch accents in phrase-medial position in Spanish leads to a neutralization of stress contrasts. In the following sections, we present a production experiment and a perception experiment aimed at answering this question.

2. Experiment I: Production

In Experiment I, we investigate the extent to which Spanish speakers maintain lexical stress contrasts in words lacking a pitch accent in phrase-medial position. We elicit unaccented verbal forms contrasting in the position of stress (e.g. ***ta**po* vs. *ta**pó***) located in phrase-medial position, and examine their phonetic realization in terms of f0, duration, intensity, formant values, and two forms of consonantal lenition using regression modeling and a cross-validation procedure. We use verbal forms from the first conjugation (with infinitives ending in –ar) as targets words, because they provide a systematic contrast in stress placement (oxytones for 3rd p. sg, past tense, vs. paroxytones for, 1st p. sg, present tense).

2.1 Method

Nine native speakers of European Spanish (4 female, 5 male) were recorded in a quiet room with an Audix HT5 head-worn microphone connected to a Sound Devices MM-1 pre-amp, which was in turn connected to a Marantz PMD660 digital recorder. The signal was digitized at 44.1 kHz, 16-bit quantization. Subjects were seated in front of a computer screen and presented with strings of words consisting of a subject pronoun, a verbal tense (present or past), an interrogative pronoun, a target verb in the infinitive form, and

a noun phrase serving as a verbal object. The participants' task was to first examine the elements on the screen, and then produce a fluent wh-question. For instance, for the stimulus *PASADO/él*: *¿Cómo/tocar/tu estrella?* 'PAST/he: How/touch/your star?, the expected response was *¿Cómo tocó tu estrella?* 'How did she touch your star?'. Wh-questions were chosen as carrier sentences because they are typically produced as one intonational phrase, both in spontaneous and laboratory speech. This method elicited a large number of phrase-medial verbal forms belonging to three minimal pairs: ***to**co* 'I touch' vs. *to**có*** 'she touched'; ***ta**po* 'I cover' vs. *ta**pó*** 'she covered'; ***cor**to* 'I cut' vs. *cor**tó*** 'she cut'.

The experimental materials included 60 target sentences (3 verbs * 2 tenses * 10 repetitions) and 60 distractors containing different subject pronouns and verbal tenses. From the 540 elicited tokens (60 target sentences * 9 speakers), we analyzed 456 that were produced as fluent broad-focus intonational phrases. The remaining utterances, exhibiting disfluencies, and less frequently, other phrasing and accentual patterns, were discarded. The majority of the analyzed utterances (74%) were produced with an initial rising accent (L*+>H* in SpToBI notation) on the question word (e.g. *cómo*) and a valley on the stressed syllable of the object (*estrella*) followed by a final rise (L* H%), with falling f0 throughout the phrase-medial verb (*toco* or *tocó*). The second most common intonational contour (24%) consisted of a low accent on the question word (L*), and a rise-fall on the object noun (L+H* L%), with flat f0 throughout the phrase-medial verb.

Acoustic analyses were performed as follows: as a first step, the two syllables of the target verb forms were segmented following standard criteria, with stop consonants starting and vowels ending at the stretches of silence attributable to oral stop closures. The following acoustic measurements were then taken with Praat, and registered as differentials between the first and the second syllable: (a) f0 at the midpoint of the vowel, in Hz; (b) syllable duration, in ms; (c) peak intensity within the syllable, in dB (d); first and second formant values (F1, F2) at the midpoint of the vowel, in Hz. According to several recent studies, Spanish /p t k/ consonants are often voiced and spirantized (realized as approximants, without a complete stop closure) in intervocalic position. These lenition phenomena may be conditioned by prosodic factors such as lexical stress or pitch accent [11, 12, 13]. For this reason, we also annotated (e) the presence of uninterrupted periodicity throughout the closure of each of the two stops in each utterance, and (f) whether vowel formants could be observed throughout the consonantal closure. In cases of spirantization, that is, without clear stop closures, segmentation of the syllables in the target words could not be performed, and the differentials mentioned above were left undefined.

2.2 Results

We first fitted a series of mixed-effects regression models with the f0, duration, F1, and F2 differentials between the first and second syllable of the target verbs as responses, stress pattern (paroxytone, oxytone) as the main predictor, verb type (*cortar*, *tapar, tocar*), and contour type as covariates, and speaker as a random factor. Positive differentials indicate that the first syllable has higher values than the second, while negative differentials indicate the opposite. If lexical stress in phrase-medial unaccented words is distinguished by speakers, some or all of these acoustic differentials should differ between the two stress patterns, with higher differentials for paroxytone words (i.e. with more phonetically prominent first syllables).

Consistent with our impressions that the phrase-medial target words did not carry a pitch accent, no f0 difference was found between the two stress patterns (p = .46). A small difference was observed for F1, with paroxytones having a slightly more open first vowel than oxytones (β = 21.51, t = 4.1, p < .0001), but no difference was observed for F2. Greater differences were observed for duration and intensity. In line with the possibility that a stress distinction is maintained in the absence of pitch accents, paroxytones tended to exhibit longer and more intense first vowels than oxytones (duration: β = 22.3, t = 12.4, p < .0001; intensity: β = 1.77, t = 7.85, p < .0001). Despite these statistical differences, however, we observed a considerable amount of overlap between the two stress patterns, in particular for the intensity differential. This can be appreciated in Figures 1 and 2, which show verb-normalized duration and intensity differentials as a function of stress pattern.

As mentioned above, we also examined consonant voicing and spirantization as possible cues to stress. Around a third (33%) of the consonants in the first syllable of the target word were fully voiced when the syllable was lexically stressed. This percentage was considerably higher (46.4%) in unstressed syllables. As for the consonant onset of the second syllable, it was voiced in 33% of tokens when this syllable was stressed vs. 53.7% when unstressed. Both differences were statistically significant in mixed-effects logistic regression models with consonant voicing as the response, stress pattern as the main predictor, verb type as a covariate, and speaker as a random factor (1st syllable: β = 0.74, z = 3.12, p < .005; 2nd syllable: β =-0.75, z = -3.24, p < .005).

Regarding spirantization, we observed that, in general, it was more frequent in the second consonant (n = 43; 9.4% of the data) than in the first one (n = 10; 2.1%). However, it was only in the first consonant that spirantization seemed to be affected by lexical stress, as suggested by a slight statistical trend in a regression model with spirantization as response, stress pattern as predictor, verb as a covariate, and speaker as a random factor (β = -1.41, z = -1.77, p = .07). In the second consonant, where spirantization was more frequent, no statistical difference was observed (8.3% in stressed syllables vs. 9.5% in unstressed syllables; p = .63).

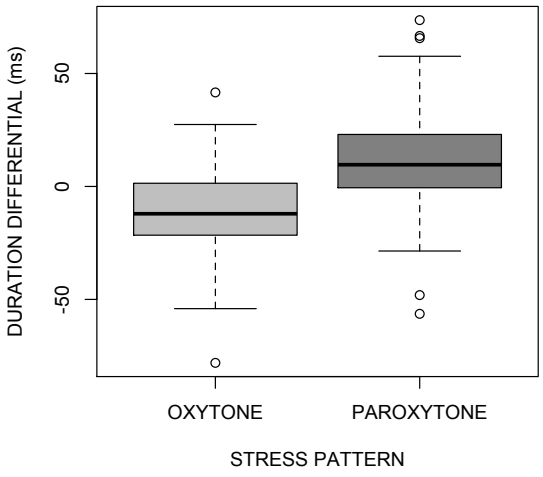


Fig. 1. *Verb-normalized duration differential between first and second syllables in our target words as a function of stress pattern.*

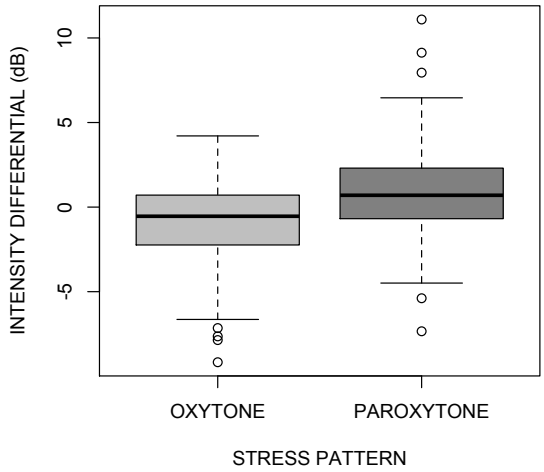


Fig. 2. *Verb-normalized peak intensity differential between first and second syllables as a function of stress pattern.*

To assess the strength of the different identified cues to stress position, we subjected our data to a leave-one-out cross-validation procedure. We simulated predicting stress patterns for unknown data in the following way: for each token in the dataset, we predicted its stress pattern with logistic regression models trained on the rest of the dataset. These models included different combinations of relevant features identified above (i.e. F1, intensity, and duration differentials, and consonantal voicing), plus speaker and verb type.

Table 1 shows percentages of correct classification obtained with five different models. Duration offered the best cue to the stress contrast, achieving 73.8% of correct classification, almost as much as a model containing all features (75.8%). Intensity provided a moderate gain over chance level (62.4%), whereas consonantal voicing and F1 only allowed for results slightly above chance level (55.2% and 51.8%).

We conclude from this analysis that duration and intensity, in this order, provide the best cues to lexical stress in Spanish unaccented words in phrase-medial position. It is also worth noting, however, that almost a quarter of the dataset could not be classified correctly even when all relevant cues where used. This suggests that in such cases there may be little acoustic information that listeners could use to distinguish stress contrasts. We address this issue in the following section.

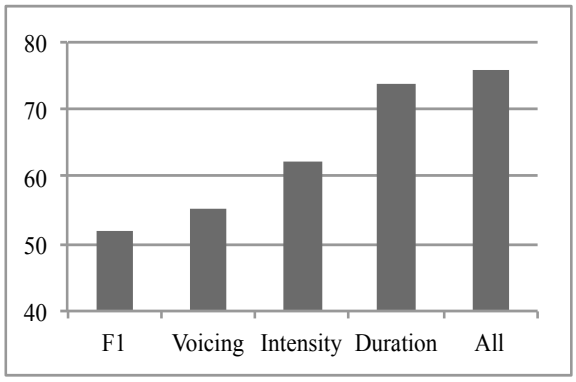


Fig. 3. *Percentages of correct stress pattern classification in Exp. I obtained with different acoustic features in a cross-validation procedure (see text for details).*

3. Experiment II: Perception

In Experiment II, we investigate the extent to which native Spanish listeners can distinguish different stress patterns in phrase-medial unaccented words. We also investigate what phonetic cues are used by listeners to discriminate between stress patterns.

3.1 Method

A random subset of 100 utterances was selected from the 456 utterances analyzed in Experiment I, and used as stimuli in Experiment II. We decided to use only a subset of 100 stimuli in order to avoid fatigue in the participants, which could lead to unreliable results. The experiment consisted of a two-alternative forced-choice task, in which listeners had to classify auditory stimuli in one of two groups according to the tense and person of the verb in the utterance, which, as explained in the previous sections, are distinguished only by the stress pattern of the verbal form (oxytone: 3rd p. sg., past vs. paroxytone: 1st p. sg, present).

Thirteen listeners, all of them native speakers of European Spanish, participated in the experiment. They wore closed headphones and sat in front of a desktop computer in a quiet office. In each trial, two written options appeared on the screen upon presentation of an auditory stimulus: *ÉL/PASADO* 'he/past', in the left part of the screen, and *YO/PRESENTE* 'I/present', in the right part of the screen. The message in the left part of the screen was congruent with an oxytone verb (e.g. ta**pó**), whereas the message in the right part of the screen was congruent with a paroxytone verb (e.g. **ta**po). Participants were asked to press the left or right-arrow key on a computer keyboard according to which message on the screen was congruent with each heard utterance.

3.2 Results

Participants correctly classified the stress pattern of the presented target verbs in 62.9% of the cases. Responses were correct significantly above chance level according to a mixed-effects logistic regression model with stress pattern as dependent variable, the participant's response as the only fixed predictor, and participant as a random factor ($\beta = 1.08$, $z = 9.2$, $p < .0001$). Note, however, that the automatic classification carried out in Experiment I outperformed participants in Experiment II by a considerable margin (75.8 % vs. 62.9%).

An analysis by participants revealed that all of them were able to identify the stress pattern of the verb in the stimuli slightly or moderately above chance level, with percentages of correct classification ranging from 55% to 70%. In an analysis by items ($n = 100$), we found that many of them tended to be correctly classified significantly above chance level (47 above 70%; 27 above 85%), but we also observed a considerable number of items below chance level (16 below 50%; 6 below 30%). Taken together, these results indicate that the stress pattern in the target words could be identified above chance level in most cases, but also that identification errors were common in the data.

We then investigated which cues participants followed when responding to the stimuli in the experiment. We subjected the data to a leave-one-out cross-validation procedure similar to the one employed in Experiment I. This time, we fitted logistic regression models with the participant's response (not the correct answer) as dependent variable, and different combinations of relevant acoustic cues as predictors (i.e. F1, intensity, and duration differentials, consonantal voicing), plus information about participant and verb type.

Figure 4 shows percentages of correct classification of participants' responses obtained with five different models. The model including all acoustic cues achieved 65.5% of correct classification. Interestingly, when only one cue was used, intensity offered the best performance, achieving 59.7% of correct classification. Duration followed closely, with 59.2%. Although in the production data from Experiment I duration clearly offered a better cue to the stress contrast than intensity, these perception data suggest that listeners use both intensity and duration cues in a similar degree when discriminating stress patterns. Regarding F1 and voicing, we found that, as in Experiment I, theses cues played a lesser role than intensity and duration, with percentages of 52.9% and 55.1% respectively.

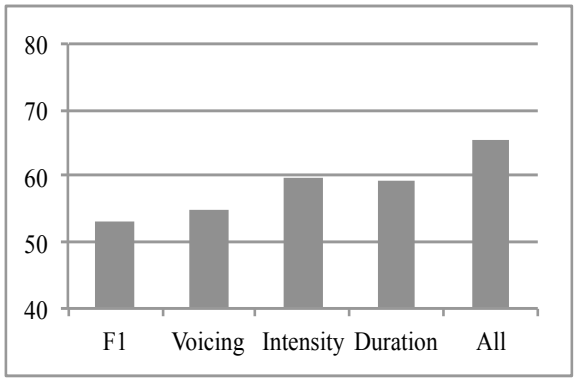


Fig. 4. *Percentages of correct classification of subjects' responses in Experiment II obtained with different acoustic features in a cross-validation procedure (see text for details).*

4. Discussion and conclusion

The previous two sections have presented two experiments aimed at investigating the production and perception of lexical stress contrasts in unaccented words in Spanish. More particularly, we have examined the case of unaccented words in phrase-medial position, a very, and possibly the most, frequent context of deaccenting in Spanish.

Our experiments have shown that, both in production and perception, a contrast in the position of lexical stress is maintained in spite of the lack of prominence-lending pitch cues. In our production experiment, lexically stressed syllables tended to be longer and more intense than their unstressed counterparts. Duration cues allowed for the correct classification of almost three quarters of the data in a cross-validation procedure simulating the discrimination of new unseen data, and they provided a considerably better cue than intensity cues. In perception, we found that all speakers could distinguish the position of lexical stress above chance level, and that, interestingly, their perception of the position of lexical stress was guided as much by intensity cues as by duration cues. Other phonetic cues, such as F1 and the presence of lenitory voicing in consonants also appeared to play a role in distinguishing lexical stress contrasts, but, both in production and perception, in a lesser degree than duration and intensity. Our results for phrase-medial unaccented words are therefore in line with those reported in [8] for parenthetical reporting clauses.

Despite the identified phonetic differences in Experiment I, and the listeners' performance above chance level in Experiment II, it should be noted nevertheless that we observed a considerable amount of phonetic overlap between stress patterns. In production, roughly a quarter of the data could not be classified correctly by a model containing several relevant features such as duration, intensity, voicing and F1; and, in the perception experiment, listeners made identification errors in a two-alternative forced-choice task in 37.1% of the trials. If we take into account that our speech materials were produced in a laboratory setting, and that the perception experiment explicitly required participants to choose between two alternative choices, we may wonder to what extent native Spanish speakers use stress-related phonetic contrasts such as the one examined in this study during online language use. In everyday conversation, it is likely that contextual information is necessary for discriminating between stress patterns, since phonetic cues to stress do not appear to be very robust even in highly controlled data such as ours.

The lack of across-the-board robust cues to stress in Spanish may be related to the low information load that the position of lexical stress has in this language. Except in verb forms, the location of stress in Spanish words is largely predictable: although lexical stress may fall in principle on any of the last three syllables of the word in theory (e.g. ***lám****para* 'lamp', *mam****pa****ra* 'screen', *Pana****má)***, there is an overwhelming tendency for consonant-final words to be stressed on the final syllable, and for vowel-final words to be stressed on the penultimate syllable, with more than 95% of all words in the Spanish lexicon following this rule. Regarding verbs, for which several kinds of minimal pairs exist, ambiguities are likely to be rare in discourse, since these minimal pairs involve verb forms in different tenses (e.g. present vs. past) and with different subjects (e.g. 1st sg. vs. 3rd p. sg.). Robustly conveying information on lexical stress pattern is therefore of little communicative value in Spanish, which may explain the significant amount of phonetic overlap between stress patterns observed in our data.

On the other hand, the fact remains that Spanish words are specified for the position of stress. Even if there are general rules of stress assignment, exceptions must be learned by children during language acquisition. In this regard, the considerable amount of phonetic neutralization that we find in phrase-medial position is not likely to pose a serious challenge to learners with continued exposure to the language, since the position of stress is often conveyed more robustly in other prosodic contexts thanks to the presence of intonational pitch cues [8]. In this sense, the location of lexical stress is not very different from other features that create contrasts between words, and that are sensitive to contextual and realizational factors. For instance, the fact that Spanish /ptk/ consonants are frequently voiced and spirantized in intervocalic position [11, 12, 13], giving rise to occasional ambiguities, does not prevent the /ptk/-/bdg/ opposition from being fully functional in the language. Nevertheless, the fact that Spanish stress contrasts are often lost under specific, but common, conditions, such as the context examined in this study, should be taken into account in descriptions of the prosody of this language.

5. Acknowledgements

The contribution of the first author was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 "INTERACT") to Stephen C. Levinson.

6. References

[1] Nadeu, M. "The effects of lexical stress, intonational pitch accent and speech rate on vowel quality in Catalan and Spanish". Doctoral dissertation, Univ. of Illinois at Urbana-Champaign. 2013.

[2] Ferreira, L. "High Initial Tones and Plateaux in Spanish and Brazilian Portuguese Neutral declaratives: Consequences to the relevance of f0, duration and vowel quality as stress correlates". Doctoral dissertation, Univ. of Illinois at Urbana-Champaign. 2008.

[3] Barbosa, P., Eriksson, A. & Åkesson, J. 2003. "On the Robustness of some Acoustic Parameters for Signaling Word Stress across Styles in Brazilian Portuguese". Interspeech 2013, Lyon, 25-29 August.

[4] Saalfeld, A. 2009. "Stress in the beginning Spanish classroom: an instructional study". Doctoral dissertation, Univ. of Illinois at Urbana-Champaign. 2009.

[5] Ortega-Llebaria M., Hong, G. & Fan, J. "English speakers' perception of Spanish lexical stress: Context-driven L2 stress perception". Journal of Phonetics, 41:186-197, 2013.

[6] Sluijter, A. M. C. & Heuven, V. J. van. "Spectral balance as an acoustic correlate of linguistic stress". J. Acoust. Soc. Am. 100:2471–2485, 1996.

[7] Beckman, M. & Edwards, J. "Articulatory evidence for differentiating stress categories" in P. Keating [Ed], Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III, 7-33, Cambridge University Press, 1994.

[8] Ortega-Llebaria, M. & Prieto, P. "Acoustic correlates of stress in Central Catalan and Castilian Spanish". Language and Speech, 54(1):73–97, 2010.

[9] Torreira, F. & Ernestus, M. "Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish". Phonetica, 69:124–148, 2012.

[10]http://corpus1.mpi.nl/ds/imdi_browser/?openhandle=1839/00-0000-0000-001B-8046-2 (to access a file, select it in the left panel and click on the URL link appearing in the right panel).

[11] Hualde, J. I., Simonet, M. & Nadeu, M. "Consonant lenition and phonological recategorization". Laboratory Phonology, 2:301-329, 2011.

[12] Kim, M. The phonetics of stress manifestation: Segmental variation, syllable constituency and rhythm. Doctoral dissertation, Stony Brook Univ. 2011.

[13] Torreira, F. & Ernestus, M. "Realization of voiceless stops and vowels in conversational French and Spanish". Laboratory Phonology, 2:331-353, 2011.

The effects of stress/accent on VOT depend on language (English, Spanish), consonant (/d/, /t/) and linguistic experience (monolinguals, bilinguals)

Miquel Simonet, Joseph V. Casillas, Yamile Díaz

Department of Spanish & Portuguese
University of Arizona, Tucson, Arizona U.S.A.
{simonet,jvcasill,ydiaz44}@email.arizona.edu

Abstract

This study examines Voice Onset Times of coronal stops in utterance-initial position in two languages. Crucially, the effects of lexical stress (stressed, unstressed syllable) on VOT are analyzed. The study investigates aspirated stops (English /t/), short-lag voiceless stops (English /d/, Spanish /t/) and prevoiced stops (Spanish /d/). Three groups of speakers provide data: English monolinguals, Spanish monolinguals, and proficient Spanish-English bilinguals. The study finds that lexical stress lengthens aspiration (English /t/) and prevoicing (Spanish /d/) but it does not alter significantly short-lag stops (Spanish /t/, English /d/). Monolinguals and bilinguals differ slightly in their phonetic behavior. Implications for gestural coordination as well as for feature theory are discussed.
Index Terms: VOT, stress, Spanish, English, bilingualism

1. Introduction

The present study is concerned with the effects of lexical stress on Voice Onset Times (VOT) of /d/ and /t/ in two languages, Spanish and English. The study examines the productions of monolingual and bilingual speakers and assesses the potential impact of linguistic experience on lexical stress effects for Spanish and English /d/ and /t/.

While both Spanish and English possess a contrast between *fortis* (/p t k/) and *lenis* stops (/b d g/), the phonetic implementation of the contrast differs for the two languages. English presents a contrast between a set of aspirated, voiceless stops (/p t k/) and a set of unaspirated, voiceless stops (/b d g/). The Spanish contrast is that between a set of unaspirated, voiceless stops (/p t k/) and a set of voiced stops (/b d g/) [1]. (Note that this description applies to utterance-initial position because Spanish /b d g/ are regularly spirantized in other positions). According to one view, the phonological opposition in English depends on the feature [spread glottis], which is specified for the *fortis* stops but not for the *lenis* ones; on the other hand, the Spanish contrast depends on [voice], which is specified for the *lenis* stops but not for the *fortis* ones [2].

A large body of research has demonstrated that prosodic structure affects fine-phonetic detail in segments. Experimental studies have shown that consonants in prominent syllables (stressed, accented) are hyperarticulated relative to those in non-prominent syllables (unstressed, unaccented) in a number of languages ([3–7, among others]). Hyperarticulation is captured by articulatory and acoustic measurements. One of the acoustic correlates that manifest hyperarticulation is VOT. In a seminal study, [7] found that VOTs in English aspirated stops (/p t k/) were longer when these consonants were found at the onset of lexically stressed syllables than when they were found at the onset of unstressed syllables. This effect was corroborated, for this language, in [3] and, for phrase-accented syllables in connected speech, in [5]. In sum, for consonants with long-lag VOT, such as English /p t k/, the effects of lexical stress are straightforward, and they seem to be of an additive nature—stress lengthens the aspiration period of aspirated consonants.

Stress effects also have been investigated for consonants with short-lag VOTs, although the results are less straightforward than those for aspirated consonants [8–10]. For instance, [7] reported a trend for English /b d g/ according to which these consonants showed a *shorter* VOT when stressed than when unstressed. On the other hand, [5]'s findings are not in line with those in [7] regarding English /b d g/. In [5]'s study, VOT is lengthened when the consonant bears prominence, and that applies to both /p t k/ (English aspirated stops) as well as to /b d g/ (English short-lag, unaspirated stops). The results of a study of Dutch *fortis* (/t/) and *lenis* (/d/) stops are also informative in this regard [4]. Dutch presents a *fortis-lenis* contrast of the "Spanish type", rather than the "English type"; that is, Dutch /t/ is a voiceless, unaspirated stop—a stop with a short-lag VOT. [4] explores the effects of lexical stress and phrasal accent separately. In this study, VOT length of Dutch /t/ is longest when unstressed and unaccented and shortest when stressed and accented. Thus, the effects of stress on this consonant are subtractive (in terms of VOT), rather than additive, and the effects of lexical stress and phrasal accent are cumulative.

No study seems to have straightforwardly addressed the effects of lexical stress on negative VOT or prevoiced stops [8, 9, cf.]. [4] analyzed the duration of voicing during closure in Dutch /d/. In the data in this study, /d/ always appeared in intervocalic, rather than utterance-initial, position. Although this measurement might differ from lead VOT (since voicing during closure in intervocalic /d/ could be due partially to carryover laryngeal coarticulation), the findings in [4] with respect to /d/ are perhaps relevant—lexical stress was found to affect voicing length. In particular, prominent syllables displayed longer voicing periods during closure than unstressed syllables. The fact remains that, to our knowledge, the effects of stress on prevoiced stops remains to be investigated in depth.

The present study analyzes lexical stress effects on long-lag stops (English /t/), short-lag stops (English /d/, Spanish /t/) and prevoiced stops (Spanish /d/). Thus, we explore a language with a [spread glottis] contrast (English), one with a [voice] contrast (Spanish), and, importantly, we study speakers possessing both features and thus all three stop types (Spanish-English bilinguals). The goal of the present study is to gain a broader understanding of stress effects on different VOT categories in order to begin to grasp the impact of prosody on segments more generally.

2. Methods

In order to collect the production data, we used a delayed shadowing task, widely used in the literature on second-language speech learning (e.g., [11]). In this task, speakers listen to and then repeat out loud target phrases.

2.1. Materials

2.1.1. Target phrases

The materials were a list of words beginning with /d/ or /t/. Half of these words were stressed on the word-initial syllable, and the other half were stressed on the second syllable. Therefore, stops could appear in a stressed syllable or unstressed syllable. There were 40 words total, 20 per language. The words were balanced for Consonant (/d/, /t/) and Stress (stressed, unstressed). There were five word items per design cell: 2 (languages) $\times$ 2 (consonants) $\times$ 2 (stress configurations) = 8 design cells $\times$ 5 items = 40 words. The target words were interspersed among many fillers or distractors (n = 36).

2.1.2. Auditory stimuli

The 40 words were printed out on two lists as a function of language—20 English and 20 Spanish words. Three male native speakers of each language (six 'talkers' in total) were asked to read their corresponding word list out loud. Each 'talker' received a different randomization of the list. Words were produced in utterance-initial position in the carrier sentences '__ is the word' and '__ es la palabra.' In this way, three auditory models were recorded for each word item, one per 'talker.' This amounted to 120 (20 items $\times$ 2 languages $\times$ 3 auditory models or 'talkers') different auditory stimuli to be used in the delayed shadowing task.

2.2. Speakers

A total of 47 volunteers participated in this experiment. All speakers were female, and they were all between 18 and 23 years of age. There were three groups of speakers: (i) a group of Spanish-English bilinguals (n = 19), (ii) a group of monolingual Spanish speakers (n = 22), and (iii) a group of monolingual English speakers (n = 7). The monolingual Spanish speakers were recruited from among the student body of the Universitat de les Illes Balears on Majorca, Spain. These participants are bilingual in Catalan. (Note that there are no reported differences between Catalan and Spanish stop consonants with respect to their VOT [12].) The monolingual English speakers, as well as the Spanish-English bilinguals, are/were undergraduate students at the University of Arizona in Tucson, Arizona. The English speakers consider English to be their native language, and they were exposed to English exclusively in their family circle while growing up. The bilinguals, Mexican-Americans born and raised in Southern Arizona, were brought up by Spanish-speaking families and were schooled mostly in English. They use both languages daily both in the classroom as well as with their friends and relatives. A bilingual profile questionnaire was used to establish the groups [13]. In this report, data from the three groups are analyzed separately.

The Spanish-English bilinguals were recorded in both of their languages. The English monolinguals were recorded in English. The Spanish monolinguals were recorded in Spanish.

2.3. Procedure

The 47 female speakers were asked to listen to and then repeat out loud the auditory stimuli. The English monolinguals heard and produced only the English materials (60 tokens). The Spanish monolinguals heard and produced only the Spanish materials (60 tokens). The Spanish-English bilinguals, however, were recorded in both of their languages—they listened to and repeated the English as well as the Spanish materials (120 tokens). Each speaker heard a different randomization of the stimuli.

Importantly, the bilinguals heard and produced the English and Spanish materials in a single block, in random order. In other words, the Spanish and English auditory stimuli were randomized in one single production session so that the English items were interspersed amongst the Spanish items. This hypothetically maximizes the degree of interlingual transfer of phonetic characteristics for these bilinguals [14], and it thus explores this population in a situation diametrically opposed to the one researched in [15]. Speakers were asked to listen to the entire utterance first ('__ is the word' or/and '__ es la palabra') and then to repeat the entire sequence aloud. In this way, shadowing of the target consonant was slightly delayed relative to the time in which it was heard.

Recordings were made through a head-mounted dynamic microphone (Shure SM10A), a pre-amp (Sound Devices MM-1) and a digital voice recorder (Marantz PMD660). Digitization was at 44.1 kHz, 16-bit. The Spanish monolinguals were recorded in a quiet laboratory at the Universitat de les Illes Balears. The English monolinguals and the Spanish-English bilinguals were recorded in a sound-attenuated booth at the Applied Phonetics Laboratory of the University of Arizona. Speakers were recorded one at a time.

2.4. Analysis

A total of 4,020 word tokens were collected for the present study. The English speakers provided 420 tokens, the Spanish speakers provided 1,320 tokens, and the Spanish-English bilinguals provided 2,280 tokens, 1,140 in Spanish and 1,140 in English. The study considered two within-speaker factors: (i) Consonant: /d/, /t/; and (ii) Stress: first syllable in word is either stressed or unstressed. For one of the speaker groups (Spanish-English bilinguals) a third within-speaker factor was added—Language spoken: Spanish, English.

The target /d/ and /t/ tokens were measured for VOT [16]. VOT is an acoustic metric that captures, for stops, the time lag between the burst (i.e., the release of the articulators after the stop closure) and the initiation of periodicity (i.e., the onset of modal voicing). VOT is positive if the burst occurs before the onset of periodicity (since burst = 0 ms) and negative if it occurs after it. In this study, bursts and voicing onsets were hand-located by exploring spectrographic and sound wave displays. The onset of voicing was marked at a zero-crossing on a sound wave display. Voicing onset in lag VOTs was marked by taking into account the initiation of $F2$ in order to make sure that we labeled the initiation of modal voicing rather than voicing *per se*.

3. Results

3.1. Spanish speakers

The Spanish materials produced by the Spanish speakers were analyzed through a repeated-measures ANOVA with VOT (ms) as the dependent variable and Consonant (/t/, /d/) and Stress

	/d/	/t/
stressed	-69.2 (3.52)	15.3 (1.04)
unstressed	-50.4 (3.51)	16.3 (1.38)

Table 1: Mean (+ SE) VOT (ms) as a function of Consonant (/t/, /d/) and Stress (stressed, unstressed) in the Spanish data.

	/d/	/t/
stressed	21.9 (3.66)	78.4 (6.72)
unstressed	26.3 (4.81)	71.4 (5.89)

Table 2: Mean (+ SE) VOT (ms) as a function of Consonant (/t/, /d/) and Stress (stressed, unstressed) in the English data.

(stressed, unstressed) as within-subjects factors. Individual speaker ($n = 22$) was the error term. The descriptive statistics of these data are shown in Table 1. The statistical test revealed significant effects of Consonant ($F(1, 21) = 511.7, p < 0.001$) and Stress ($F(1, 21) = 123, p < 0.001$), as well as a two-way significant interaction ($F(1, 21) = 135.7, p < 0.001$). For these Spanish speakers, /d/ has a lead, negative VOT with a mean of -59.8 ms and /t/ has a short-lag, positive VOT with a mean of 15.8 ms.

In order to explore the interaction, the main effects of Stress were investigated for the two stop consonants separately; this was done by means of paired t-tests. The alpha level was adjusted accordingly (0.05/2 = 0.025) On the one hand, /d/ was shown to present longer prevoicing when stressed than when unstressed by a factor of 1.373 ($t(21) = -12.02, p < 0.001$). On the other hand, /t/ was found not to be affected by Stress ($t(21) = -1.85, p > 0.05$).

3.2. English speakers

The English VOT data were explored by means of a repeated-measures ANOVA with Consonant (/t/, /d/) and Stress (stressed, unstressed) as main factors. Subject ($n = 7$) was the Error term. The ANOVA yielded significant Consonant effects ($F(1, 6) = 244.9, p < 0.001$), but no significant effects of Stress ($F(1, 6) = 1.27, p > 0.05$). The model revealed a significant two-way interaction ($F(1, 6) = 19.22, p < 0.01$). Descriptive statistics are provided in Table 2.

The interaction was examined by means of two separate paired t-tests, and the alpha level was adjusted accordingly. The interaction was due to the fact that Stress effects were robust for /t/ ($t(6) = 3.8, p < 0.01$), but not for /d/ ($t(6) = -2.7, p = 0.03$). At most, one could claim that a marginal trend exists for /d/ as well, although in the opposite direction. The effects for /t/ were due to the fact that stressed /t/ displays a longer period of aspiration than unstressed /t/ by a factor of 1.098.

3.3. Spanish-English bilinguals

The bilingual VOT data were submitted to a repeated-measures ANOVA with Language (Spanish, English), Consonant (/t/, /d/) and Lexical Stress (stressed, unstressed) as fixed factors. All these were within-subjects factors since bilinguals were recorded in both languages. Subject ($n = 19$) was the error term. The statistical model yielded significant effects of all three main effects: Language ($F(1, 18) = 249.5, p < 0.001$), Consonant ($F(1, 18) = 152.4, p < 0.001$), and Stress ($F(1, 18) = 7.58, p = 0.01$). More relevant was the fact that the model revealed three significant two-way interactions (Consonant by Language, $F(1, 18) = 5.3, p = 0.03$; Language by Stress, $F(1, 18) = 54.5, p < 0.001$; Consonant by Stress, $F(1, 18) = 16.95, p < 0.001$) and a significant three-way interaction ($F(1, 18) = 31.6, p < 0.001$). In order to explore the interactions, the dataset was divided into subsets.

Figure 1: Effects of lexical stress on VOT in English /t/ and /d/ in the bilingual speakers' productions.

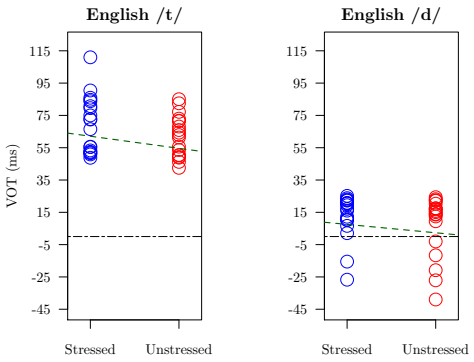
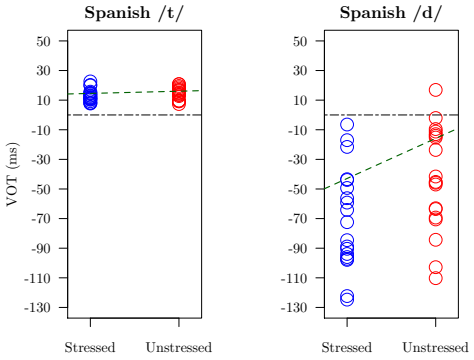


Figure 2: Effects of lexical stress on VOT in Spanish /t/ and /d/ in the bilingual speakers' productions.



Firstly, the effects of Consonant and Stress were examined for the two languages separately. The first ANOVA investigated the effects of these two fixed, within-subjects factors on VOT for the English data only. These data are plotted in Figure 1. This model yielded significant effects of Consonant ($F(1, 18) = 140, p < 0.001$), as well as Stress ($F(1, 18) = 17.3, p < 0.001$). Importantly, there was no interaction between the two factors ($F(1, 18) < 1$). The effects of Stress were due to the fact that stressed consonants had a longer VOT (a VOT further away from zero) than unstressed consonants by a factor of 1.18. Note that both consonants (/t/, /d/) were similarly affected by the effects of stress in this data subset.

The second follow-up ANOVA explored the effects of Consonant and Stress for the Spanish materials only. (See Figure 2.) In this analysis, Consonant and Stress were within-subjects factors and speaker was the error term. The model revealed significant effects of Consonant ($F(1, 18) = 99.25, p < 0.001$) and Stress ($F(1, 18) = 35.38, p < 0.001$), as well as a significant two-way interaction ($F(1, 18) = 28.42, p < 0.001$). The interaction was due to the fact that /d/ was affected by Stress ($t(18) = -5.73, p < 0.001$) while /t/ was not ($t(18) = -1.6, p > 0.1$). In these data, stressed /d/ had longer prevoicing than unstressed /d/ by a factor of 1.63.

4. Discussion and conclusions

The present study has reported on the results of a production experiment in which the effects of lexical stress on VOT length are examined. VOTs were measured for /d/ and /t/ in both Spanish and English. Three groups of speakers were recorded: Spanish and English monolinguals, and Spanish-English bilinguals. The three groups were investigated separately.

Regarding English /d/ and /t/, the study found that the *fortis*, aspirated stop (/t/) was affected by stress while the *lenis* stop (/d/) was not (or marginally so). The effects of stress on English /t/ were additive—lexical stress lengthened aspiration in this consonant. Therefore, it could be said that stress enhances the acoustic difference between the two members of the /d/-/t/ contrast by affecting one of the two.

The situation for Spanish was different from that for English. In Spanish, it was the *lenis* stop (/d/), rather than the *fortis* one (/t/), the one to be impacted by lexical stress. Spanish /t/ was not modulated by stress. The effects of stress on Spanish /d/ were straightforward—stress lengthens prevoicing in this consonant. Once again, therefore, it could be claimed that stress enhances the acoustic opposition between /d/ and /t/ by displacing one of the two members of the contrast.

The findings of the present study regarding the effects of lexical stress on three VOT categories of two different languages are reminiscent of the literatures on the effects of (i) clear speech and (ii) of speech rate on these sound categories. For instance, [17] found that, in clear speech, aspiration is lengthened in English *fortis* stops and prevoicing is lengthened in Croatian *lenis* stops—short-lag VOT categories are largely unaffected by clear speech.

With respect to speech rate effects, studies typically find that as speech rate decreases (i.e., speech becomes slower) VOTs lengthen, but not for all stop consonants [18, 19]. In parallel to the studies on clear speech (and stress effects), the literature on speech rate has shown that aspiration is lengthened in English *fortis* stops while VOT is not impacted (or only minimally so) in stops characterized by short-lag VOTs [18]. In a fundamental cross-linguistic study, [19] researched the effects of speech rate on Thai, French and English stops. French has a contrast of the "Spanish-type"—*fortis* stops are voiceless, unaspirated and *lenis* stops are prevoiced. Thai has a three-way contrast—it possesses sets of (i) voiceless, aspirated stops, (ii) voiceless, unaspirated stops, and (iii) prevoiced stops. The study found that slow speech lengthened aspiration in English *fortis* stops and French *lenis* stops while short-lag VOT (English *lenis*, French *fortis*) stops were not affected. Regarding the case of Thai, the study found that short-lag VOT categories were not modulated by speech rate while the other two VOT categories were—VOTs were lengthened in the predictable directions.

A review of the literature suggests that the effects of speech rate, clear speech and lexical stress on VOT categories are very similar. What these studies have in common is that they find that, in two-way contrasts, short-lag VOT categories are unaffected while the other category is affected. In three-way contrasts, short-lag VOT categories are also the ones that remain largely unmodified. This could be summarized with the statement that short-lag VOT stop categories act as anchors in situations that lead to contrast enhancement, such as slow and/or clear speech and prosodic prominence.

The "contrast-enhancement" perspective, however, has been challenged by a study that investigated the effects of speech rate on Central Standard Swedish [2]. This language has a two-way contrast. It opposes a set of aspirated, voiceless stops with a set of prevoiced stops—a short-lag VOT category is not found in this language. In Swedish, both stop consonants are affected by slow speech. In sum, the same categories that are affected in French, English and Thai are also affected in Swedish—the only difference is that Swedish does not have an "anchor". The reasoning in [2] is that the Swedish contrast is already acoustically large in faster speech and does not need to be enhanced. [2] propose that speech rate, together with clear speech (and lexical stress), affects VOT categories that *are specified for a phonological feature*. In this view, aspirated stops are specified for [spread glottis], prevoiced stops are specified for [voice] and, importantly, short-lag VOT stops are left unspecified. Contrast enhancement is viewed as an artifact of feature modulation.

An alternative view that does not exploit feature theory is one that depends on articulatory gesture coordination. The VOT metric measures the time lag between two acoustic events that stand for two articulatory gestures, a laryngeal and a supralaryngeal gesture. As such, the VOT metric is a measure of gesture coordination. It is possible that speech rate, clear speech and prosodic prominence do not affect gestural coordination patterns for sounds whose two gestures are robustly synchronized (short-lag VOT stops), but do for those that are loosely synchronized (prevoiced and aspirated stops).

An analysis of the bilingual data found the following: (i) Spanish /d/ is robustly affected by lexical stress while Spanish /t/ is not, and (ii) both English consonants are slightly affected by lexical stress. Therefore, for these bilingual subjects, consonants with lead VOT (Spanish /d/) and long-lag VOT (English /t/) are impacted by lexical stress, although the former more so than the latter. This is the same result that was found for the native English- and Spanish-speaking groups. From the two short-lag VOT consonant (Spanish /t/, English /d/), only English /t/ is affected by lexical stress. While English monolinguals did not display robust stress effects on their English /d/, the bilinguals did.

A possible interpretation of the bilingual findings is suggested by the fact that a subgroup of the Spanish-English bilinguals produced English /d/s with some prevoicing—see Figure 1. This could be a transfer effect from Spanish /d/ for this subgroup. Since negative VOT in prevoiced stops is indeed affected by lexical stress (Spanish /d/, for instance) it is perhaps not surprising that English /d/, as produced by this particular group, is somewhat modulated by stress. Alternatively, English, but not Spanish, *lenis* stops, in general, could be impacted by lexical stress in this population. A strong interpretation of the "featural specification hypothesis" of [2], according to which only consonants with specified features are affected by prosodic modulations, is therefore not possible for the bilingual group. It could be concluded that the Spanish-English bilinguals in this study posess a phonological system consisting of [voice] (Spanish /d/), ∅ (Spanish /d/), and [spread glottis] (English /t/). The nature of the English *lenis* stop would deserve more attention. If [voice] were to be chosen as the featural specification for English /d/, the question arises as to why the phonetic facts for this consonant differ from those for Spanish /d/.

5. Acknowledgments

The authors wish to express their gratitude to the following: Mark Amengual, Melinda Porta, Olivia Obeso, Miquel Llompart, and the 47 volunteers who participated in the study. The authors are also grateful to three anonymous reviewers, and to the SP-7 audience, for constructive criticism.

6. References

[1] Rosner, B and López-Bascuas, L and García-Albea, J and Fahey, R, "Voice-onset times for Castilian Spanish initial stops," *Journal of Phonetics*, vol. 28, pp. 217–224, 2000.

[2] Beckman, J and Helgason, P and McMurray, B and Ringen, C, "Rate effects on Swedish VOT: Evidence for phonological overspecification," *Journal of Phonetics*, vol. 39, pp. 39–49, 2011.

[3] Cho, T and Keating, P, "Effects of initial position versus prominence in English," *Journal of Phonetics*, vol. 37, pp. 466–485, 2009.

[4] Cho, T and McQueen, J, "Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress," *Journal of Phonetics*, vol. 33, pp. 121–157, 2005.

[5] Cole, J and Choi, H and Kim, H and Hasegawa-Johnson, M, "The effect of accent on the acoustic cues to stop voicing in Radio News Speech," in *Proc. of the International Congress of Phonetic Sciences*, 2003.

[6] de Jong, K, "The suppraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *The Journal of the Acoustical Society of America*, vol. 97, pp. 491–504, 1995.

[7] Lisker, L and Abramson, A, "Some effects of context on Voice Onset Time in English stops," *Language and Speech*, vol. 10, pp. 1–28, 1967.

[8] M. Castañeda, "El VOT de las oclusivas sordas y sonoras españolas." *Estudios de Fonética Experimental*, vol. 2, pp. 91–110, 1986.

[9] D. Poch, "Caractérisation acoustique des occlusives de l'espagnol: Le problème du VOT." *Revue de Phonétique Appliqueé*, vol. 77, pp. 477–490, 1985.

[10] M. Troya, "El VOT de las oclusivas sordas en la norma culta de Las Palmas de Gran Canaria." *Boletín de Lingüística*, vol. 17, pp. 31–38, 2005.

[11] Guion, S, "The vowel systems of Quichua-Spanish bilinguals," *Phonetica*, vol. 60, pp. 98–128, 2003.

[12] Amengual, M, "Interlingual influence in bilingual speech: Cognate status effect in a continuum of bilingualism," *Bilingualism: Language and Cognition*, vol. 15, pp. 517–530, 2012.

[13] Birdsong, D and Gertken, L and Amengual, M, "Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism," 2012.

[14] Olson, D, "Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production," *Journal of Phonetics*, vol. 41, pp. 407–420, 2013.

[15] Magloire, J and Green, K, "A cross-language comparison of speaking rate effects on the production of voice onset time in English and Spanish," *Phonetica*, vol. 56, pp. 158–185, 1999.

[16] Lisker, L and Abramson, A, "Crosslanguage study of voicing in initial stops," *Word*, vol. 20, pp. 384–422, 1963.

[17] R. Smiljanic and A. Bradlow, "Stability of temporal constrats across speaking styles in English and Croatian," *Journal of Phonetics*, vol. 36, pp. 91–113, 2008.

[18] J. Allen and J. Miller, "Effects of syllable-initial voicing and speaking rate on the temnporal characteristics of monosyllabic words," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2031–2039, 1999.

[19] R. Kessinger and S. Blumstein, "Effects of speaking rate on voice-onset time in Thai, French and English," *Journal of Phonetics*, vol. 25, pp. 143–168, 1997.

Is Syllable Stress Information Robust for ASR in Adverse Conditions?

Bogdan Ludusan, Stefan Ziegler, Guillaume Gravier

CNRS - IRISA Rennes, France

bogdan.ludusan@ens.fr, {stefan.ziegler, guillaume.gravier}@irisa.fr

Abstract

This paper presents a study on the robustness of stress information for automatic speech recognition in the presence of noise. The syllable stress, extracted from the speech signal, was integrated in the recognition process by means of a previously proposed decoding method. Experiments were conducted for several signal-to-noise ratio conditions and the results show that stress information is robust in the presence of medium to low noise. This was found to be true both when syllable boundary information was used for stress detection and when this information was not available. Furthermore, the obtained relative improvement increased with a decrease in signal quality, indicating that the stressed parts of the signal can be considered islands of reliability.

Index Terms: speech recognition, prosody, syllable stress, noise

1. Introduction

Prosodic information has already been used successfully in large vocabulary speech recognition (e.g. [1, 2]). Prosodic information (duration, pauses, F0, etc) was integrated in automatic speech recognition (ASR) systems both at the acoustic and language model level, being posited that prosodic features are robust to noise and unaffected by channel condition [1].

Among the major prosody components, stress seems to present several characteristics which are particularly helpful for speech recognition tasks under different conditions. Studies examining the role of stressed syllables showed that they provide salience in terms of their acoustic attributes, they are less likely to suffer phonological modification or to be misinterpreted and they are detected more consistently than unstressed syllables in *noisy* environments [3].

Indirect evidence supports the relevance of stress information in noisy conditions, not only for humans, as shown by psycholinguistics studies [4] but also for machines [5]. A human perception in noise study [4] examined the intelligibility of speech at very low signal-to-noise ratios (SNR), by either masking or unmasking the stressed syllables. It showed that the SNR required to identify the consonants of the unstressed syllables increased when the stressed syllables were masked and it decreased when the stressed syllables were unmasked. The author concluded that the listener relies on prosody to achieve robust speech understanding and that the information in the stressed syllables helps predict the neighbouring unstressed syllables. Further evidence can be found in an ASR study investigating the effects of phonetic information reduction on recognition performance [5]. When phone identity was substituted with manner feature, performance dropped in both clean and noisy conditions, but no significant difference was observed when phone identity only inside unstressed syllables was replaced by manner information. This suggests that the information carried by the stressed syllables has a higher importance for speech recognition in noisy environments than the one present in non-stressed syllables.

Based on the fact that humans exploit stress information not only under normal acoustic conditions, but also in the presence of noise, and that this information is salient for them, we are interested in investigating whether adding stress information to ASR in adverse conditions would be useful. One would expect it, as stressed syllables display higher energy than non-stressed syllable and in noisy environments they would exhibit a higher SNR than the neighbouring syllables, thus helping to their recognition.

Stress information has been used before in several speech recognition systems [6, 7, 8, 9, 10]. The systems employed different methods to add this new information to the recognition process: at the lexicon level [7, 8], as a separate model in the decoding process [6, 9], or to guide paths during search [10]. Most of these studies reported statistically significant improvements when including stress, but the role of stress was examined exclusively under normal acoustic conditions. To our knowledge, there is only one study in the literature which reported results for speech under adverse conditions [10]. In that study, experiments both on normal and strong accented speech were performed and the same level of improvement was obtained for the two types of speech, when stress information was used. A close analysis of the system showed that the integration of stress information improved speech recognition due to its interaction with the pruning process, by helping prune away some of the wrong hypotheses.

The current paper builds upon a previous mentioned study [10] and aims at the following two aspects: to enlarge the investigation of stress robustness for ASR in adverse condition to a new case (additive noise) and to explore the behaviour of the system when there is no syllable boundary information available for the computation of the stress score. Because we wanted to examine only the effect of stress on the recognition performance, the same system was used in the clean and noisy condition, with no speech enhancing pre-processing. Further details on the recognition system employed and on the stress detection procedure can be found in section 2. For the recognition experiments presented in section 3, we used white and pink noise at various SNRs and we tested two conditions based on whether syllable boundary information was available for stress detection or not. We have chosen to use coloured noise in this study as a starting point in our research, an initial test, but we envisage our future work to include more complex types of noise.

2. System Presentation

The system used in this paper is composed of two components: a stress detection procedure and a speech recognizer. The first component computes syllable-level stress scores which are sub-

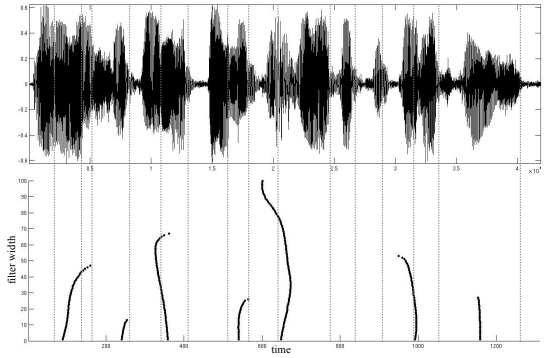


Figure 1: Speech waveform and corresponding rhythmogram.

sequently integrated in the search process of the ASR system.

2.1. Stress Detection

For the computation of the stress score, an unsupervised method based on the auditory primal sketch (APS) [11] was used. The APS is a model of rhythm perception which seeks to identify strong acoustic events present in speech, in a similar way to edge detection in vision theory [12]. In order to accomplish this, the speech signal is filtered with a bank of Gaussian filters with different filter widths. The peaks of the obtained functions are stacked onto each other, with the peaks of the minimum filter width at the bottom. The stacking of the maxima forms contiguous lines which we call "events" (as exemplified in Figure 3). Thus, a hierarchical representation of speech, called a rhythmogram, is obtained. The rhythmogram contains the time on the abscissa and the event height on the ordinate and can be condensed into a two-line vector, containing the time instants of the events on the first line and their corresponding values on the second line.

The procedure used for obtaining the rhythmogram, as well as the values of the parameters needed for its computation, are the one proposed in a previous study [13]. It consists of the following steps:

1. resample at 500 Hz,
2. perform full wave rectification,
3. take the cubic root, to model the ears' loudness function,
4. apply one hundred logarithmically-distanced Gaussian filters,
5. stack the maxima of the obtained function in a 2-D representation, with time on the x-axis and filter width on the y-axis.

Once the rhythmogram is computed, information about syllable boundaries is needed in order to obtain a syllable-level stress score. Then, a search within each syllable is performed and the value of the highest event is taken as the stress score. For the experiments conducted in section 3.2 we had this information available, while the experiments presented in section 3.3 made use of an approximation. Further details will be given in the respective sections.

Figure 1 illustrates the waveform of a sentence from the corpus used in the experiments along with its corresponding rhythmogram. Each point of the rhytmogram is associated to a time instant (t) and a filter width (i) and it represents a peak at time t in the function obtained with the ith filter. For this particular example we can observe that for the minimum filter width (bottom of lower panel) the function had 7 maxima, for the 40th filter 4 maxima, while applying the filter with the highest width returned a function with one maximum (top part of lower panel). By plotting these points in space, we obtain the contiguous lines observed in the figure (the events). In order to compute the stress score we determine to which syllable each event belongs, based on the start time of the event, obtained with the lowest filter width. The stress score will be equal to the number of points which form the event, in this case a stress score of 0 is obtained for the first syllable, a stress score of 47 for the second one, etc.

2.2. Speech Recognition System

For all the experiments we used a two-pass recognition system [10]. The recognizer, produced a word graph after the first pass, graph which will be rescored using more complex acoustic models in the second pass. It uses in the first pass word-internal triphone acoustic models with 4,019 distinct states and 32 Gaussians per state and word trigrams as language model. The rescoring pass has 4-grams as language model and cross-word triphone models with 6,000 states and 32 Gaussians each.

In order to integrate the stress information in the recognition system, the Viterbi search was modified as shown in Equation 1.

$$Q(j,t) = \max_i Q(i,t-1) + log(a_{ij}) + log(b_j(y_t)) + \\ + str(t) \cdot R \quad (1)$$

The first three terms are also present in the classical Viterbi decoding: $Q(j,t)$, the score of the path up to state j at time t, $log(a_{ij})$, the transition probability between states i and j, and $log(b_j(y_t))$, the observation probability of y_t when in state j. The last term is a product between $str(t)$, the stress score of the syllable at time t (taking values between 0 and 1), and R, a weighting factor which represents the contribution of the stress information to the decoding process. The value of R was determined by optimizing the recognition performance on the development set.

Thus, by adding the new term in the search equation, the decoding procedure reinforces all the phonemes belonging to the stressed syllables with a value proportional to the syllable stress score. This implementation choice was made based on the fact that stressed syllables are more stable and are distinguished better [3] and, by giving them a higher weight in the search process, improvements can be obtained [10].

3. Experiments

Speech recognition experiments were conducted on a corpus of broadcast news to which coloured noise, at different SNRs, was added. We investigated the role of syllable boundary information for the computation of the stress score as well as the effect of optimizing the stress weighting factor for each noise level. For each of the experiments, the recognition performance was evaluated by considering the reduction brought in terms of word error rate.

3.1. Materials

The materials used for the experiments presented here are the files of the ESTER 2 evaluation campaign corpus [14]. The corpus consists of mainly broadcast news recordings from French radio stations, although it also contains some more spontaneous radio shows as well as recordings from French-speaking African radio stations which exhibit strong accents.

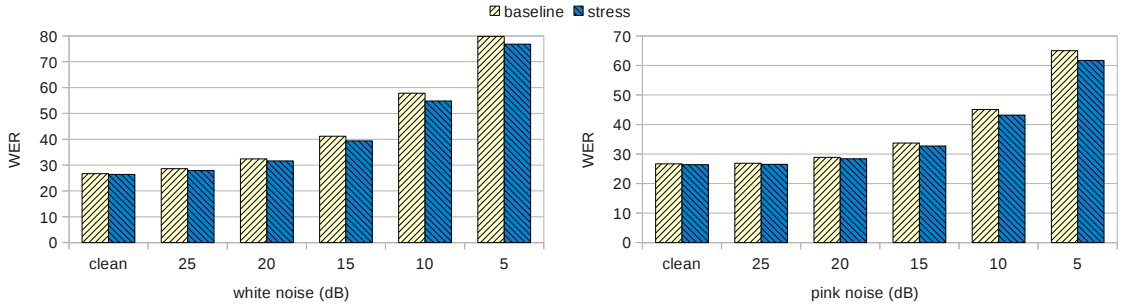


Figure 2: The WER obtained for various SNRs of additive white noise (left panel) or additive pink noise (right panel).

The entire training set, approximately 180 hours of data, was used for the estimation of the acoustic models. Its corresponding transcriptions, along with articles from newspapers, were employed in the computation of the language model. The rest of the corpus forms the development set, used for tuning the system parameters, and the test set, on which the performance of the system was evaluated. These subsets contain circa 6 and 7 hours of recordings, respectively.

For the experiments, two types of additive noise were added to the speech files: white noise and pink noise, ranging from 25 to 5 dB SNR. The white noise was generated using the MATLAB function *wgn*, while the pink noise was obtained by filtering the previously generated white noise [15]. For each SNR, the gain of the filter was determined in order to obtain the desired SNR after the filtering operation. The noise was added only to the files belonging to the development and evaluation sets, the acoustic models having been previously trained with the original training set files.

3.2. Experiment 1

As mentioned in section 2.2, in the first experiment we had knowledge of syllable boundaries for computing the syllable stress score. The boundaries were determined by force aligning the data at the phoneme level and then applying French syllabification rules [16]. Although the quality of the syllabification was not as good as when manual syllables would have been used, this was the best available option for computing a syllable-level stress score. The search for the parameter R in Equation 1 was performed only on the clean data and the obtained value was used in all experiments, regardless of the SNR. The stress score was determined at each step from the noisy data.

The recognition results for various levels of noise are presented in Figure 2: for white noise (WN) in the left panel and for pink noise (PN) in the right panel, with the reported measure being the word error rate (WER). The results obtained with the baseline system are represented by the lighter coloured columns, while those of the recognizer employing stress knowledge are represented by the darker coloured columns. The clean condition is illustrated in both panels for an easier comparison. It can be observed that the importance of stress information in the recognition process increases with the decrease in SNR.

$$WER_{rel} = \frac{WER_{stress} - WER_{base}}{WER_{base}} \quad (2)$$

Next, we used the relative WER (see Equation 2) to compare the performance at different SNRs of the speech signal. As can be seen in Figure 2, the results of the system using stress information were always better than those of the baseline. Thus, we show in Table 2 the values of $abs(WER_{rel})$, which represent the relative improvement (in %), at each SNR, with respect to the baseline. A Wilcoxon signed rank test was used to determine the statistical significance of the results. All the differences were found to be *significant* at the $p < 0.001$ level, except for PN25 ($p < 0.01$).

Noise	Clean	25 dB	20 dB	15 dB	10 dB	5 dB
WN	1.1	2.5	2.5	4.4	5.2	3.8
PN		1.5	1.7	3.0	4.2	5.1

Table 1: *Relative improvement (in %) when syllable boundaries are known.*

The same tendency is observed for both types of noise: the relative improvement increases with the decrease in signal quality. This suggests that the information in the stressed parts tends to have a higher weight on the recognition process as conditions deteriorate. Also, because we were using a value for R determined on clean data it also proves the robustness of the information added. The results obtained in this experiment can be considered as an upper bound for the improvement brought by adding stress information, as it uses almost ideal syllable boundary information.

3.3. Experiment 2

In the second experiment no syllable boundary information was used. Instead we define a time interval around each rhythmogram event which will act as a "pseudo-syllable", i.e. the whole region will be considered as one entity and it will be assigned a stress score equal to the height of its corresponding rhythmogram event. The gaps between the obtained regions will be assigned a zero stress score and, thus, they will have no effect on the decoding process. Similarly to the first experiment, the value of R was determined on the clean development set, while the stress score was obtained from the noisy data.

The "pseudo-syllable" approach is illustrated in Figure 3. The curly braces under each event correspond to the size of the region considered, while the intervals delimited by dashed lines and labeled PS_n represent the final entities. Their size might be lower than the chosen region size, due to events being too close to each other, or too close to the beginning or the end of an utterance. In order to choose the size of the region we took a look at the average size of the dev set syllables having non-zero stress scores. The value obtained, 197 ms, can be compared to the 141 ms average length of syllables having stress scores equal to zero. By rounding the 197 ms to the closest odd number of frames (the central frame and an equal number of frames on

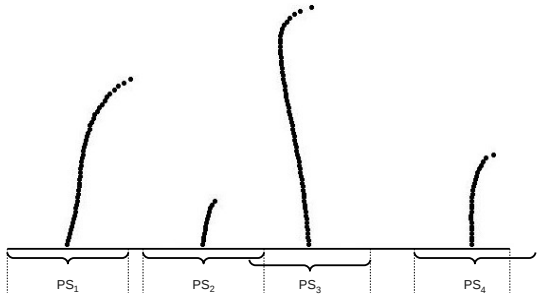


Figure 3: The approach used for hypothesizing syllables, when syllable boundaries are unknown.

each side) we get regions 19 frames wide. This value was used in the present experiment.

Table 2 shows the results given by the recognizer when the previously explained approach is employed. All differences except for PN25 and WN20, were found to be statistically *significant* (WN25 $p < 0.05$, WN15-WN5 $p < 0.01$, PN20 $p < 0.01$, PN15 $p < 0.05$, PN10-PN5 $p < 0.001$). Although the performance increase obtained is not as important as in in the case when syllable boundaries were known, a similar trend in the results can be observed.

Noise	Clean	25 dB	20 dB	15 dB	10 dB	5 dB
WN	0.8	0.7	0.3	1.2	1.4	1.0
PN		0	0.7	0.6	1.1	1.9

Table 2: *Relative improvement (in %) when syllable boundaries are unknown.*

3.4. Experiment 3

As a final test, we examined the effect of parameter optimization on the recognition performance. The experiment was run only on the data containing additive white noise and an optimum value of the R parameter was obtained for each SNR, on the dev set. Both conditions used in the previous two experiments, with or without knowledge of syllable boundary information, were considered and the results obtained are presented in Table 3.

Condition	25 dB	20 dB	15 dB	10 dB	5 dB
Exp 1	2.5	2.5	3.2	5.5	5.0
Exp 2	1.1	1.9	3.4	3.3	0.5

Table 3: *Relative improvement (in %) for an optimized value of the R parameter, in the case of white noise.*

Comparing the results in the first row with those illustrated in section 3.2, one can see that they are quite *similar*, the only case where a significant improvement was obtained is for the 5 dB level ($p < 0.001$). Furthermore, the performance advantage obtained in Experiment 1 for 15 dB is statistically significant ($p < 0.01$). This shows that, when the syllable boundaries are known, the values of R obtained in clean conditions give a good performance also for noisy speech. And, by not needing to optimize R for each SNR, we avoid having to estimate the noise level prior to the actual recognition process.

For the conditions described in Experiment 2, the optimization of the parameter R gives instead *significant* improvements for all noise levels (WN25, WN5 $p < 0.05$, WN20-WN10 $p < 0.001$). An interesting results was obtained for the 5 dB WN condition: although the performance was lower in all the experiments conducted, here the difference was the highest. This might be due to the length of the region considered for our syllable approximation. In case of low SNR it would probably more appropriate to consider smaller regions. While this will decrease the effect that stress has on recognition, by reinforcing a smaller area it is more likely that this area will fall inside the boundaries of the actual syllable and it will not introduce any other errors.

4. Conclusions

In this study we investigated the robustness of stress information in the recognition process, when speech is corrupted by additive noise. To our knowledge this is the first study in the literature aimed at investigating this issue. Using white and pink noise in the experiments conducted, we have observed the same behaviour in both cases: for medium to low SNRs, higher relative improvements are obtained with the increase in the noise level, when stress knowledge is integrated into the recognizer. These results support the view that stressed syllables represent the reliable regions of the speech signal and that the information they carry is important for speech recognition. Besides agreeing with the role given to stress by psycholinguistic studies [3, 4] as well as to the indirect evidence coming from other ASR studies [5] the results of this investigation also encourage the use of such information in speech recognition.

Stress information is robust in the presence of additive noise, especially when syllable boundaries are known for the computation of the stress score. This finding is supported by the small difference observed when the R parameter was optimized for each noise level, compared to the case when the value for R obtained on the clean data was used in the experiments. While improvements are obtained also when a syllable approximation is used instead of the actual syllables, they are significantly lower and depend more on the value of the R parameter, for different noise levels. This might suggest that the approximation used is not suitable for calculating the stress score and that a new approach should be sought. A possible alternative would be to hypothesize syllable boundaries based on the transcription obtained after the first pass of the recognizer. Unfortunately this approach would then limit the use of stress information to the second step only and it was shown in [10] that a big part of the improvement brought by stress information was due to its use in the first pass.

In this work we examined the role of stress information only in the case of coloured noise, but, in the future, we plan to extend the study to also include speech in the presence of competing talkers. Further lines of research to follow include the search for a better syllable approximation or the use of stress only in the rescoring step, as syllable information can be extracted from the word graph produced by the first pass. Also, we used a 19-frame representation in the present work, but a search for the optimum size of the region to be considered might give better results.

5. Acknowledgements

This work was partially funded by the Agence Nationale de la Recherche, through the ASH project. The first author is now affiliated with LSCP - EHESS/ENS/CNRS, Paris.

6. References

[1] D. Vergyri, A. Stolcke, V. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. of IEEE ICASSP 2003*, 2003, pp. 208–211.

[2] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavarria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus," *Speech Communication*, vol. 46, pp. 418–439, 2005.

[3] S. Mattys, "The use of time during lexical processing and segmentation: A review," *Psychonomic Bulletin and Review*, vol. 4, pp. 310–329, 1997.

[4] P. Divenyi, "Humans glimpse, too, not only machines (hommage à Martin Cooke)," in *Forum Acusticum 2005*, 2005, pp. 1533–1538.

[5] E. Fosler-Lussier, A. Rytting, and S. Srinivasan, "Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance," in *Proc. of INTERSPEECH-2005*, 2005, pp. 1249–1252.

[6] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain," in *Proc. of EUROSPEECH-2001*, 2001, pp. 2761–2764.

[7] H. van den Heuvel, D. van Kuijk, and L. Boves, "Modelling lexical stress in continuous speech recognition," *Speech Communication*, vol. 40, pp. 335–350, 2003.

[8] R. van Dalen, P. Wiggers, and L. Rothkrantz, "Lexical stress in continuous speech recognition," in *Proc. of INTERSPEECH-2006*, 2006, pp. 2382–2385.

[9] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," in *Proc. of INTERSPEECH-2007*, 2007, pp. 1813–1816.

[10] B. Ludusan, S. Ziegler, and G. Gravier, "Integrating stress information in large vocabulary continuous speech recognition," in *Proc. of INTERSPEECH-2012*, 2012.

[11] N. Todd, "The auditory "primal sketch": A multi-scale model of rhythm grouping," *Journal of New Music Research*, vol. 23, pp. 25–70, 1994.

[12] D. Marr, *Vision*. New York: Freeman Education, 1982.

[13] B. Ludusan, A. Origlia, and F. Cutugno, "On the use of the rhythmogram for automatic syllabic prominence detection," in *Proc. of INTERSPEECH-2011*, 2011, pp. 2413–2416.

[14] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French broadcasts," in *Proc. of INTERSPEECH-2009*, 2009, pp. 1149–1152.

[15] J. Smith, *Spectral Audio Signal Processing*. W3K Publishing, 2011, ISBN: 978-0-9745607-3-1.

[16] F. Dell, "Consonant clusters and phonological syllables in French," *Lingua*, vol. 95, pp. 5–26, 1995.

Prosodic Differences between Taiwanese L2 and North American L1 speakers—Under-differentiation of Lexical Stress

Chiu-yu Tseng & Chao-yu Su

Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

Assuming that categorical differentiation is major acoustic characteristics of English lexical stress through binary instead of more complex 3-way distinction, we investigated lexical stress in broad and narrow focus positions and found how binary distinction is achieved by the concomitancy of secondary stress defined by its position and distance in relation to primary stress. Similar results are found in broad (sentence initial) and narrow focus as well. These results suggest that binary categorical contrast is the optimal choice while differentiation is dependent on robust contrast patterns in the speech signal. Comparison between Taiwanese L2 and North American L1 speakers revealed how L2 speakers' realization of the binary opposition is of a lesser degree. The results explain why L2 speech is less differentiable and not as expressive.

Index Terms: English stress, primary, secondary, binary, contrast pattern, differentiation

1. Introduction

Earlier studies of L2 or foreign accents that concentrated mainly on segmental features [1, 2, 3], however, more recent shift to prosodic features has led to the discovery that prosodic deviations have as much an effect on the intelligibility and comprehensibility of L2 speech, and contribute significantly to perceived foreign accent as well. For example, Field [4] showed that shift of lexical stress has a strong effect on the intelligibility of native vs. non-native speech group. Mixdorff et al. [5] further showed how the speech rhythm in L2 Vietnamese (tone and syllable-timed) and Japanese (pitch accent and mora-timed) Australian English differed from L1 speakers; both Vietnamese and Japanese speakers produce longer and more equal syllable durations than Australian English speakers. However, we see instead the more even syllable duration of L2 speech not as a rate issue but as lack of the required long/short contrast for categorical stress differentiation. Our hypothesis is that patterns of robust contrast in the speech signal are directly correlated to linguistic categorical contrasts, while lack of or under-differentiation is a major feature of L2 speech.

In the case of English lexical stress, while it is necessary for L2 speakers to learn where the stressed syllable of a word is, it is as important for them also to learn how to maintain the contrast patterns between stressed/unstressed syllables. In other words, even when the correct syllable is stressed, insufficient contrast degree would still result in less differentiable perception and impair intelligibility. Based on the rationale of contrast robustness, we studied the contrast patterns of English lexical stress of all three acoustic correlates the F0, duration and amplitude between English stressed/unstressed syllables produced by L1 American vs. L2 Taiwan Mandarin (TM) speakers and found that in L1 English the most significant contrast is in F0 (pitch contrast), not duration (rhythmic contrast). And as expected, contrast by lesser degree is found in both F0 and amplitude in TM L2 English. TM L2 speakers were able to maintain similar rhythmic contrast as L1speakers do but still sounds flatter and foreign due to lack of pitch contrast [6]. Our results of TM L2 speakers differ considerably from Vietnamese and Japanese L2 speakers in [5], thus reduces possible generalization of how syllable-timed L1 may affect L2 English rhythm in general.

In a subsequent study of stress contrast, we further discovered that the 3-way primary/secondary/tertiary stress contrasts as English lexical stress is defined were not found in both L1 and L2 speech [7]. Instead, significant difference is only found in a 2-way contrast between stressed/unstressed syllables in both speaker groups, and again TM L2 speech exhibited less degree of contrasts. In addition, we found no significant contrast between secondary and tertiary stress across L1 and L2 speech and further discovered that 6 of the 20 tested words differ in where the secondary stress should be in three dictionaries consulted. Nevertheless, our results do suggest that 2-way contrast seem most optimal. This has lead us to further hypothesize that (1) in forming the optimal 2-way contrast the role of secondary stress is a concomitant one; its varied realization a surface phenomenon and should predictable. (2) The same rationale of maintaining the optimal 2-way contrast can also be applied to other prosodic categories such as broad sentential focus and narrow focus induced by context or syntactic structure. Prosodic contrasts of larger sized units are even coarser ones to facilitate long distance prediction, providing contextual and pragmatic information that distinguishes speech from text most significantly.

In the following study, we will analyze English secondary stress under the assumption that its concomitancy is dependent on two factors: (1) its linear order (before or after) the primary stress and its distance from it as well. Namely, if a secondary stress appears immediately BEFORE the primary one it is likely to be assimilated to the target primary stress. However, if it appears AFTER the primary stress then it is likely to be assimilated to the following tertiary stress and be reduced to a lower level in order to create the robust contrast patterns. (2) However, if there is more distance between the primary and secondary stress, such as BEFORE but intercepted by a tertiary stress, then the secondary stress stands more chance to be more differentiable from either the primary or the tertiary counterparts.

In the following analysis, we will compare English secondary stress in different positions in a word, in sentence/broad focus position and in narrow focus position in relation to categorical stress differentiation.

2. Method

2.1. Speech Materials

A subset of the AESOP-ILAS (Asian English Speech cOrpus Project—Institute of Linguistics Academia Sinica) corpus was

used for the present study. AESOP is a multinational collaboration whose aim is to build up English speech corpora across Asia that would represent the varieties of English spoken in that region while ILAS is part of the consortium that specifically collects L2 English of Mandarin L1 speakers in Taiwan. The materials used here are 20 frequently used words from 2-, 3- and 4-syllables categorized according to syllabicity and stress type: (1) 2-syllable initial stress 2, (2) 3-syllable initial stress, (3) 3-syllable medial stress, (4) 3-syllable final stress, (5) 4-syllable initial stress, (6)4-syllable medial 1 stress, (7) 4-syllable medial 2 stress, (8) 4-syllable final stress, (9) left-headed compounds (e.g orange juice), (10) right-headed compounds (e.g. afternoon). The chosen words are money, morning, white wine, hospital, apartment, department, tomorrow, video, overnight, January, supermarket, elevator, available, Japanese, afternoon, misunderstand, information, experience, California and Vietnamese. These words are then embedded in two conditions: (1) in a fixed sentence-medial broad-focus position two words removed from any phrase boundary, i.e., "I said OVERNIGHT five times." for the purpose of baseline comparison as well as broad focus. (2) As elicited narrow focus to create phrasal and sentential prominence in broad and narrow focus positions. For example, *Context: Will 3-day delivery be fast enough?* Reply: "No. We need OVERNIGHT delivery" where the provided context requires the answer to disambiguate. As illustrated, the same target word in the previous broad focus position would now re-appear as narrow focus. At the same time, the sentence-initial word "we" may receive sentential prominence, thus providing both narrow focus and sentential prominence in the same sentence.

Speech data were recorded by trained proctors in quiet rooms directly into a laptop computer, using a recording platform developed specifically for AESOP. Experimental sentences and context were preloaded and appeared individually on a computer screen. Participants wore head-mounted Sennheiser PC155 microphones positioned 2 cm away from their mouths; they were instructed to speak naturally at a normal rate and volume. The speech data of a total of 25 speakers were analyzed: 11 L1 North American English speakers (5 male and 6 female), 16 Taiwan L2 speakers (8 male and 8 female)

2.2. Data Analysis

Prosodic contrast is presented by F0, duration and intensity using Z-score normalization by each sentence first. In order to extract F0 due to lexical stress without intonation effect for subsequent analysis, a straight line with minimal distance (RMSE) to original F0 contour is derived to represent intonation and subtracted, the residual is regarded as F0 without intonation effect. In turn, duration extraction is also refined to remove the effect of inherent segmental duration and boundary lengthening using a multi-layered normalization method shown below[8], in which *factor1* represents information at the segmental level, *factor2* represents respective syllable position within the word (to remove word-final boundary lengthening effects), and ε_i represents all other unpredictable values. Extracted values μ_i thus represent duration values which have been normalized for inherent segmental duration and boundary effect:

$$x_i = \mu_i + factor_1 + factor_2 + ... + \varepsilon_i$$

3. Results and Discussion

3.1. Contrast patterns--primary vs. secondary stress

This study examines prosodic contrast of secondary stress by linear order (before or after) to the primary stress and its distance from primary stress to test (1) if a secondary stress is assimilated to the target primary stress when it immediately precedes the primary stress; if a secondary stress immediately follows the primary stress it is reduced to tertiary stress in order to create more robust contrast patterns. (2) If there is more distance between the primary and secondary stress due to intermediate tertiary stress, then the secondary stress is more differentiable from either the primary or the tertiary counterparts. Furthermore, this study tests if these characteristics could help distinguish L1 and L2 English.

3.1.1. F0 without intonation effect—pitch contrast

Figure2 shows F0 of secondary stress without intonation effect by L1/L2, linear order (preceding or following) regarding the primary stress and its distance from the primary stress. When the primary stress precedes the secondary stress, no significant difference between L1 and L2 English is found. However, the most distinct difference between L1 and L2 English is found when the secondary stress is immediately BEFORE the primary stress (1-syllable distance). The F0 of primary stress in L1 English is higher than secondary stress while L2 English lacks the same contrast pattern and highly varied. When secondary stress appears BEFORE primary stress (2-syllable distance), both L1 and L2 English show how secondary stress "stands out" and becomes more differentiable from the primary stress.

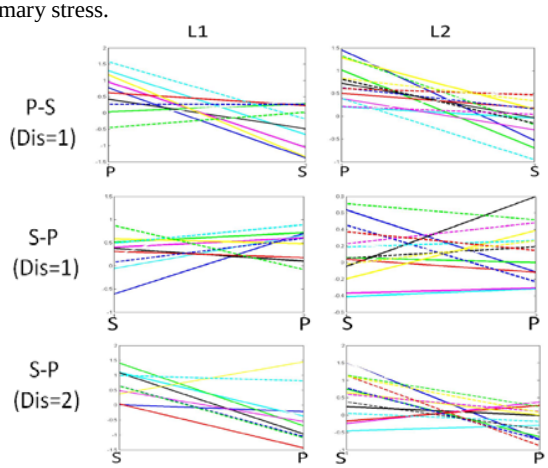


Figure 2: *F0 of secondary stress without intonation effect by L1/L2, linear order (before or after) the primary stress and its distance from primary stress. Each color line denotes F0 patterns of one speaker.*

The above results suggest that the F0 realization of secondary stress without intonation effect is context-dependent, concomitant but predictable. When secondary stress precedes the primary one, the F0 distinction does not always exist; the two categories are often under-differentiated. However, in reversed positions when secondary stress follows the primary one, it is lowered significantly, thus creating a sharper pitch contrast as shown in upper left panel. It is therefore no surprise why secondary stress is annotated differently in

different dictionaries. It is therefore also true that it is more difficult for L2 speakers to master.

3.1.2. Duration without segmental duration—rhythm contrast

Figure3 shows normalized duration of secondary stress by L1/L2, linear order (before or after) the primary stress and its distance from primary stress. The most distinct difference between L1 and L2 English is found when secondary stress appears immediately BEFORE primary stresses. By this context, the primary stress is distinctly longer than secondary stress while L2 English is again highly varied. For L1 English, primary stress is always longer than secondary stress thus the positions of primary stresses could be indicated by systematic rhythm/beat variation while L2 English exhibited no similar rhythm patterns.

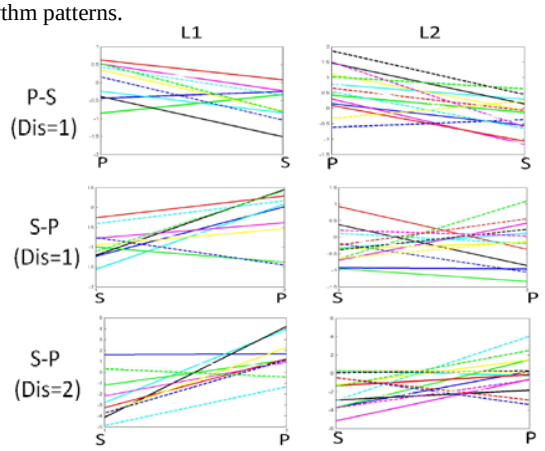
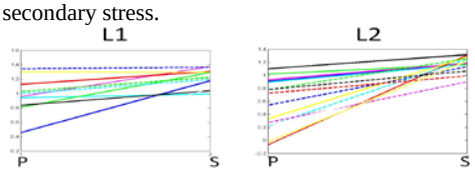


Figure 3: *Normalized Duration of secondary stress by L1/L2 and primary/secondary stress context which is represented by order and distance between primary and secondary stress. Each color line denotes F0 patterns of one speaker.*

The above results suggest the most robust feature to distinguish L1 and L2 English is normalized duration. For L1 English, primary stressed syllables are always longer than secondary stresses, creating systematic rhythmic patterns that imply distinctly the position of primary stresses in a word. However, L2 English lacks the same rhythmic contrast even some duration difference is exhibited. In other words, no categorical rhythm differentiation can be found in L2 English.

3.1.3. Intensity—loudness contrast

Figure4 shows intensity of secondary stress by L1/L2, linear order (before or after) to the primary stress and its distance from primary stress. For both L1 and L2 English, no significant difference is found between primary stress and secondary stress.



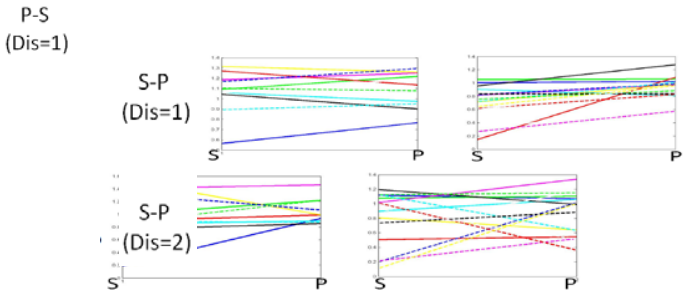


Figure 4: *Normalized Duration of secondary stress by L1/L2, linear order (before or after) the primary stress and its distance from primary stress. Each color line denotes F0 patterns of one speaker.*

3.2. Contrast patterns--narrow focus vs. sentential prominence

This study examines prosodic contrast of broad focus (sentential prominence) by linear order (before or after) to the narrow focus and its distance from narrow focus to test if (1) a sentence-initial broad focus is assimilated to the narrow focus when it is immediately BEFORE the narrow focus; or if the broad focus is reduced to a lower level in order to create more robust contrast patterns when it is AFTER the narrow focus. (2) However, if the same rationale can be applied to distance as well. Furthermore, this study tests if these characteristics could also help distinguish L1 and L2 English.

3.2.1. F0 without intonation effect—pitch contrast

Figure5 shows F0 of broad focus without intonation effect by L1/L2, linear order (before or after) to the narrow focus and its distance from narrow focus. When narrow focus precedes prominent word, the narrow focus is higher than broad focus for both L1 and L2 English. When broad focus precedes narrow focus, the contrast between narrow focus/broad focus is not clear for both L1 and L2 except for distance=8. It denotes broad focus “stand out” and is more differentiable from narrow focus when there is more distance between them. In addition, the contrast degree of L1 English is slightly higher than L2 English.

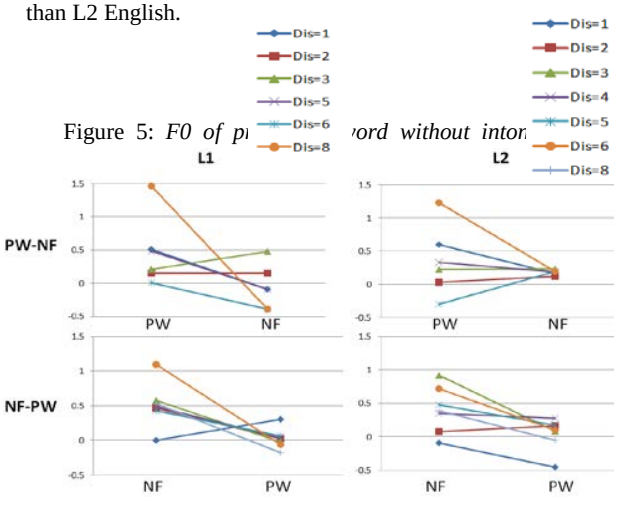


Figure 5: *F0 of pi ord without intor effect by L1/L2, linear order (before or after) the narrow focus and its distance from narrow focus PW, NF and Dis represent narrow focus, prominent word and distance by words.*

The above results show how the contrast patterns of broad/narrow focus distinction is similar to patterns found between primary and secondary stress, as shown in 3.1. Broad focus is dependent on position and distance from narrow focus. The concomitant and more subtle differentiation again proves to be difficult for L2 speakers. L2 speech sounds flatter in melody.

3.2.2. Duration—Tempo contrast

Figure6 shows duration patterns of broad focus by L1/L2, linear order (before or after) to the narrow focus and its distance from narrow focus. For narrow focus before or after broad focus, L1 English shows significant difference with L2 English, namely, a distinct contrast between narrow focus and broad focus. Narrow focus is always slower than broad focus thus the position of narrow focus could be indicated by systematic change of tempo. For L2 English, the tempo pattern between broad and narrow focus is more monotonous and exhibits no systematic tempo variations as found in L1.

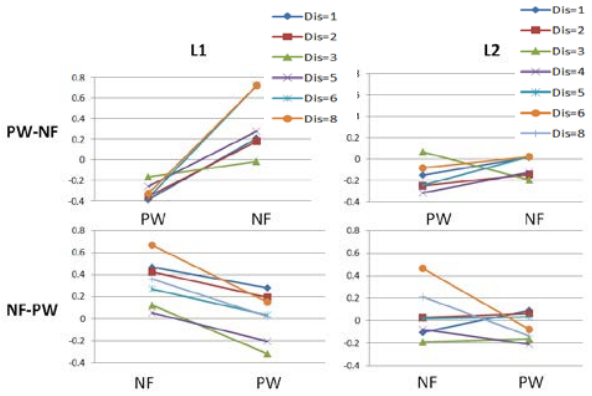
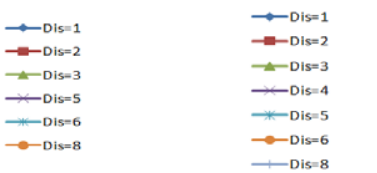


Figure 6: *Tempo of prominent word by L1/L2, linear order (before or after) the narrow focus and its distance from narrow focus. PW, NF and Dis represent narrow focus, prominent word and distance by words.*

The results above showed that the most significant difference of broad and narrow focus between L1 and L2 English is tempo patterns. While L1 speakers maintain distinct differentiating patterns as shown in left panels in Figure 6, L2 speakers could not realize the same tempo contrast patterns. As a result, L2 speech sounds more monotonous.

3.2.3. Intensity

Figure7 shows intensity patterns of broad focus by L1/L2, linear order (before or after) to the narrow focus and its distance from narrow focus. Difference between L1 and L2 English is found when narrow focus precedes broad focus. By this context, a lager degree of contrast in L1 English is found than L2 English.



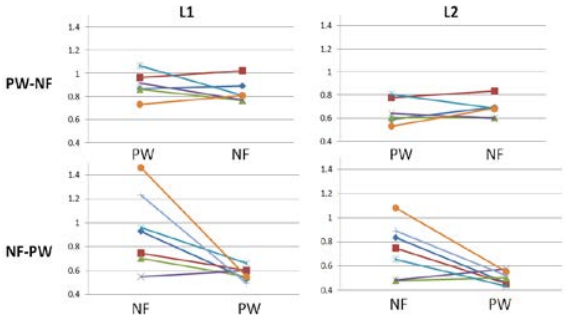


Figure 7: *Intensity of prominent word by L1/L2, linear order (before or after) the narrow focus and its distance from narrow focus. PW, NF and Dis represent narrow focus, prominent word and distance by words.*

The results above showed that the strong/weak contrast of broad/narrow focus is only differentiable for L1 when distance factor is bigger. L2 speech is less differentiable as expected.

4. General Discussion and Conclusion

Following our previous studies that showed the major acoustic characteristics of English lexical stress is F0 (pitch) contrast [6, 7], we further found in the present study that the major acoustic characteristic of primary vs. secondary stress is duration (rhythm) contrast, as shown in L1 speech. Though the same rationale is also found in the differentiation of broad vs. narrow focus, the patterns are additional contrast patterns on top of word level distinctions. The added results collectively suggest that binary contrast is the optimal choice of differentiating patterns, thus providing evidence of binary opposition, a crucial phonological concept, in the physical sense. Furthermore, these relative opposition patterns appeared to be quite difficult for L2 speakers to produce, suggesting on how perceptual sensitivity of such relative contrast patterns may be language dependent, and why L2 speech is less differentiable in production and flatter sounding in perception. Hence, under-differentiation of the necessary contrasts is a major feature of Taiwanese L2 English. We believe category related differentiation that require robust but sometime concomitant contrasts have not been properly addressed in language teaching in general and could be implemented to CALL technologies.

5. References

[1] Magen, H.S., "The perception of foreign-accented speech", Journal of Phonetics, vol. 26, 381-400, 1998.

[2] Anderson-Hisieh, J., Johnson, R. and Koehler, K. "The relationship between native speakers judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure", Language Learning 42: 4 529-555, 1992.

[3] Tajima, K., Port, R., and Dalby, J. "Effects of temporal correction on intelligibility of foreign-accented English", Journal of Phonetics, 25, 1-24, 1997.

[4] Field, J. "Intelligibility and the listener: The role of lexical stress", TESOL Quarterly, 39(3), 399– 423, 2005.

[5] Mixdorff, H. and Ingram, J. “Prosodic analysis of foreign-accented English”, Proc. Interspeech 2009, 6-10 Sep. Brighton UK, 2009.
[6] Tseng, C. Y. Su, Z. Y. and Visceglia, T. “Underdifferentiation of English Lexical Stress Contrasts by L2 Taiwan Speakers”, Slate 2013 164-167. Grenoble, France, 2013.
[7] Tseng, C. Y. Su, Z. Y. and Visceglia, T. “Levels of Lexical Stress Contrast in English and their Realization by L1 and L2 Speakers”, KIIT Gurgaon, India, 2013.
[8] Tseng, C. Y. and Su, Z. Y. “Dynamic Discourse Speech Tempo and Phonological Timing”, The 7th International Congress of Phonetic Sciences. Hong Kong, China, 2011.

Crowdsourcing regional variation in speaking rate through the iOS app 'Dialäkt Äpp'

Adrian Leemann[1], Marie-José Kolly[1], Volker Dellwo[1]

[1]Phonetics Laboratory, Department of General Linguistics, University of Zurich
{adrian.leemann, marie-jose.kolly}@pholab.uzh.ch, volker.dellwo@uzh.ch

Abstract

It is a common stereotype in Switzerland that speakers from Bern speak slowly and speakers from Zurich speak quickly. Are these differences in perception at all mirrored in production? We present a new method of crowdsourcing speaking rate through a free of charge iOS application. Astonishingly, results indicate that the temporal structure of a few words alone – as spoken by a few hundred speakers – are sufficient to tell apart the two dialects in speaking rate. In line with previous literature, females articulate more slowly than males. Further potential fields of application of the introduced method are discussed.

Index Terms: Speaking rate, crowdsourcing, dialectology, Swiss German, iOS application

1. Introduction

Swiss German dialects are spoken by roughly 4.5 million people [1] and enjoy high prestige in Swiss society [2, 3, 4]. Speakers of Swiss German (SwG) are well aware of regional variation and many dialects are stereotyped: Zurich Swiss German (ZH SwG), for example, is perceived as fast. Bern Swiss German (BE SwG), which enjoys the status of being Switzerland's most popular regional variety [5], is perceived as very slow [6, 7, 8]. Whether these differences in perception are reflected in production has been examined in passing by [9, 10]. Based on a corpus of spontaneous speech for ten speakers per dialect, [9, 10] reported that ZH SwG speakers articulate nearly one syllable more per second than BE SwG speakers (5.8 syll./sec. vs. 5.0 syll./sec – excluding pauses), thereby corroborating the previously mentioned stereotypes. Possible reasons for these differences in speaking rate were given in [10], who showed that BE SwG speakers produced distinctly longer mean durations of vowels and, in particular, exhibited more distinct phrase-final lengthening.

The studies mentioned present one major weakness: while clearly highlighting trends in regional variation in speaking rate, these tendencies have yet to be validated on a large set of speakers and on controlled material. We aim to alleviate this issue: based on crowdsourced data from an iOS application, we provide a more precise estimate of regional variation in speaking rate on the basis of nearly 250 speakers who articulated a controlled set of words. The term 'crowdsourcing' refers to "the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers" [11].

The use of crowdsourcing applications for studying linguistic phenomena has, until recently, received relatively little attention. This is somewhat surprising given that iPhone microphones, for example, feature wide frequency ranges of 50Hz-20kHz that enable high-quality audio recordings [12]. Previous research showed that a first generation iPhone from 2007 provides very useful for speech analysis and allows for reliable acoustic measurements – particularly for F1 and F2 [13]. Currently, a number of smartphone applications are in use or in development for crowdsourcing linguistic data. [14, 15] developed Android applications as a means to collect speech for the training of acoustic models. [16, 17] are applications currently in development for the purpose of documenting endangered languages, putting language documentation in the hands of the speakers. The mentioned apps are primarily used for acoustic modeling, dictionary building, text collection, translation, as well as dialect mapping. We present a novel method for crowdsourcing data to conduct research on prosodic features of dialects.

2. Data and methods

2.1. iOS application: 'Dialäkt Äpp'

'Dialäkt Äpp' [18] capitalizes on the Swiss' public interest in dialectology. It provides functionality that allows users (1) to localize their own Swiss German dialect by indicating – i.e. listening to pre-canned recordings and then tapping on the screen – their dialectal pronunciation of 16 tokens, i.e. words, and (2) to articulate and anonymously record these 16 tokens in their dialect. Data used in the current study stem from this second function.

The user interface (UI) prompts the users to indicate their dialect (possible localities are those used in [19]), age, and gender (Figure 1, left panel) before they proceed to the recording instructions screen (Figure 1, right panel), see Figure 1.

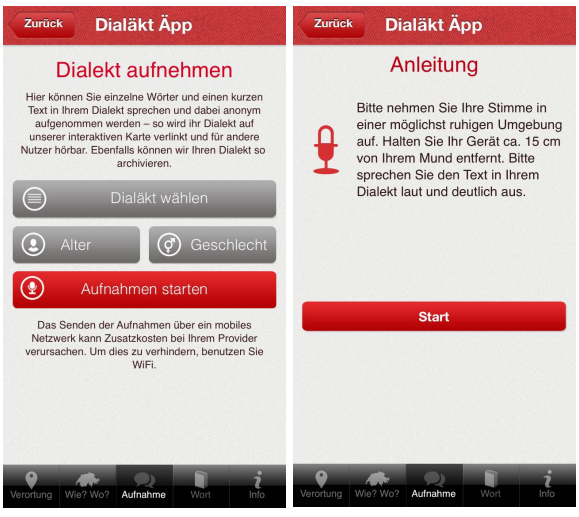


Figure 1: *UI for dialect, age, and gender selection (left panel) and recording instructions (right panel).*

The right panel in Figure 1 reads: "Please record your voice in as quiet an environment as possible. Keep an approximate distance of about 15 cm between your device and your lips. Please articulate the text loudly and clearly in your own dialectal pronunciation". Next, the user articulates and records the token shown on the screen (see Figure 2, left panel). The 16 tokens in this recording function are the same as those used for the localization function. Once the recordings are finished they are anonymously uploaded on our servers where each audio file is given a unique ID. Following the upload, users can navigate to an interactive map of Switzerland where they can listen to their own recordings and those of other users (Figure 2, right panel, green and purple pins).

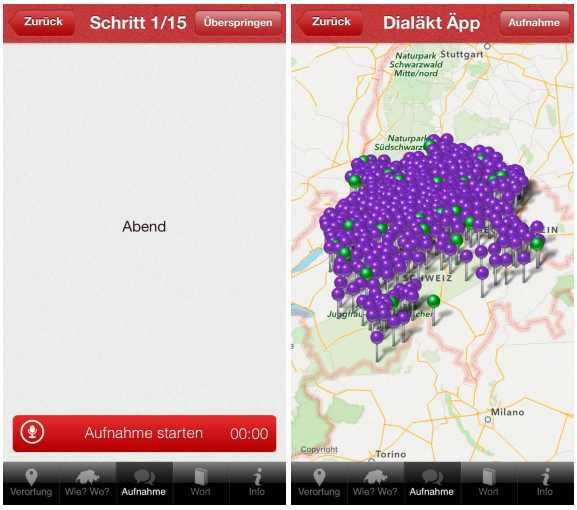


Figure 2: *UI for token recording (left panel) and audio playback map of one's own and other users' recordings (right panel).*

In Switzerland, 'Dialäkt Äpp' was the number one downloaded free app for iPhones after its release on March 22, 2013 [20]. It received major media attention and, so far, has >58,000 downloads. More than 2300 users from all over German-speaking Switzerland have uploaded voice recordings.

2.2. Subjects

Users who declared BE SwG (i.e. Bern city) and ZH SwG (i.e. Zurich city) as their local dialect served as subjects. In total there were 115 unique BE SwG speakers and 205 unique ZH SwG speakers. Not all speakers read all of the presented words, which is why the number of observations varies from token to token (cf. 2.3). On average speakers were 32-years-old, ranging between 4 years of age and 75 years of age, with 60% males and 40% females.

2.3. Material

We selected six out of a total of 16 'Dialäkt Äpp' tokens (cf. 2.1) for analysis of speaking rate. Selection criteria were that each token consisted of two syllables, given that we measured the temporal distance between adjacent vowel onsets (cf. 2.4). Half of our selected words further featured phonologically Middle High German long vowels or diphthongs while the other half featured underlying short vowels. The selected words with underlying long vowels were *Abend* 'evening', *Augen* 'eyes', and *fragen* 'to ask'; those with underlying short vowels *Donnerstag* 'Thursday', *heben* 'to lift', and *trinken* 'to drink'. Typically, these words are articulated as follows:

Long vowels/diphthongs:
Abend: BE SwG: [ˈɑːb̥ə], ZH SwG: [ˈɒːb̥ig̥]
Augen: BE SwG: [ˈɔug̥ə], ZH SwG: [ˈæug̥ə]
fragen: BE SwG: [ˈfrɑːg̥ə], ZH SwG: [ˈfrœːg̥ə]

Short vowels:
Donnerstag: BE SwG: [ˈd̥ɔnʃt̥i], ZH SwG: [ˈd̥unʃt̥ig]
heben: BE SwG: [ˈlʏp͡fə], ZH SwG: [ˈlup͡fə]
trinken: BE SwG: [ˈtrɪŋk͡xə], ZH SwG: [ˈtriŋk͡xə]

Table 1 presents the total number of observations, i.e. speakers, with figures on gender.

	ZH SwG	BE SwG
Abend 'evening'	n=188 (114m, 74f)	n=100 (58m, 42f)
fragen 'to ask'	n=193 (118m, 75f)	n=103 (64m, 39f)
Augen 'eyes'	n=186 (113m, 73f)	n=96 (60m, 36f)
Donnerstag 'Thursday'	n=186 (114m, 72f)	n=100 (63m, 37f)
heben 'to lift'	n=194 (118m, 76f)	n=104 (64m, 40f)
trinken 'to drink'	n=199 (120m, 79f)	n=105 (66m, 39f)

Table 1: *Summary of the number of total observations, i.e. speakers.*

The majority of recordings were usable, i.e. demonstrated little background noise interference nor were the speakers goofing off. Instances of unfavorable audio quality or otherwise unusable material were disregarded from the analyses (percentage of discarded tokens: approximately 20%).

2.4. Procedure

There are various approaches to measuring speaking rate. Most commonly one measures a linguistic unit per second (words, syllables, segments, consonantal intervals, vocalic intervals; cf. [21]). Since our corpus contains words that exhibit cross-dialectal differences in syllable structure (e.g. *Abend*: BE SwG V.CV [ˈɑːb̥ə] vs. ZH SwG V.CVC [ˈɒːb̥ig̥] or *Donnerstag*: BE SwG CVC.CCV [ˈd̥ɔnʃt̥i], vs. ZH SwG CVC.CCVC [ˈd̥unʃt̥ig]), we refrained from applying conventional speaking rate measures such as number of syllables per second, and instead measured the temporal duration between the two vowel onsets. In theory, this is motivated by [22]'s findings that vowel onsets represent perceptually prominent centers of a syllable. We call this measure of vowel-onset-to-vowel-onset duration *durVonVon*. Figure 3 schematically shows the measurement technique applied in the present study

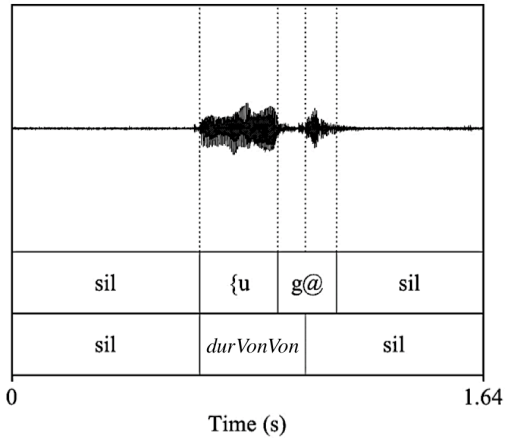


Figure 3: *Schematic of vowel-onset-to-vowel-onset measurement (2nd tier).*

Figure 3 shows the oscillogram of a ZH SwG speaker articulating the token *Augen* as [ˈæugə] (cf. 1st tier). The 2nd tier shows the boundaries placed at the vowel onsets. 'sil' indicates silence. Altogether there were 2920 measurement points (1460 intervals). Temporal duration between these two vowel onsets was measured in Praat [23].

3. Results

3.1. Statistical analyses

All data were analyzed using R [24] and the R packages *lme4* [25] and *languageR* [26, 27]. If not indicated otherwise, we analyzed data using linear mixed effect models (LMEs). Normality was checked by visual inspection of quantile plots. *Dialect*, *gender*, and *vowel type* were treated as fixed effects, *token* and *age* as random effects. Effects were tested by model comparison between a full model, in which the factor in question is entered as either a fixed or a random effect, and a reduced model without this effect. p-Values were obtained by comparing the results from the two models using ANOVAs. For the assessment of the relative goodness of fit, we report *AIC* (Akaike Information Criterion) values that decrease with goodness of fit. Only p-values that are considered significant at the α= 0.05 level are reported.

3.2. Overall regional differences

Figure 4 shows the boxplots representing the two dialects' *durVonVon* data.

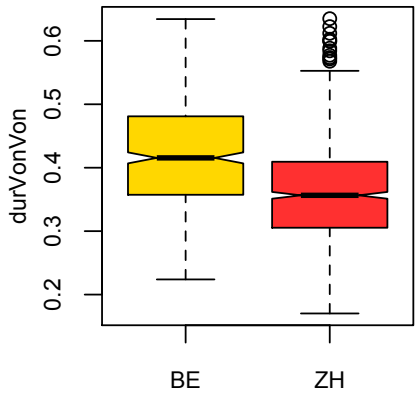


Figure 4: *Boxplots of the dialects' durVonVon.*

The yellow boxplot indicates the values for BE SwG, the red boxplot those of ZH SwG. Visually, the two boxes' notches do not overlap, which can be taken as strong evidence that their medians differ. The comparison between the full and reduced models showed a significant difference for dialect, with the full model exhibiting an increased goodness of fit (BE SwG M=.42, SD=.08; ZH SwG M=.36, SD=.08; p<.0001; *AIC*=-3806). There was thus a significant difference in *durVonVon* between the two dialects. BE SwG speakers showed longer *durVonVon* intervals than ZH SwG speakers.

3.3. Cross-gender differences by dialect

Figure 5 shows the distribution of *durVonVon* across *dialect* and *gender*.

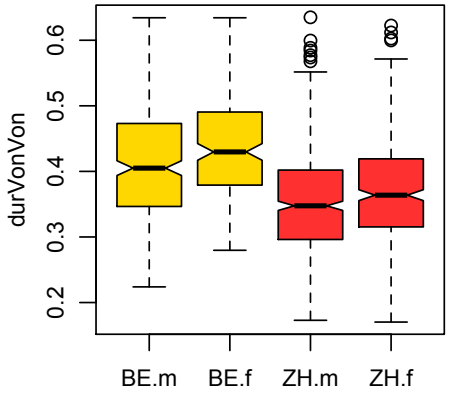


Figure 5: *Boxplots of durVonVon across dialect and gender.*

The differences between the genders were significant in both dialects (Bonferroni adjusted for *gender*, α=0.025; both p<.0001; BE SwG: *AIC*=-1244, ZH SwG: *AIC*=-2598). The boxplots in Figure 5 indicate that for both BE SwG and ZH SwG, females demonstrated significantly longer vowel-onset-to-vowel-onset durations. There was no significant interaction of *dialect*gender* (p=.20; *AIC*=-3799).

4. Discussion

Based on a controlled set of words spoken by a large number of speakers, the current study found that BE and ZH SwG strongly differ in terms of speaking rate. For the sake of

illustration, let us extrapolate these results to a more realistic scenario. Say a BE SwG and a ZH SwG speaker read Aesop's fable "The North Wind and the Sun". In Zurich German, the fable consists of 129 syllables, i.e. approximately 128 vowel-onset-to-vowel-onset intervals (cf. [28]). Based on raw findings of the current study – disregarding contextual factors such as phrase-final lengthening – the BE SwG speaker would need 54 seconds to read the text while the ZH SwG speaker would only take 46 seconds.

This finding is intriguing in a number of ways. [10] notes that BE SwG speakers speak more slowly particularly because they exhibit more distinct phrase-final lengthening. Results of the current study show, however, that the two dialects strongly vary from one another in speaking rate irrespective of such contextual factors. Our findings are unique in just this sense: the temporal information contained in a few words alone is already sufficient to tell apart the two dialects (cf. Figure 4) – regardless of contextual factors such as phrase-final lengthening. In future studies it would be interesting to test whether vowel length differences are the major influential factor. Moreover, it will be interesting to examine if we find these between-dialect differences in each of the 6 tokens individually.

Concerning gender differences in speaking rate. The result that female speakers articulate more slowly than males is in line with previous studies on cross-gender differences in speaking rate on British English dialects [29] and on American English dialects [30, 31]. It is further in line with [32] who shows that in German, durations of female vowels are systematically longer than durations of male vowels.

Several questions remain unanswered at present. Is it conceivable that listeners can tell apart the two dialects based solely on time domain information in perception experiments? To answer this question one would require tokens that are identical in segmental, syllabic, and prosodic structure, which (deliberately) does not apply to the tokens used in 'Dialäkt Äpp'. One could, however, use 'Dialäkt Äpp' tokens that are identical in syllabic structure and delexicalize them (e.g. *sasasa*-delexicalization, where every consonantal interval is replaced with a pre-recorded [s] and every vocalic interval with a pre-recorded [a] (cf. [33]).

Speaking rate represents only one of countless areas of application in speech prosody where crowdsourcing is useful. In the present region-wide 'Dialäkt Äpp' corpus, more than 2300 speakers have uploaded voice recordings, and in most cases speakers recorded all 16 words. This amounts to approximately 36,000 voice recordings. In future studies we will further examine fundamental frequency distributions, temporal, stress and intonational patterns, and generate vowel plots. These phenomena can be explored in multiple dimensions, enabling us to test for effects of speaker, age, gender, locality or region. Crowdsourcing applications for American English, German regional varieties, and British English are currently being developed, inspired by the 'Dialäkt Äpp' framework. The acoustic data crowdsourced through the 'Dialäkt Äpp' is further used to train an automatic speech recognition system. This system will be part of a follow-up smartphone application [34] that will perform dialect localization based on spoken language input.

5. Conclusion

Based on crowdsourced utterances from a large number of speakers, results of the current study corroborate previous impressionistic and empirical observations that ZH SwG is fast and BE SwG slowly spoken [6, 7, 8, 9, 10]. Results of the current study revealed that regional differences in speaking rate are prevalent on the basis of a few words alone. We further showed that female speakers articulate more slowly than male speakers, which is in line with other research findings.

6. Acknowledgments

We thank Daniel Wanitsch for server-side technical assistance and audio data extraction and Ingrid Hove for database maintenance. We are indebted to 65 backers who made 'Dialäkt Äpp' possible through crowdfunding. Thank you!

7. References

[1] BFS = Bundesamt für Statistik, Statistisches Lexikon der Schweiz. Personenverkehr: Entwicklung der Tagesmobilität, 2005, http://www.bfs.admin.ch/.

[2] Werlen, I., "Zur Sprachsituation in der Schweiz mit besonderer Berücksichtigung der Diglossie in der Deutschschweiz", Bulletin VALS-ASLA (Vereinigung für angewandte Linguistik in der Schweiz), 79:1-30, 2004.

[3] Christen, H., "Was Dialektbezeichnungen und Dialektattribuierungen über alltagsweltliche Konzeptualisierungen sprachlicher Heterogenität verraten", in C. Anders, M. Hundt and A. Lasch [Eds], "Perceptual dialectology". Neue Wege der Dialektologie, 269-290, Berlin/New York: de Gruyter, 2010.

[4] Hotzernköcherle, R., Die Sprachlandschaften der deutschen Schweiz. Ed. by N. Bigler, R. Schläpfer, Aarau: Sauerländer, 1984.

[5] Schwarzenbach, R., Die Stellung der Mundart in der deutschsprachigen Schweiz. Studien zum Sprachgebrauch der Gegenwart (= Beiträge zur schweizerdeutschen Mundartforschung XVII), Frauenfeld: Huber, 1969.

[6] Ris, R., "Innerethik der deutschen Schweiz", in P. Hugger [Ed], Handbuch der schweizerischen Volkskultur, vol. II, 749–766, Zürich: Offizin, 1992.

[7] Berthele, R., "Wie sieht das Berndeutsche so ungefähr aus? Über den Nutzen von Visualisierungen für die kognitive Laienlinguistik", in H. Klausmann [Ed], Raumstrukturen im Alemannischen. Beiträge der 15. Arbeitstagung zur alemannischen Dialektologie, Schloss Hofen (Vorarlberg) vom 19.-21.9.2005 (= Schriften der VLB 15), 163–176, Graz-Feldkirch: Neugebauer, 2006.

[8] Werlen, I., "Zur Einschätzung von schweizerdeutschen Dialekten", in I. Werlen [Ed], Probleme der schweizerdeutschen Dialektologie. 2. Kolloquium der Schweizerischen Geisteswissenschaftlichen Gesellschaft 1978, 195–257, Fribourg, 1985.

[9] Leemann, A., Swiss German Intonation Patterns, Amsterdam/Philadelphia: Benjamins, 2012.

[10] Leemann, A. and Siebenhaar, B., "Statistical Modeling of F0 and Timing of Swiss German Dialects", Proceedings of Speech Prosody, 2010.

[11] crowdsourcing, Merriam-Webster.com, Retrieved December 14, 2013, http://www.merriam-webster.com/dictionary/crowdsourcing

[12] http://blog.faberacoustical.com/2009/ios/iphone/iphone-microphone-frequency-response-comparison/.

[13] De Decker, P. and Nycz, J., "For the Record: Which Digital Media Can be Used for Sociophonetic Analysis?," University of Pennsylvania Working Papers in Linguistics 17(2):51-59, 2011.

[14] Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. and LeBeau, M., "Building transcribed speech corpora quickly and cheaply for many languages", Proceedings of Interspeech 26.-30.10.2010, Makuhari, Chiba, Japan: 1914-1917.

[15] de Vries, N., Davel, M. H., Badenhorst, J., Basson, W. D., de Wet, F., Barnard, E. and de Waal, A., "A smartphone-based ASR data collection tool for under-resourced languages", Speech Communication, 56: 119-131, 2014.
[16] Ma! Iwaidja, https://itunes.apple.com/au/app/ma-iwaidja/id557824618?mt=8.
[17] Hanke, F. R., Byrd, S., "Large-scale text collection for unwritten languages", International Joint Conference on Natural Language Processing, 1134-1138; http://lp20.org/aikuma/, 2013.
[18] Leemann, A., and Kolly, M.-J., Dialäkt Äpp. https://itunes.apple.com/ch/app/dialakt-app/id606559705?mt=8, 2013.
[19] SDS = Sprachatlas der deutschen Schweiz, Bern (I-VI), Basel: Francke (VII-VIII), 1962-2003.
[20] http://www.appannie.com/,
[21] Roach P., "Myth 18: Some languages are spoken more quickly than others", in L. Bauer and P. Trudgill [Ed], Language Myths, 150–158, London: Penguin 1998.
[22] Allen, G. D., "The location of rhythmic stress beats in English: An experimental study I.", Language and Speech, 15:72–100, 1972.
[23] Boersma, P. and Weenink D., Praat: doing phonetics by computer, www.praat.org, 2013.
[24] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Version 3.0.0. http://www.R-project.org, 2013.
[25] Bates, D. M. and Maechler, M., lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-32, 2009.
[26] Baayen, R. H., Analyzing Linguistic Data: A Practical introduction to statistics using R, CUP, Cambridge, 2008.
[27] Baayen, R. H., languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics using R", R package version 0.955, 2009.
[28] Fleischer, J. and Schmid, S., "Zurich German", Journal of the International Phonetic Association, 36(2):243-253, 2006.
[29] Whiteside, S. P., "Temporal-based acoustic-phonetic patterns in read speech: some evidence for speaker gender differences", Journal of the International Phonetic Association, 26:23-40, 1996.
[30] Jacewicz, E., Fox, R. A., O'Neill, C. and Salmons, J., "Articulation rate across dialect, age, and gender", Language Variation and Change, 21:233-256, 2009
[31] Byrd, D., "Preliminary results on speaker-dependent variation in the TIMIT database", Journal of the Acoustical Society of America, 92(1):593-596, 1992.
[32] Simpson, A. P., „Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung", Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 33, 1998.
[33] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: A study based on speech resynthesis", Journal of the Acoustical Society of America, 105(1):512–521, 1999.
[34] Swiss Voice App, www.voiceapp.ch

Are gesture and prosodic prominences always coordinated? Evidence from perception and production

Núria Esteve-Gibert[1], *Ferran Pons*[2], *Laura Bosch*[2], *Pilar Prieto*[3,1]

[1]Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
[2] Department of Basic Psychology, University of Barcelona, Barcelona, Spain
[3] ICREA – Institució Catalana de Recerca i Estudis Avançats

nuria.esteve@upf.edu, ferran.pons@ub.edu, laurabosch@ub.edu, pilar.prieto@upf.edu

Abstract

This study explores the temporal coordination between gesture and speech by addressing two main questions: (1) Are speakers sensitive to the misalignment between gesture prominence and prosodic prominence? (2) Is this sensitivity modulated by the semantic information conveyed by gesture and speech modalities in production? Experiment 1 tested question (1) and Experiment 2 tested question (2). Results from Experiment 1 revealed that the combinations in which prominences were misaligned were less acceptable than combinations with aligned prominences, and that the metrical pattern of the target word had an effect on the speakers' sensitivity: unsynchronized trochees (with the gesture prominence at the post-tonic syllable) were frequently accepted, while unsynchronized iambs (with the gesture prominence at the pre-tonic syllable) were rejected. Results from Experiment 2 revealed that when the pointing gesture adds information to speech, i.e. it is supplementary to speech, the prominences are frequently misaligned (with gesture occurring after the speech), as if two different speech acts were produced. These findings suggest that the semantic content of gesture-speech combinations might influence the speakers' sensitivity of the misalignment between prosodic and gesture prominences.

Index Terms: gesture-speech synchronization, audiovisual prosody, multimodal prominence

1. Introduction

There is ample evidence in the literature that humans coordinate gesture movements with speech, suggesting that both modalities are in fact part of an integrated system [1-3]. This coordination is evidenced from both semantic and temporal points of view. What speakers express with their hands is semantically related with what they express with their speech (what could be called 'semantic coordination'). Also, gesture and speech timings are coordinated, since the most prominent part of the gesture co-occurs with the most prominent part of speech ('temporal coordination') [3].

1.1. Temporal and semantic coordination

Studies investigating the temporal coordination of gesture and speech have found convincing evidence that gesture and speech co-occur in time in the sense that the point of maximal expression of a gesture (hereafter 'gesture prominence') coincides with the moment of maximal prosodic prominence in speech [4]. In order to define gestural prominence, most studies use either the stroke of the gesture (the interval involving the greatest physical effort in the gesture) or the apex of the gesture (the point in time in which the gesture reaches its maximal extension). As for the prominent feature of speech, a growing body of research has found that the speech landmark with which the gesture prominence aligns is the lexical stress [5, 6] and even the pitch peak within the stressed syllable when it is uttered in a contrastive focus situation [7, 8].

However, it has also been proposed that the temporal synchronization between gesture and speech may depend on their semantic coordination: when the meanings expressed by the co-speech gesture and by the accompanying lexical affiliate are complementary, the onset of the gesture stroke is closely aligned with its lexical affiliate; but when the two modalities express supplementary semantic features, stroke onset and lexical affiliate are not so closely aligned [9]. But more evidence is needed to corroborate this hypothesis.

1.2. Perception of temporal asynchrony

But how important is this tight temporal coordination? As interlocutors, do we expect the gesture apex to co-occur with the lexical stress? Do we perceive misalignments in their temporal coordination?

Most of the studies examining the perception of audio-visual asynchrony have focused on the human ability to perceive unsynchronized audiovisual events in articulatory gestures of a person producing syllables or a list of words. They found that adults can detect an audiovisual asynchrony of around 200 ms when the visual attributes of an audiovisual event precede the auditory attributes, and around 100 ms when the auditory attributes precede the visual attributes [10]. However, the articulatory synchronization patterns tested in these experiments did not answer the question of whether the temporal coordination of prominences found for co-speech gestures is relevant in perception.

Few studies have examined the effects of the gesture-prosodic misalignment in the perception of the lexical stress [11-13]. Results seem contradictory, some finding a clear influence [12, 13] and some not [11]. From these, only in [11] the authors analyzed pointing gestures and they did not find a clear influence and the authors suggest that their results might be influenced by some methodological problems with the procedure. Thus, the influence of the timing of the gesture prominence with respect to the speech prominence needs to be further analyzed.

1.3. Aim of the study

The aim of the present study was two-fold: first, to investigate speakers' ability to perceive a temporal asynchrony between gesture and speech prominences (Experiment 1); second, to investigate whether this perceptual ability is related to how

speakers align gesture and speech when the semantic information expressed by gesture supplements what is expressed in speech (Experiment 2).

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty-two adult Catalan-speakers took part in an online acceptability judgment task. They were unaware of the purpose of the study and participated voluntarily.

2.1.2. Materials

An online survey was prepared using the SurveyGizmo application. Participants watched a series of video clips each showing a woman producing a disyllabic word accompanied by a deictic pointing gesture. The woman appeared sideways in the right part of the screen and pointed to the left part of the screen. In order to prevent participants from looking at her lip movements, the woman covered her mouth with the hand not used for pointing (see Figure 1, left panel).

Sixteen disyllabic words were used, half of them iambs (with stress on the second syllable) and the other half trochees (with stress on the first syllable). They were all common words, such as "miRALL" ('mirror'), "ioGURT" ('yogurt'), "AIgua" ('water'), or "COtxe" ('car').[1] Words were pronounced in an exaggerated manner so that syllable duration values were longer and pitch range values were higher than in spontaneous speech (see Figure 1, right panel).

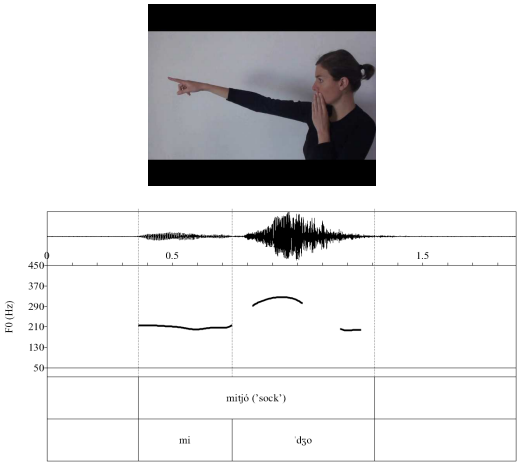


Figure 1: *Top panel, visual stimulus presented in the survey: a video clip showing a woman pointing while saying a word (frame showing the apex of the gesture). Bottom panel, waveform and pitch contour of the word produced while pointing, with the F0 peak coinciding with the gesture apex.*

Of the total number of video clips that participants observed (N = 32), half were synchronized (gesture apex coinciding with lexical stress) and half were unsynchronized (gesture apex not coinciding with lexical stress). Using Adobe Premiere Pro, all clips were constructed with the same pointing gesture and then the various audio inputs were juxtaposed on it, either synchronized or not. To create the synchronized stimuli, we combined the audio track of the different target words with the video track of the pointing movement so that the apex of the gesture movement coincided with the pitch peak of the target word (see the frame in Figure 1). To create the unsynchronized stimuli, we combined the audio track of each target word with the video track of the pointing movement in such a fashion that the apex of the gesture movement occurred in the middle of the unaccented syllable. Synchronized and unsynchronized stimuli were randomly mixed during the survey.

2.1.3. Procedure

Participants were asked to rate the acceptability of the video clips containing either synchronized or unsynchronized gesture-speech combinations on a 5-point Likert scale (1 = totally unnatural; 2 = quite unnatural; 3 = slightly unnatural; 4 = quite natural; 5 = totally natural).

Before the survey, participants were asked to imagine that the person in the videos was pointing at an object while naming it because she wanted to show them where the object was. Also, they were told that they had to base their acceptability judgments on the degree of coordination between gesture and speech that they perceived. The duration of the experiment was approximately 6 minutes.

2.2. Results

The total number of ratings obtained were 736 (23 participants x 32 clips), but 20 clips were found to have been left unrated by one or the other participant, so the total number of ratings analyzed was 716 (179 ratings for each of the four stimulus types, i.e. synchronized trochee, synchronized iamb, unsynchronized trochee, and unsynchronized iamb). An ANOVA analysis was carried out with acceptability rate as the dependent variable and stimulus type as the independent variable (four levels: synchronized trochee, synchronized iamb, unsynchronized trochee, unsynchronized iamb). The statistical analysis revealed that stimulus type significantly affected the acceptability rate ($F(3,715)=73.778$) = $p< .001$). Bonferroni post-hoc comparisons showed that, as expected, ratings for synchronized and unsynchronized trochees were significantly different ($p < .01$), and ratings for synchronized and unsynchronized iambs were also significantly different ($p < .001$), while synchronized trochees and synchronized iambs were rated similarly ($p> .05$). Surprisingly, ratings for unsynchronized trochees were also significantly different from ratings for unsynchronized iambs ($p< .001$). As Figure 2 shows, the mean acceptability rating for synchronized stimuli was very close to '4 = quite natural' (M = 3.79, *SD* = 0.983 for trochees; M = 3.89, *SD* = 0.963 for iambs). Unsynchronized iambs were rated very close to '2 = quite unnatural' (M = 2.36, *SD* = 1.331). However, participants judged unsynchronized trochees between '3 = slightly unnatural' and '4 = quite natural' (M = 3.39, *SD* = 1.050), thus more acceptable than unsynchronized iambs.

The results from Experiment 1 indicate that speakers detect the asynchrony between gesture and speech prominence, but it is more acceptable to them when the gesture apex occurs during an unaccented syllable in word-final (and also phrase-final in our stimuli) position (trochees) than during an unaccented syllable in word-initial position (iambs). Experiment 2 aimed at investigating the reason why misaligned trochees are more accepted than misaligned iambs.

[1]Capital letters indicate the accented syllable.

We hypothesized that when the pointing gesture conveys supplementary information to speech, speakers may misalign both modalities such that the gesture prominence can occur in post-tonic position but not in a pre-tonic one.

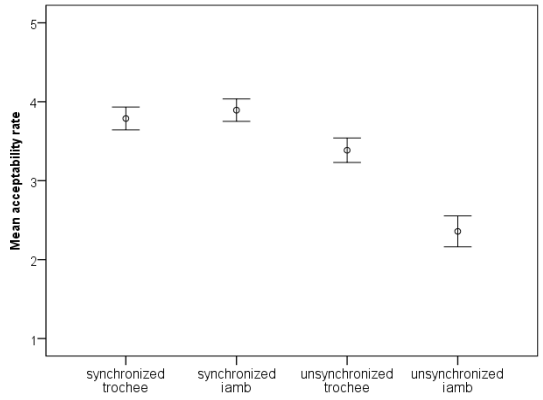


Figure 2: *Error bars of the mean acceptability rating as a function of stimulus type in Experiment 1.*

3. Experiment 2

In the second experiment we explored whether when the gesture is supplementary to speech speakers produce misaligned trochees (with gesture apexes in post-tonic position, i.e., phrase-final positions) but not misaligned iambs (with the apex in pre-tonic position).

3.1. Methods

3.1.1. Participants

Six Catalan-speakers participated in a pointing task. They were unaware of the purpose of the study and participated voluntarily.

3.1.2. Materials

In this pointing task, participants were asked to teach the experimenter the name of eight strange objects that were lined up in a row on a table (see Figure 3). The names of these objects were disyllabic nonsense words, half trochees (CVcv) and the other half iambs (cvCV), but all consisting of combinations of the same vowels and consonants, e.g. 'DUbi', 'duBÍ', 'BIdu' 'biDÚ'. Nonsense words were used to give meaning to the act of teaching and they were similar to make the game more challenging for the participants. Crucially, the participants had to name the object in the context of the sentence "*Agafa el* [target name]" ('Pick up the [target name]') and they were instructed not to produce any other kind of speech. Since the experimenter did not know which name referred to which object, participants were offered the possibility of using gestural strategies to indicate which object they were referring to.

3.1.3. Procedure

During the task, participants were recorded using a Panasonic HD AVCCAM recording at 25 frames per second. The sound was recorded through a small microphone that was placed somewhere on their clothing and as close as possible to their mouth.

Figure 3: *Setting of Experiment 2.*

At the beginning of the experiment, participants were given a legend in which the objects were labeled with their names. Participants were instructed to keep it hidden on their lap during the experiment and were then told that they were going to play a game in which they had to teach the experimenter the name of each object. In this teaching phase, the participant indicated the name of an object and its location to the interlocutor, then the interlocutor picked up that object, held it for a couple of seconds, and then put it back on the table. The participant then moved on to the next object. The task continued until the participant thought that the interlocutor would now be able to remember all the objects' names and locations. At that point the task ended and the interlocutor attempted to name all the objects.

3.1.4. Coding

All gesture-speech combinations that appeared in the video recordings were annotated using ELAN software in terms of the temporal features of both speech and pointing gestures. For speech, we annotated the temporal limits of the target name within the sentence, the metrical pattern of the name (either trochaic or iambic), and the temporal limits of the accented syllable within it. For pointing gestures, we annotated the preparation, stroke, and retraction phases of the gesture, and the location of the gesture apex [3].

3.2. Results and discussion

To examine whether participants produced unsynchronized trochees but not unsynchronized iambs, we calculated the location of the apex with respect to the end of the accented syllable as a function of the two metrical patterns. In total, 147 instances of items were analyzed, 73 with trochaic words and 74 with iambic words. Figure 4 and 5 illustrate the position of all the gesture apexes with respect to the accented syllable in trochaic and iambic words, separated by participant. In both figures, the solid horizontal line indicates the end of the accented syllable and the dotted line indicates the beginning of the accented syllable. Thus, circles occurring below the dotted line are cases in which the gesture apex occurs in the pre-tonic position and circles occurring above the solid horizontal line are cases in which the gesture apex occurs in the post-tonic (phrase-final) position.

Despite the high variability within and across participants, some patterns can be observed: (1) apexes occurring during the pre-tonic material are extremely scarce (3 cases in trochees and 4 cases in iambs, i.e. 4% and 5.4% respectively), and crucially all of them contain a pause between the pointing gesture and the upcoming speech; (2) in around one third of all instances, gesture apexes occur within the accented syllable (19 cases in trochees and 27 cases in iambs, i.e. 26.1% and 36.6% respectively); and (3) more than half of the participants produced the gesture apexes in phrase-final position,

irrespective of the metrical pattern (51 cases in trochees and 43 cases in iambs, i.e. 69.9% and 58% respectively).

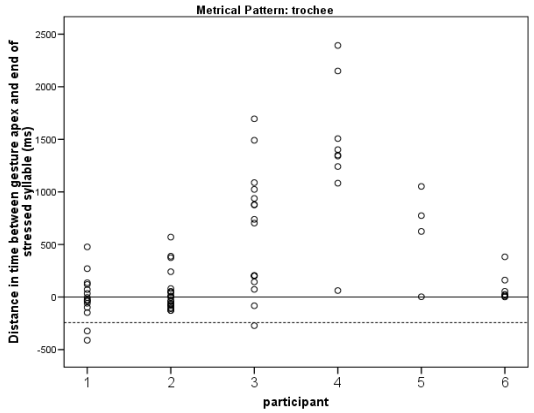


Figure 4: *Dispersion graph of the distance between gesture apex and end of the stressed syllable (in milliseconds) in trochaic words as a function of each participant.*

Chi-square tests indicated that the proportion of gesture apex occurring at a pre-tonic, tonic, or post-tonic position did not change across the two metrical patterns ($\chi^2(2) = 2.597$, $p > .05$). They also showed that the proportion of apexes at a pre-tonic position differed significantly from the proportion of apexes at tonic ($\chi^2(1) = 27.769$, $p < .001$) and post-tonic positions ($\chi^2(1) = 74.941$, $p < .001$), and a significant difference was also seen when comparing the proportion of apexes occurring at the tonic and post-tonic positions ($\chi^2(1) = 17.273$, $p < .001$).

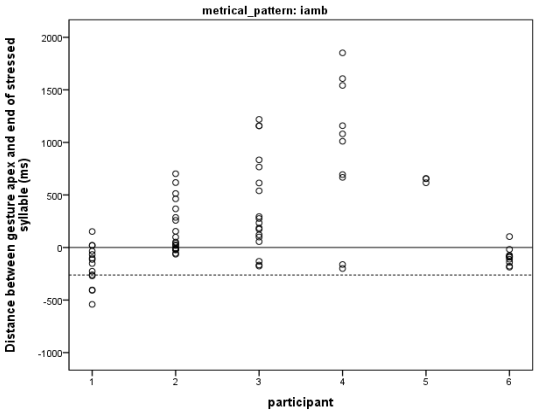


Figure 5: *Dispersion graph of the distance between gesture apex and end of the stressed syllable (in milliseconds) in iambic words as a function of each participant.*

We observed three strategies regarding the use of gesture and speech. The most frequent strategy was to utter a sentence and follow it by a pointing gesture, e.g. "Take the" [speech] + "object's name" [speech] + *it is this one* [gesture]. The second most frequent strategy was to utter unsynchronized pointing plus speech combinations, e.g. "Take the" [speech] + "object's name/ *it is this one* [gesture-speech combination]. And finally, there were few instances where the gesture apexes occurred during pre-tonic material, and these were produced with a pause between the pointing and the following word, e.g. "Take" [speech] + *this one* [gesture] + "which is called object's name" [speech]". These results show that pointing gestures can be produced before or after the target words, i.e. they are positioned at the edges of prosodic phrase boundaries, provided that they are perceived as separate speech acts carrying different semantic information.

4. Discussion

The purpose of this study was to investigate whether speakers detect temporal asynchrony between gesture and speech prominences (Experiment 1) and whether this perceptual ability is related to how they actually align gesture and speech in natural interactions (Experiment 2).

The results of Experiment 1 indicated that speakers do indeed detect asynchrony between gesture and speech prominences. However, surprisingly, unsynchronized trochees were perceived as more natural than unsynchronized iambs. More research is needed to investigate whether this effect is also found in trisyllabic words in which the misalignment of prominences can lead to an apex occurring at the pre-tonic or at the post-tonic position. This unexpected finding was further explored through a production experiment which elicited pointing gestures with the goal of teaching the name of the object and at the same time indicating its location. Our hypothesis was that speakers would rate unsynchronized trochees as fairly natural because in natural interactions speakers frequently align gesture prominences with phrase-final positions, especially when the semantic information conveyed by gesture is supplementary to the one conveyed in speech. Results of the production experiment (Experiment 2) confirmed this hypothesis: speakers produced practically no apexes during the pre-tonic material while apexes aligned during the post-tonic material were fairly frequent.

In our production study participants signaled the object they were referring to through a pointing gesture that frequently occurred after the object naming. It seems that speakers were actually saying "Pick up the object" using speech strategies + "that is there" using a pointing strategy. Thus, the gesture supplemented the meaning of speech and this affected the temporal coordination of the two modalities. This is not the first study showing evidence for the interrelation between semantic and temporal synchrony [9]. In [9] the authors found that gesture and speech timings were better aligned in complementary gesture-speech combinations than in supplementary gesture-speech combinations.

In sum, our results suggest that speakers perceive the alignment of gestural prominences by taking into account the temporal coordination of these gestures to prosodic heads (i.e. stressed syllables) or prosodic edges (i.e. phrase boundaries), and also by taking into account the semantic coordination of those gestures. Although further research is needed, this study has attempted to contribute to gain a better understanding of the temporal coordination between gesture and speech.

5. Acknowledgments

We thank Alfonso Igualada, Rafèu Sichel, and Santiago González for help with running the studies, and also the participants in the experiments. This research has been funded by grants FFI2012-31995, PSI-2011-25376, 2009SGR-701, and by the RECERCAIXA 2012 grant "Els precursors del llenguatge: una guia TIC per a pares i educadors".

6. References

[1] Birdwhistell, R. L. “Introduction to kinesics: An annotated system for analysis of body motion and gesture”. Washington, DC: Department of State, Foreign Service Institute, 1952.

[2] Kendon, A. “Gesticulation and speech: Two aspects of the process of utterance”. In M. R. Key (Ed.), The relationship of verbal and nonverbal communication (pp. 207–227). The Hague, the Netherlands: Mouton, 1980.

[3] McNeill, D. “Hand and Mind: What Gestures Reveal About Thought”. The Chicago University Press, Chicago, 1992.

[4] Wagner, P., Malisz, Z., and Kopp, S. “Gesture and speech in interaction: An overview”. Sp. Comm 57:209-232, 2014.

[5] Loehr, D. “Aspects of rhythm in gesture and speech”. Gesture 7: 179-214, 2007.

[6] Rusiewicz, H., Shaiman, S., Iverson, J., Szuminsky, N., Smith, A., and van Lieshout, P. “Effects of Prosody and Position on the Timing of Deictic Gestures”. J. Speech Lang. Hear. Res. 56(2):458-470, 2013.

[7] De Ruiter, J. P. “Gesture and speech production”. Doctoral dissertation. Katholieke Universiteit, Nijmegen, 1998.

[8] Esteve-Gibert, N. and Prieto, P. “Prosodic structure shapes the temporal realization of intonation and manual gesture movements”. J. Speech Lang. Hear. Res. 56(3): 850-864, 2013.

[9] Bergmann, K., Aksu, V., and Kopp, S. “The Relation of Speech and Gestures: Temporal Synchrony Follows Semantic Synchrony” in Proceedings of the 2nd Workshop on Gesture and Speech in Interaction. Bielefeld, Germany, 2011.

[10] Vatakis, A., and Spence, C. “Audiovisual temporal integration for complex speech, object-action, animal call, and musical stimuli”. In M. J. Naumer & J. Kaiser (Eds.), Multisensory Object Perception in the Primate Brain. New York, Springer, 2010.

[11] Jesse, A., and Mitterer, H. “Pointing gestures do not influence the perception of lexical stress” in Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011): 2445-2448, 2011.

[12] Treffner, P., Peter, M., and Kleidon, M. “Gestures and phases: the dynamics of speech-hand communication”. Ecol. Psychol. 20:32-64, 2008.

[13] Leonard, T. and Cummins, F. “The temporal relation between beat gestures and speech”. Lang. Cognitive Proc. 20(1):32-64, 2008.

Prominence and Coreference – On the Perceptual Relevance of F0 Movement, Duration and Intensity

Stefan Baumann[1], *Anna Roth*[2]

[1] I*f*L-Phonetik, Universität zu Köln, Germany
[2] Institut für Linguistik, Goethe-Universität Frankfurt am Main, Germany
stefan.baumann@uni-koeln.de, a.roth.unifrankfurt@gmail.com

Abstract

We conducted a web-based experiment on German testing the perception of an element's prosodic prominence in relation to its status as a potential coreferent of an antecedent. Data were elicited by asking subjects to judge the probability of a coreference relation between a context noun (antecedent) and a target word (anaphor), whose lexically stressed syllable was manipulated as to the parameters F0 movement, duration and intensity. Results suggest a direct but inverse relationship between prominence and coreference judgements indicating that the likelihood of a coreference interpretation decreases with increasing prosodic prominence. F0 movement turned out to be the dominant cue for prominence – as the main trigger for the perception of pitch accents – with rises being perceived as more prominent than falls. In turn, lack of tonal movement probably led to perceived deaccentuation and thus favoured the evaluation of a target word as being coreferential with an antecedent. Duration was found to be a significant factor as well, while intensity did not prove to be relevant for the task given. Thus, the present study with its revised methodology adds new aspects to the debate of which parameters are crucial for prominence perception, directly linking it to the investigation of information structure.

Index Terms: prosody, coreference, prominence, perception, pitch accent, givenness

1. Introduction

A central aspect in the analysis of an utterance's information structure is the information status of its elements, defining whether an element can be regarded as Given, Accessible or New (e.g. [1]). In many accounts, Givenness is equated with coreference, i.e. referential identity (e.g. [2]) (although it can be shown that a lexical level of description is also relevant for a comprehensive account of information status; see [3]). In (1a), e.g., *a lasagne* introduces a New referent, while in (1b), *the tasteless stuff* is Given, since it stands in a coreference relation to the previously mentioned *lasagne*:

(1) a. Yesterday, a friend of mine prepared a laSAgne for me.
b. I found it hard to enJOY the tasteless stuff. [4:16]

Prosodically, New referents are mostly marked by pitch accents in West Germanic languages (in the example, nuclear accents are indicated by capital letters, as in *lasagne* in (1a)), whereas coreferential anaphors are often deaccented (as *the tasteless stuff* in (1b)) or at least marked by a clearly attenuated, i.e. less prominent, prosody. Thus, there seems to be a more or less direct link between the information status of an element and the prominence of its prosodic marking. We will base our study on this general assumption.

There has been a long-standing debate on which acoustic parameters are most relevant for the perception of post-lexical prominence, in particular change in fundamental frequency (F0) (perceptual level: pitch movement), longer duration (perceptual level: increased length), or higher intensity (perceptual level: increased loudness). Another relevant parameter is vowel quality, differentiating between full and reduced vowels. Previous perception studies presented quite different results, assigning the greatest importance for prominence judgements either to F0 variation (e.g. [5], [6]), duration (e.g. [7]), intensity (e.g. [8]), or to a combination of duration and intensity (a factor called 'total amplitude' by [9]). Furthermore, there is conflicting evidence for the question of whether rising or falling pitch accents are preceived as more prominent (cf. [4] and [10]). See also [11] for an overview of acoustic and perceptual correlates of prosodic prominence.

There are a number of recent empirical (perception or production) studies on prosody and information status in German but they either do not differentiate between various levels of Givenness (conflating coreference, semantic-pragmatic accessibility and lexical repetition, e.g. [12]) or they do not relate an element's information status directly to its degree or level of prosodic prominence ([13], but see [4]). In the experiment we report on in the present paper we explicitly ask subjects to judge the probability of a coreference relation between an antecedent and a potential anaphor, the latter being manipulated as to its degree of prosodic prominence. This elicitation technique allows us to analyze the relation between coreference and perceived prominence, expressed in the following hypotheses:

Hypothesis 1: Acoustic/prosodic prominence correlates with perceived (non-)coreference – here: the more prominence-lending parameters (e.g. longer duration, higher intensity) are present, the less likely is a referent to be perceived as coreferential with an antecedent.

Hypothesis 2: Prosodic parameters vary in their relevance for the perception of prominence/non-coreference: F0 movement > duration > intensity.

2. Method

2.1. Test material

Three female proper names were chosen as target words: *Tamara*, *Pamela* and *Simone*. All of them are trisyllabic in German, with lexical stress on the second syllable. The stressed syllables contain the bilabial nasal /m/ in the onset and one of the long vowels /a:/, /e:/, /o:/ in the syllable rhyme. We created three test sentences. These contain a time specification (weekday), the subject pronoun *ich* ('I'), one of the three proper names as an accusative object, and a predicate

composed of the auxiliary verb *haben* ('have') and a past participle of one of the transitive irregular verbs *getroffen* ('met'), *gesprochen* ('talked') or *gesehen* ('seen'):

(2) Montag habe ich Tamara getroffen.
'On Monday I met Tamara.'

(3) Dienstag habe ich Pamela gesprochen.
'On Tuesday I talked to Pamela.'

(4) Freitag habe ich Simone gesehen.
'On Friday I saw Simone.'

The sentences were spoken by a female native German speaker and were recorded with an Edirol R-44 in a sound attenuated booth. The prosody of the recorded sentences was controlled for. In particular, three prosodic parameters of the stressed syllables in the target words were manipulated, namely F0 movement, duration and intensity.

F0 was varied to create rising, falling and level contours. For the rising condition, the stressed syllable starts at 190 Hz and rises to 240 Hz at the end of the syllable. For the falling condition, the F0 starts at 240 Hz at target syllable onset and falls down to 190 Hz at syllable offset. In the level condition, we assigned a value of 190 Hz throughout the whole target word. Figure 1 shows a *Praat* [14] screenshot for test sentence (2) with a rising pitch (solid line), a falling pitch (dotted line) and no pitch movement (dashed line) on the stressed syllable *ma* in the target word *Tamara*:

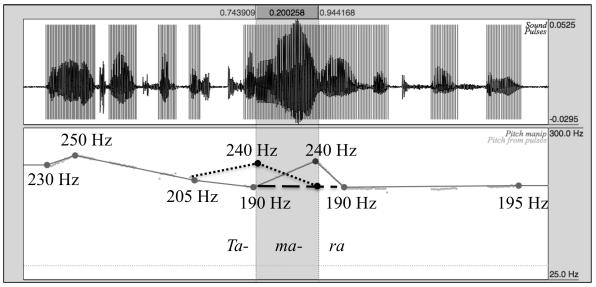


Figure 1: *Screenshot of the sentence* Montag habe ich Tamara getroffen ('On Monday I met Tamara') *in the Praat manipulation window. The stressed syllable of Tamara is shaded; it contains an F0 rise (solid line; first condition) an F0 fall (dotted line; second condition) or no F0 movement (dashed line; third condition).*

As to duration, two values were used: long syllables with 200 ms and short syllables with 150 ms. Similarly, we created two intensity levels: loud syllables with a maximum of 57 dB and soft syllables with a maximum of 47 dB. While F0 and duration were manipulated in *Praat*, we used the audio editor *Audacity* [15] to adjust the intensity of the target syllables. All combinations of the acoustic parameters were implemented in each sentence, resulting in 12 manipulated versions of each target sentence. Thus, 36 manipulated sentences for the perception experiment were generated in total. An overview of the manipulations is provided in Table 1.

The prosodic context for the target syllables was held constant throughout. That is, the weekday always carried a H* pitch accent (following [16]), with an F0 rise from 230 Hz to 250 Hz. It was crucial to create at least one pitch accent in the sentence which turns into the nuclear accent in cases of largely attenuated prominence of the target word. In this condition, the sentences would have lacked a pitch accent altogether without the early accent in the phrase, which would have sounded highly unnatural. Furthermore, the pitch height of the personal pronoun (*ich* 'I') immediately preceding the target word constantly had a value of 205 Hz, and the intonation phrase always ended at 195 Hz (cf. Figure 1).

Table 1: *Overview of manipulated prosodic parameters on target words.*

	Stressed syllable of target word	
F0 movement ('Tone')	rise	190Hz – 240Hz
	fall	240Hz – 190Hz
	none	190Hz – 190Hz
Duration	long	200 ms
	short	150 ms
Intensity	loud	57 dB
	soft	47 dB

Duration and intensity values for the rest of the target sentence were also controlled. In particular, the first and third syllable of the target word invariably had a duration of 150 ms and 130 ms, respectively, while both syllables were set to a maximum intensity of 50 dB.

2.2. Procedure

We conducted an online perception experiment via an open URL, using the professional software package *SoSci Survey* [17]. All 36 manipulated sentences were pseudo-randomized and matched with a respective context question. We created 12 different context questions for each target sentence, in order to diversify the task for the subjects. All questions followed the same structure; an example is given in (5):

(5) Hast du deine Cousine getroffen / gesprochen / gesehen?
'Have you met / talked to / seen your cousin?'

Twelve concrete, female nouns (as accusative objects) were used, which served as potential antecedents for the proper names in the target sentences: *Cousine* 'cousin', *Schwester* 'sister', *Nachbarin* 'neighbour', *Klassenkammeradin* 'classmate', *Tennispartnerin* 'tennis partner', *Arbeitskollegin* 'co-worker', *Mitbewohnerin* 'roommate', *Kommilitonin* 'fellow student', *Schulfreundin* 'schoolmate', *Teamkollegin* 'teammate', *Bekannte* 'acquaintance' and *Trainerin* 'trainer'.

The context questions were presented only visually, while the manipulated test sentences were presented auditorily. The participants controlled when to start a stimulus and were free to listen to it as many times as they chose. Afterwards, the subjects had to answer a question by rating the probability of a (non-)coreference relation between the proper name in the target sentence and the noun in the context question. An example of the questions is shown in (6):

(6) Für wie wahrscheinlich halten Sie es, dass es sich bei der Cousine um Tamara handelt?

'How likely do you think it is that the cousin is Tamara?'

Judgements were given via a horizontal scroll bar whose poles were labelled *sehr wahrscheinlich* 'very likely' (left side) and *sehr unwahrscheinlich* 'very unlikely' (right side). Figure 2 gives an example of the setup.

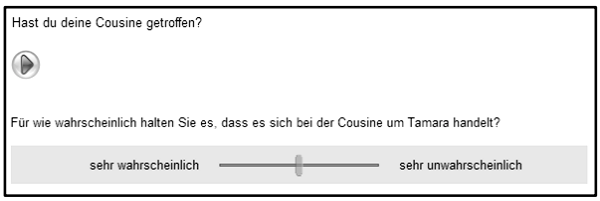


Figure 2: *Screenshot of the response submission setup in the online questionnaire, including a context question* ('Have you met your cousin?')*, the button for playing the target sentence, and the rating task* ('How likely do you think it is that the cousin is Tamara?') *with the scroll bar.*

The pole 'very likely' (= low values on the scale) corresponds to the judgement of coreference between the context noun (e.g. *Cousine*) and the proper name (e.g. *Tamara*). That is, the target word is perceived as Given information. In contrast, the pole 'very unlikely' (= high values on the scale) corresponds to the perception of a non-coreference relation between the context noun and the proper name. This means that the target word displays New information.

The experimental setup also included an instruction and a short practice section prior to the main experiment. The subjects were asked to conduct the experiment in a quiet environment and to wear headphones when listening to the stimuli. The entire procedure took approximately 10 to 15 minutes per subject.

2.3. Subjects and analysis

We collected judgements of 40 native speakers of German (32 female, 8 male), aged between 19 and 62 years. The mean age was 27 years. They grew up in nine different German Federal States. The subjects were no experts in the analysis of spoken language and did not report any hearing impairment.

The elicited judgements were encoded on an interval scale, illustrated as a horizontal continuous line in the experimental condition, ranging from 1% at the left pole ('very likely') to 100% at the right pole ('very unlikely'). For the statistical analysis, we used these percentage values as the dependent variable in a linear mixed model, with TONE (i.e. F0 movement), DURATION, INTENSITY and TEST WORD as fixed effects, and SUBJECT as a random effect.

3. Results and discussion

First of all, the subjects covered the whole range of the scale when judging the probability of corefence relations. This suggests that the stimuli were reasonably balanced for the type of task presented. Nevertheless, there is a slight bias towards the left pole of the scale (coreference judgements), maybe because a syntactically unmarked answer to a polar question (i.e. lacking an explicit 'yes' or 'no') tends to be interpreted as a confirmation.

Results reveal a significant main effect for TONE ($p<0.001$) in the coreference judgements. Figure 3 shows that the highest values were assigned to the target words if they carried a rising F0 movement, and that a falling contour still triggered clearly higher values than no F0 movement. That is, non-coreference (Newness) judgements correlate with tonal movement, and more so with rises (presumably more prominent; mean = 66.62%) than with falls (presumably less prominent; mean = 51.51%) whereas coreference (Givenness) judgements correlate with – the least prominent – lack of tonal movement (mean = 24.05%).

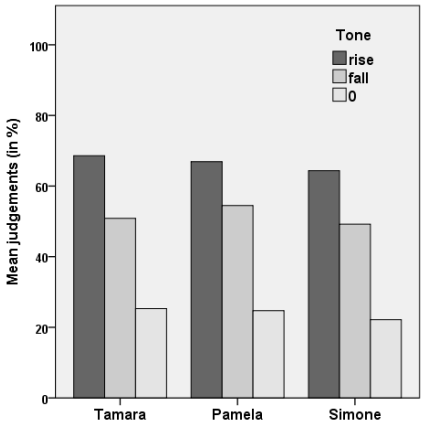


Figure 3: *Bar plot of mean judgements on (non-)coreference-probability scale with respect to the factor TONE; '0' stands for lack of tonal movement; all subjects pooled.*

The factor DURATION shows a significant main effect as well ($p<0.01$), with longer syllables favouring the subjects' impression that a coreference relation between context noun and proper name is unlikely (means = 49.63% for long syllables, 45.16% for short syllables; see Figure 4). Thus again, an increase in prosodic prominence leads to a decrease in coreference judgements.

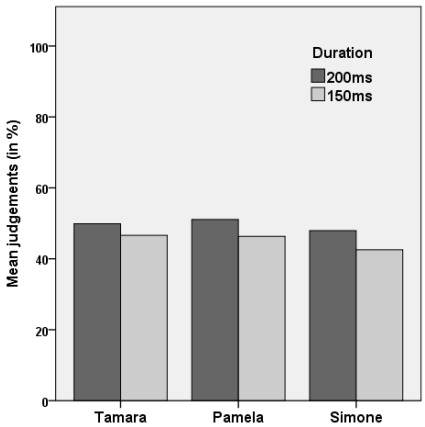


Figure 4: *Bar plot of mean judgements on (non-)coreference-probability scale with respect to the factor DURATION; all subjects pooled.*

No significant effect on coreference evaluations has been found for the parameter INTENSITY (p=0.131), although the judgements for two test words display the expected tendency, namely higher values (indicating lower probability of coreference) for test words carrying a louder stressed syllable (see Figure 5). In fact, the manipulation method for intensity may have been particularly prone to result in an unnatural outcome of this parameter (e.g. no adequate variation of spectral tilt) which may explain the somewhat inconsistent judgements.

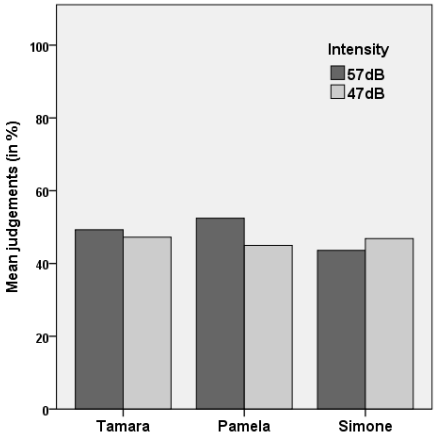


Figure 5: *Bar plot of mean judgements on (non-)coreference-probability scale with respect to the factor INTENSITY; all subjects pooled.*

The factor TEST WORD does not show a significantly distinct distribution over all three parameters either (p=0.085). Thus, the quality of the stressed vowel in the target word did not have an influence on the task.

Finally, no significant interactions were found between the fixed factors investigated. This result came as a surprise since an incremental effect of the prominence-lending parameters with respect to the (non-)coreference judgements would have been expected. That is, e.g., a target word with a rising F0 movement was not found to be interpreted as 'less coreferential' with an antecedent if it was *accompanied* by increased duration and intensity.

This notwithstanding, both hypotheses could generally be confirmed. As to Hypothesis 1, there are clear indications for an inverse but straightforward relation between perceived prosodic prominence and listeners' judgements of coreference. For example, if a target word such as *Tamara* in *Montag habe ich Tamara getroffen* ('On Monday I met Tamara') is realized in a prosodically attenuated manner, e.g. by lack of tonal movement or relatively short and soft syllables (in particular the lexically stressed syllable), the likelihood is high to interpret *Tamara* as the same referent as the cousin in the context sentence *Hast du deine Cousine getroffen?* ('Have you met your cousin?'). In contrast, a rising tone as well as increased duration on the lexically stressed syllable (*Tamara*) enhances the likelihood of the word to be interpreted as New information, which thus cannot be coreferential. However, as mentioned above, an additive effect of the parameters could not be found.

As to Hypothesis 2, the factor *tonal movement* is clearly dominant, while duration – which shows a significant effect as well – and intensity are much weaker cues (contrary to the claims proposed e.g. by [7] or [8]). This primacy of F0 variation is an old assertion already made by Fry [5], who based his findings on perception experiments in English, and Bolinger [6], who closely linked prominence with pitch accents (at least for West Germanic languages). Although pitch accents are combinations of F0 movement and increased duration and intensity, an F0 change in the vicinity of a stressed syllable is the defining characteristic of a pitch accent.

Moreover, the present experiment adds to the findings of more recent studies on West Germanic languages by suggesting that the shape or *type* of pitch accent does not only play a crucial role for the evaluation of Given or New information but also for the perception of prosodic prominence – with low and falling accents being perceived as less prominent than high and rising accents ([4], [13], [18]). In particular, there is evidence that the most important factor is the question of whether the *onglide* to the accented syllable is rising or falling ([19]).

4. Conclusions

The present study relates the interpretation of information status to the perception of prosodic prominence, assuming a basic link between them. Methodologically, it proved to be possible to use a gradient probability measure based on distinctions at various levels of prosodic parameters to evaluate the intrinsically categorical distinction between coreference (i.e. referential identity) and lack of coreference.

Results suggest that tonal movement is the central cue to perceived prominence, since it is characteristic of a pitch accent. Nevertheless, it is generally agreed on today that non-tonal cues such as duration, intensity and vowel quality are highly relevant for indicating the presence of *rhythmic* prominences (e.g. [20]) as well, which, however, may not contribute to the marking of information status in the same way as pitch accents do.

5. References

[1] Chafe, W., "Discourse, Consciousness, and Time", University of Chicago Press, Chicago/London, 1994.

[2] Prince, E. F., "Toward a Taxonomy of Given-New Information", in P. Cole [Ed], Radical Pragmatics, 223-256, Academic Press, New York, 1981.

[3] Baumann, S. and Riester, A., "Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects", in G. Elordieta and P. Prieto [Eds], Prosody and Meaning. Mouton De Gruyter, Berlin, New York, [Interface Explorations 25], 119-162, 2012.

[4] Baumann, S. and Riester, A., "Coreference, Lexical Givenness and Prosody in German", in J. Hartmann, J. Radó and S. Winkler [Eds], [Special Issue "Information Structure Triggers"], Lingua 136:16-37, 2013.

[5] Fry, D.B., "Experiments in the Perception of Stress", Language and Speech 1:126-152, 1958.

[6] Bolinger, D., "A Theory of Pitch Accent in English", Word 14:109-149, 1958.

[7] Cole, J., Mo, Y. and Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence", Laboratory Phonology 1:425-452, 2010.

[8] Kochanski, G., Grabe, E., Coleman, J. and Rosner, B., "Loudness Predicts Prominence; Fundamental Frequency Lends Little", J. Acoustical Society of America 11(2):1038-1054, 2005.

[9] Beckman, M. E., "Stress and Non-Stress Accent", Foris, Dordrecht, 1986.

[10] Hermes, D. and Rump, H., "Perception of prominence in speech intonation induced by rising and falling pitch movements", Journal of the Acoustical Society of America 90:97-102, 1994.
[11] Terken, J., and Hermes, D., "The perception of prosodic prominence", in M. Horne [Ed] Prosody: Theory and experiment, Kluwer, Dordrecht, 89-127, 2000.
[12] Féry, C. and Kügler, F., "Pitch accent scaling on given, new and focused constituents in German", Journal of Phonetics 36:680-703, 2008.
[13] Röhr, C. and Baumann, S., "Decoding information status by type and position of accent in German", Proceedings 17th ICPhS, Hong Kong, China, 1706-1709, 2011.
[14] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", Version 5.3.60, retrieved 4 March 2012 from http://www.praat.org/.
[15] "Audacity" [Computer software audio editor], Version 2.0.5, retrieved 30 October 2013 from: http://audacity.sourceforge.net/.
[16] Grice, M., Baumann, S. and Benzmüller, R., "German Intonation in Autosegmental-Metrical Phonology", in S.-A. Jun [Ed.] Prosodic Typology. The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, 55-83, 2005.
[17] Leiner, D. J., "SoSci Survey" (Version 2.3.05-i) [Computer Software]. Available from https://www.soscisurvey.de, 2013.
[18] Pierrehumbert, J. B. and Hirschberg, J., "The Meaning of Intonational Contours in the Interpretation of Discourse", in P.R. Cohen, J. Morgan, M.E. Pollack, [Eds] Intentions in Communication. MIT Press, Cambridge, 271-311, 1990.
[19] Ritter, S. and Grice, M., "The Role of Tonal Onglides in German Nuclear Pitch Accents", Oral presentation at Phonetics and Phonology in Iberia, Lisbon, 26 June 2013.
[20] Calhoun, S., "The centrality of metrical structure in signaling information structure: A probabilistic perspective", Language 86:1-42, 2010.

An acoustic study of Estonian word stress

Pärtel Lippus [1,2], *Eva Liina Asu* [1], *Mari-Liis Kalvik* [1,3]

[1] Institute of Estonian and General Linguistics, University of Tartu, Estonia
[2] Institute of Behavioural Sciences, University of Helsinki, Finland
[3] Institute of the Estonian Language, Estonia

partel.lippus@ut.ee, eva-liina.asu@ut.ee, Mari-Liis.Kalvik@eki.ee

Abstract

This study investigates the acoustic correlates of word stress in Estonian. It forms part of a broader international collaboration the aim of which is to develop a universal language-independent model for evaluating lexical stress regardless of the phonological structure of a given language. To this aim the characteristics of word stress in a range of languages are studied using unified methodology. For the present study, four acoustic measures were analysed as a function of speaking style and stress: vowel duration, F0 mean, F0 standard deviation, and spectral emphasis. The results show that the strongest correlate of style and stress in Estonian is vowel duration, but stress has a strong interaction with the Estonian three-way quantity system.

Index Terms: word stress, quantity, speaking style, Estonian

1. Introduction

The present study forms part of the Word Stress Project (http://wordstress.ling.gu.se/index2.html), which addresses the acoustic description of word stress correlates and the perception of these correlates as cues to relative syllable prominence. At present 7 languages are included: Brazilian Portuguese, British English, Estonian, French, German, Italian and Swedish. The results of the project will ultimately be used to formulate a typology of word stress suggesting methods that could be applied to the study of any language.

All languages included in the project are studied using the same methodology, whereby the acoustic correlates of stress are described as a function of stress level (primary, secondary and unstressed) and speaking style (spontaneous, phrase reading and word list reading). The acoustic correlates measured for each syllable of a word are the mean F0, standard deviation of the F0, vowel duration, and spectral emphasis. So far, results have been obtained for Swedish [1], [2], Brazilian Portuguese [3], [4], and German [5]. In this paper similar methodology will be applied to Estonian – an unrelated Finno-Ugric language which is known for its three quantity degrees.

In Estonian, the main word stress is usually fixed on the first syllable. It can be elsewhere only in foreign and loan words, interjections and names. As pointed out by Lehiste, the role of stress in Estonian is primarily identificational rather than contrastive [6]. The position of stresses, and thus the division of the word into feet, follows the trochaic rhythm structure. Disyllabic or trisyllabic words typically consist of a single foot (a primary stressed syllable is followed by one or two unstressed syllables). A tetrasyllabic word is generally made up of two disyllabic metric feet. Secondary stresses normally fall on successive odd-numbered syllables (e.g. *magamata* [ˈmɑ.kɑ.ˌmɑt:.tɑ] 'sleepless', *lõpetatigi* [ˈlʏp.pe.ˌtɑt.ti.ki] 'was finished'). In longer words the stress pattern is not fully predictable, although it is often determined by the morphology.

All primary stressed syllables in Estonian are in one of the three quantity degrees: short (Q1), long (Q2) and overlong (Q3). This distinction operates only over a disyllabic trochaic foot, and the decisive factor in determining the degree of quantity is the duration ratio between the first (stressed) and second (unstressed) syllable, regardless of the stressed syllable structure [7]. An additional cue to the quantity distinction is the pitch contour which in Q1 and Q2 is a step down between the two first syllables, but in Q3 is realised as an early fall in the first syllable.

Word stress in Estonian has been subject to thorough phonological investigations [8]–[10] whereas very little is known about its phonetic characteristics. Earlier studies addressing acoustic correlates of Estonian stress are limited to a small-scale investigation into the influence of stress on nasal flow, amplitude and duration [11], where duration was shown to be the parameter on which stress had most consistent influence. The temporal patterns of stressed and unstressed syllables have also been studied in relation to the quantity system. A study of penta- and hexasyllabic words [12] showed that the secondary stressed feet were systematically shorter than the stressed feet while the stressed-to-unstressed syllable ratio was retained to a certain extent. Additionally, there is some acoustic evidence which shows that the degree of stress in Estonian has an influence on vowel quality [13], [14].

The aims of the present study are two-fold: firstly, to test the suitability of the methodology of the Word Stress Project on Estonian, which is a typologically different language from the others included in the project so far, and secondly, to gain new knowledge about the little-studied acoustic correlates of word stress in Estonian. As the main stress in Estonian is fixed it could be hypothesised that stress correlates are fewer and not as strong as in languages with a morphologically unbound stress like, for instance, English. We predict vowel duration to be the most important correlate.

2. Materials and method

Identical methodology as used by the other studies carried out in the framework of the international Word Stress Project (e.g. [1], [3], [5]) was applied. The starting point of the materials was the University of Tartu Phonetic Corpus of Spontaneous Speech (http://www.keel.ut.ee/foneetikakorpus). Spontaneous recordings of 16 native speakers of Estonian – 10 females and 6 males – were chosen from the corpus on the basis of the speaker's age and dialectal background. All speakers were between 22 and 34 years old, and represented the Standard North Estonian variety. At the time of the recording they were either students or graduates of the University of Tartu. From each (approximately 30 min long) recording 15–20 utterances were selected per speaker, and from each utterance one or two target words were chosen. The two criteria for selecting the

words were the position of the word in the utterance (prenuclear i.e. not preceded or followed by an IP-boundary) and the number of syllables (at least two or preferably more syllables long). The words therefore display various syllable structures and phonological quantities. Two lists were prepared with these utterances and target words where each phrase and word appeared three times in random order. The two lists were read by the same 16 speakers.

The final data set comprises 276 target words in three different speaking styles: spontaneous speech, phrase reading and word list reading. Most commonly, in our data, a target word is trisyllabic.

The recordings were analysed in Praat [15]. The analysis focussed on the vowels of the target words which were manually labelled. Combined labels were given for stress and quantity. Stress was marked with numbers 3 (primary), 2 (secondary), and 1 (unstressed), and depending on the quantity of each foot the relevant number was used once (Q1), twice (Q2) or three times (Q3), as for instance: 3+1+1 in case of a Q1 foot (e.g. *midagi* [ˈmi.tɑ.ki] 'something'), 33+11+11 in case of a Q2 foot (e.g. *rongile* [ˈroŋ.ki.le] 'train AllSg'), and 333+111+111 in case of a Q3 foot (e.g. *eelmine* [ˈee:l.mi.ne] 'previous'). In tetrasyllabic and longer words where successive feet carry secondary stress the labelling was following: 3+1+2+1 for two Q1 feet (e.g. *esimene* [ˈe.si.ˌme.ne] 'first'), 33+11+22+11 for two Q2 feet (e.g. *värviliste* [ˈvær.vi.ˌlis.te] 'coloured one GenPl'), and 3+1+222+111 for a combination of a Q1 and Q3 foot (e.g. *mesilasse* [ˈme.si.ˌlɑs:.se] 'beeyard IllSg').

Stress and quantity levels were identified by a trained phonetician. The measures of duration, F0 mean, F0 standard deviation, and spectral emphasis were extracted using a Praat script. For spectral emphasis the formula $SE = L - L_0$ was used, where L is the entire intensity of the segment spectrum and L_0 is its intensity in the low band from 0 to 43 % over the F0 median in Hertz (see [16] for details). The data was analysed statistically in R. The mean values of the measures were calculated for each speaker for the factors Style, Stress and Quantity. Three-way Anova tests were used to evaluate the significance of the factors, and post-hoc tests were carried out with Tukey HSD.

3. Results and discussion

The results will be presented separately for each measure studied. For duration the data from male and female speakers was pooled as there is no reason to presume any gender related differences. For F0 and spectral emphasis the analysis was carried out separately for male and female speakers.

3.1. Vowel duration

It can be seen that vowel duration is significant for all factors: Style $F(2, 399) = 97.4$, $p < 0.001$; Stress $F(2, 399) = 69.4$, $p < 0.001$; Quantity $F(2, 399)$ 36.5, $p < 0.001$, and there is a weak interaction of Style and Stress $F(4, 399) = 2.7$, $p < 0.05$, and a strong interaction of Stress and Quantity $F(4, 399) = 41.8$, $p < 0.001$. Post-hoc testing showed that Style is significant on all levels, duration being the longest in word list reading and the shortest in spontaneous speech. For Stress there are diversely different patterns for different quantity levels: in Q2 and Q3, vowel duration is the longest in primary stressed syllables and the shortest in unstressed syllables; but conversely in Q1 the unstressed syllables are the longest and the primary stressed syllables the shortest.

The strong interaction of Stress and Quantity resulting in diversely different stressed-unstressed duration patterns is in line with previous findings from the studies of Estonian quantity system, e.g. [14]. Additionally it can be seen from Figure 1 that there is a large variation of the stressed syllable durations in Q2 and especially in Q3. This is due to the fact that in the present analysis we only considered the foot level quantity but not the stressed vowels being short or long. As pointed out above, the Estonian three-way quantity operates over a disyllabic foot and the main feature of quantity is the relative temporal pattern of stressed and unstressed syllables. The stressed syllables can be filled by various segmental combinations, including a long or an overlong vowel or a combination of a short or a long vowel followed by a short or long coda consonant. Therefore, a more accurate description of stress patterns in Estonian could be obtained by taking the phonological structure into account.

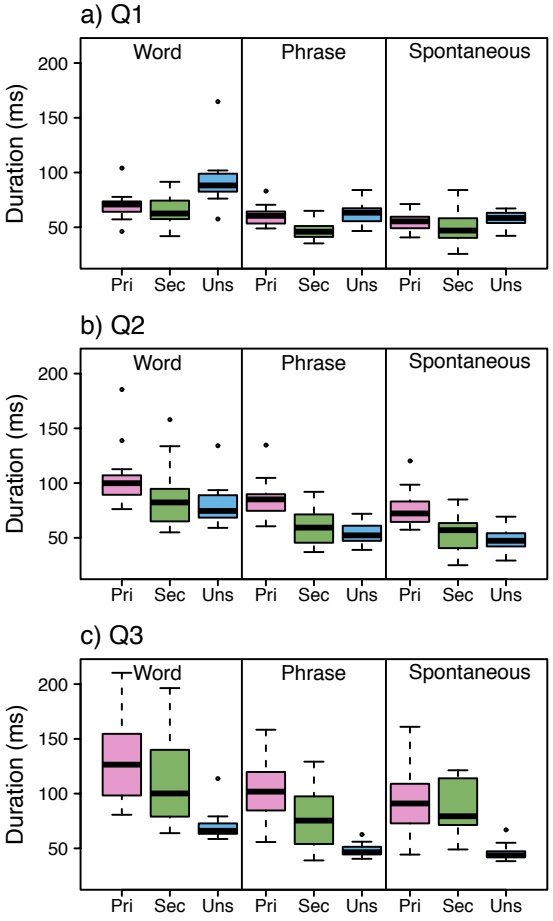


Figure 1: *Duration as a function Style (word list reading, phrase reading, spontaneous) and Stress (primary, secondary, unstressed), grouped by Quantity (Q1, Q2, Q3).*

3.2. F0 mean

For male speakers, the mean pitch is significant for Style $F(2, 151) = 14.9$, $p < 0.001$, and Stress $F(2, 151) = 43.6$, $p < 0.001$. For female speakers, there is only a significant effect of Stress

$F(2, 255) = 17.6$, $p < 0.001$. Post-hoc tests showed that for both male and female speakers, the F0 mean is higher in the primary stressed syllables, but there is no difference between secondary and unstressed vowels. For male speakers, the F0 mean is higher in read speech than in spontaneous speech, but there is no significant difference between the word list reading and phrase reading.

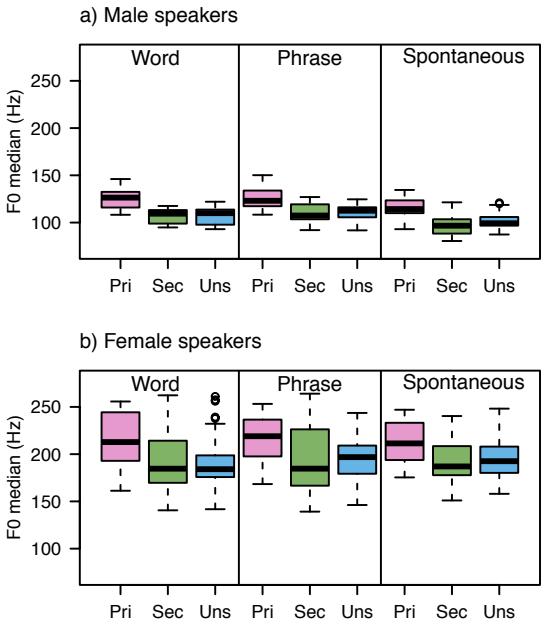


Figure 2: *F0 median as a function Style (word list reading, phrase reading, spontaneous) and Stress (primary, secondary, unstressed).*

Unexpectedly, there was no effect of Quantity on the F0 mean, although it is known from previous studies that there is more pitch movement in Q3 primary stressed syllables.

On the side we noticed a considerable amount of creaky voice in non-initial syllables that were left out of the present analysis. Creakiness usually occurs in Estonian in less prominent parts of speech and can be associated with the lack of prominence. Therefore, some measure of creakiness should be included in the stress model.

3.3. F0 variation

The F0 variation within vowels has different patterns for male and female speakers. For male speakers, there is a significant effect of Stress $F(2, 146) = 12.9$, $p < 0.001$ and a weak interaction of Stress and Quantity $F(4, 146) = 2.4$, $p < 0.05$. Post-hoc testing showed that the F0 variation is greater in primary stressed and unstressed syllables than in secondary stressed syllables. The interaction of Stress and Quantity indicates that the variation of F0 is somewhat different in primary stressed vowels of Q1 vs. Q3 ($p < 0.05$), but in secondary and unstressed syllables the quantity has no effect.

For female speakers, there is a significant effect of Stress $F(2, 249) = 7.4$, $p < 0.001$, and a weak interaction of Style and Stress $F(4, 249) = 2.7$, $p < 0.05$. Post-hoc testing showed that unlike for male speakers the variation of F0 is greater in unstressed vowels, and there is no significant difference between primary and secondary stressed vowels. The post-hoc testing of the interaction of Style and Stress did not reveal any meaningful patterns.

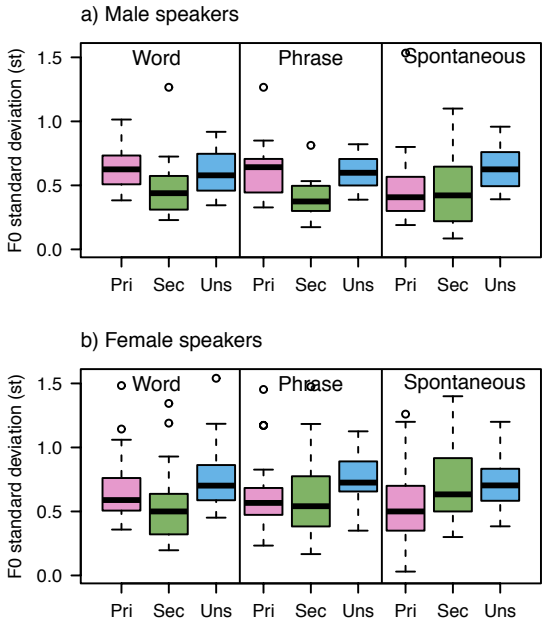


Figure 3: *F0 standard deviation as a function Style (word list reading, phrase reading, spontaneous) and Stress (primary, secondary, unstressed).*

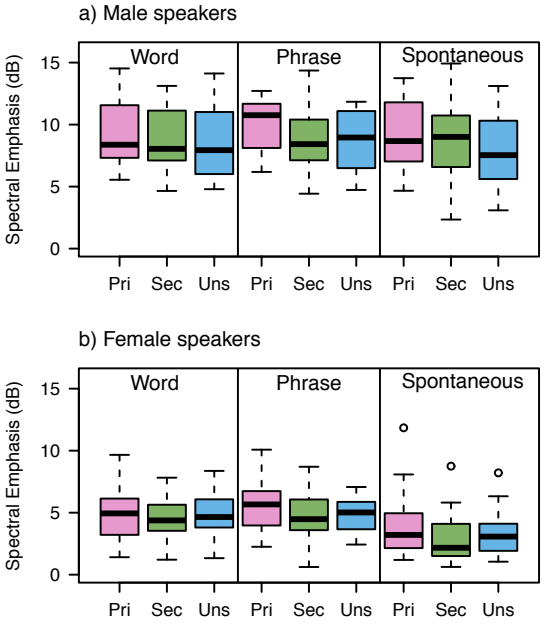


Figure 4: *Spectral emphasis as a function Style (word list reading, phrase reading, spontaneous) and Stress (primary, secondary, unstressed).*

An interaction between Stress and Quantity was expected and could have been even stronger, because it is known that the F0 in Q1 and Q2 stressed syllables is flat, whereas there is an F0 fall within the Q3 stressed syllable [14]. Also, it is

possible that a larger F0 variation in unstressed syllables is linked to the fact that a more marked F0 movement takes place in Q1 and Q2. A greater variation in unstressed syllables was also shown for German [5].

3.4. Spectral emphasis

Spectral emphasis is the weakest measure for Style and Stress in Estonian. Again there is a difference between the genders. None of the tested factors are significant for the male speakers, while the data of the female speakers only shows a significant effect of Style $F(2, 253) = 19.7$, $p < 0.001$. A post-hoc test showed that the spectral emphasis is greater in read as compared to spontaneous speech. Additionally, for female speakers, the effect of Stress closely fails to reach the significance level $F(2, 253) = 2.9$, $p = 0.0595$, the spectral emphasis being somewhat higher in primary stressed vowels than in secondary and unstressed syllables.

4. Conclusions

The application of the methodology of the Word Stress Project yielded new results about the acoustic correlates of Estonian word stress. Four measures were studied: duration, F0 mean, F0 standard deviation, and spectral emphasis.

As predicted, vowel duration turned out to be the most important stress correlate. It was shown to be the strongest for Style and Stress with there being a strong interaction with Quantity. In Q2 and Q3 the stressed vowels were longer than the unstressed ones, but in Q1 the unstressed syllables were the longest. The secondary stressed vowels were always shorter than the primary stressed ones.

The pitch in primary stressed vowels was higher than in secondary and unstressed vowels, while the variation of pitch was high in stressed and unstressed vowels and lower in secondary stressed vowels. The F0 variation could be partly quantity-related, but the nature of the data does not enable to draw a solid conclusion. Additionally, F0 mean was associated with speaking style, although this was only apparent for male speakers whose F0 mean was higher in read than in spontaneous speech.

Spectral emphasis turned out to be the weakest measure of stress for Estonian, being only significant for female speakers, distinguishing read from spontaneous speech.

The results seem to support the hypothesis about there being fewer and weaker stress correlates in Estonian. Since Estonian has a fixed word stress and very clear phonological rules determining the stress patterns, it seems logical that it is not as important to mark stress acoustically as it is in many other languages.

In this study neither the duration of the stressed vowel nor the phonological structure of the stressed syllable was taken into account. Thus, in further work on Estonian word stress, not only the foot level quantity but also the segmental structure of the foot should be factored in.

5. Acknowledgements

The authors are very grateful to the speakers who participated in this study. We would also like to thank the participants of the 2nd Word Stress Workshop for feedback and comments. The study was supported by the Estonian Research Council grant IUT2-37, the targeted financing project SF0050023s09, and the National Program for the Estonian Language Technology Project EKT4.

6. References

[1] A. Eriksson, P. A. Barbosa, and J. Åkesson, "Word stress in Swedish as a function of stress level, word accent and speaking style," in *Nordic Prosody. Proceedings of the XIth conference, Tartu 2012*, E. L. Asu and P. Lippus, Eds. Frankfurt am Main: Peter Lang, 2013, pp. 127–136.

[2] A. Eriksson, P. A. Barbosa, and J. Åkesson, "The acoustics of word stress in Swedish: A function of stress level, speaking style and word accent," in *Proceedings of Interspeech 2013*, 2013, pp. 778–782.

[3] P. A. Barbosa, A. Eriksson, and J. Åkesson, "Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese," in *Nordic Prosody. Proceedings of the XIth conference, Tartu 2012*, E. L. Asu and P. Lippus, Eds. Frankfurt am Main: Peter Lang, 2013, pp. 97–106.

[4] P. A. Barbosa, A. Eriksson, and J. Åkesson, "On the robustness of some acoustic parameters for signalling word stress across styles in Brazilian Portuguese," in *Proceedings of Interspeech 2013*, 2013, pp. 282–286.

[5] J. Behrens, "Die Prosodie des Wortakzentes in Abhängigkeit von Akzentlevel und Sprechstil," BA Thesis, Christian-Albrechts-Universität zu Kiel, 2013.

[6] I. Lehiste, "Search for phonetic correlates in Estonian prosody," in *Estonian Prosody: Papers from a Symposium*, I. Lehiste and J. Ross, Eds. Tallinn: Institute of Estonian Language, 1997, pp. 11–35.

[7] I. Lehiste, "Segmental and syllabic quantity in Estonian," in *American Studies in Uralic Linguistics*, T. A. Sebeok, Ed. Bloomington: Indiana University Publications, 1960, pp. 21–82.

[8] M. Hint, *Eesti keele sõnafonoloogia I: Rõhusüsteemi fonoloogia ja morfofonoloogia põhiprobleemid*. Tallinn: Eesti NSV Teaduste Akadeemia, 1973.

[9] M. Hint, *Häälikutest sõnadeni*, 2. ed. Tallinn, 1998.

[10] T.-R. Viitso, "Phonology, morphology and word formation," in *Estonian language*, M. Erelt, Ed. Tallinn: Estonian Academy Publishers, 2003, pp. 9–92.

[11] M. Gordon, "Phonetic correlates of stress and the prosodic hierarchy in Estonian," in *Estonian prosody: Papers from a symposium*, I. Lehiste and J. Ross, Eds. Tallinn: Institute of Estonian Language, 1997, pp. 100–124.

[12] P. Lippus, K. Pajusalu, and P. Teras, "The Temporal Structure of Penta- and Hexasyllabic Words in Estonian," in *Proceedings of the 3rd International Conference Speech Prosody*, R. Hoffmann and H. Mixdorff, Eds. Dresden: TUDpress, 2006, pp. 759–762.

[13] A. Eek and E. Meister, "Quality of standard Estonian vowels in stressed and unstressed syllables of the feet in three distinctive quantity degrees," *Linguistica Uralica*, vol. 34, no. 3, pp. 226–233, 1998.

[14] P. Lippus, E. L. Asu, P. Teras, and T. Tuisk, "Quantity-related variation of duration, pitch and vowel quality in spontaneous Estonian," *Journal of Phonetics*, vol. 41, no. 1, pp. 17–28, Jan. 2013.

[15] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. Computer program, 2013.

[16] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

Prominence Contrasts in Czech English as a Predictor of Learner's Proficiency

Lenka Weingartová,[1] Kristýna Poesová,[2] Jan Volín[1]

[1]Institute of Phonetics, Charles University, Prague, Czech Republic
[2]Department of English Language and Literature, Charles University, Prague, Czech Republic
lenka.weingartova@ff.cuni.cz, k.poesova@pedf.cuni.cz, jan.volin@ff.cuni.cz

Abstract

The study investigates prominence patterns in Czech-accented English comparing the production of non-native speakers of English at two distinct stages of phonological acquisition (beginners and intermediates) with a native performance. Word stress in Czech is entirely different from English, it has a fixed position, a delimitative function and rather impalpable acoustic manifestations. Alternations in the realization of word stress were analyzed by measuring the ratios or differences of acoustic correlates of prominence: duration, fundamental frequency, sound pressure level and spectral slope. Since word stress is a relational phenomenon, these characteristics were measured in two adjacent syllables one of which was a canonical stress bearer. The results reveal a clear difference between native and non-native treatment of word stress in all parameters examined. In the non-native sample distinct interferences of L1 across the two groups were detected: the subjects displayed different exploitation of duration, spectral slope and sound pressure level (SPL) with relation to their proficiency in L2 English. Out of these, duration ratio proves to be the most significant correlate. Furthermore, our findings indicate a strong effect of prosodic context coinciding with the prominence features, particularly in intonation declination and phrase-final lengthening.

Index Terms: Czech English, word stress, prominence, duration, F0, SPL, spectral slope

1. Introduction

Non-native accents of English have gradually ceased to be frowned upon, firstly due to the vast interconnectedness of today's world requiring the ability to understand a wide range of interlocutors and secondly thanks to the continuous research offering valuable insights into the nature of foreign-accentedness and its impact on human communication. One such groundbreaking study revealed that even heavily accented speech may be objectively intelligible [1], however the authors in the same breath warn of subjectively perceived difficulty to comprehend strongly accented productions which may consequently trigger negative attitudes or result in prejudiced judgments [2], [3]. In the field of second language pedagogy it has been repeatedly argued that appropriate prosodic features may contribute greatly to successful communication, increase intelligibility e.g., [4], [5], [6] and weaken the degree of foreignness [7]. Nevertheless, phonological transfer from the mother tongue places many constraints on the acquisition of the target sound system, especially in the area of rhythm and intonation [8].

The present study aims at mapping the relationship between L1 (Czech) and L2 (English) at different stages of phonological acquisition in the area of prominence patterning. One of the main functions of stress in English is to maintain the natural rhythm in connected speech [9], [10]. Stress misplacement may therefore bring about momentary confusion on the listeners' part as the deviated surface form does not immediately match the expected prominence scheme. Stress is a relational phenomenon and its degree of prominence cannot be determined within monosyllabic isolated words. The juxtaposition of stressed and unstressed syllables makes stressed parts stand out clearly which leads to their smoother identification and processing. The decreased degree or even absence of prominence in unstressed syllables corresponds perceptually to relative shortness, lower pitch, quietness and vowel reduction involving a centralization of peripheral vowel qualities, which is most frequently materialized as *schwa* [11]. The stress systems of English and Czech do not coincide, the former being completely unpredictable for a Czech speaker with the experience of a fixed stress mother tongue and a very limited space for reductions in standard pronunciation.

In general, word stress is not a one-dimensional phenomenon as it manifests itself in different ways in the speech signal. Several reliable acoustic correlates of word stress have been identified: F0, duration, intensity and spectral slope, e.g., [12, 13, 14, 15, 16, 17, 18, 19]. However, the way these parameters are manipulated by the speakers in order to create prominence contrasts can be highly language-specific, see e.g., [20].

In Czech, word stress remains somewhat impalpable and its comprehensive description is yet to be completed. Contrary to English, its position is fixed to the first syllable and its acoustic manifestations do not usually include an increase of F0, intensity or duration. Furthermore, unstressed syllables contain predominantly full vowel qualities. Duration is mainly used for marking phonologically long and short vowels and the stressed syllables can be shorter than the unstressed ones [21]. Also, F0 of stressed syllables is often lower due to a post-stress rise (i.e., L*+H accent) typical of Czech stress-groups [22]. The results concerning intensity and spectral slope are still preliminary and these acoustic attributes are an object of ongoing research.

The issue of prominence in Czech English has enjoyed an increased interest among researchers in both perception and production domains in the past decade. In the former, Czech listeners demonstrate a weakened perceptual sensitivity to English vowel reductions [23] and experience greater difficulties identifying strong beats when individual acoustic attributes indicating syllable prominence are in conflict [24]. As far as production is concerned, Czech-accented English can be generally characterised as lacking clear temporal contrasts, that is to say stressed syllables tend to be shorter and unstressed syllables longer than in native speech, which may, however, in certain contexts give rise to positive L1 transfer [25]. The strong inclination of Czech speakers to equalize the duration of vowels with a different degree of prominence seems to be a straightforward fact, although at a closer inspection, various strategies for signalling prominence have been detected. These largely depend on a number of factors, such as phonological structure of words, their textual frequency and resemblance to Czech counterparts [26]. Turning to weak vowels, the measurement of their spectral slope yielded quite reliable results in discriminating Czech and

British speakers of English [27]. Prominence investigation was also opted for in the assessment of strength of foreign accentedness under adverse listening conditions [28].

The current study follows and further develops the previously mentioned research in two respects. First, it compares the ratios or differences of acoustic parameters in stressed and adjacent unstressed vowels instead of comparing the values of duration, F0, intensity and spectral slope in grand means. We believe that this method better reflects the relational character of the explored phenomenon and is consistent with the idea of the speakers' effort to create local contrasts between prominent and non-prominent elements. Second, Czech respondents were subdivided into two groups according to the achieved language competence (beginners A1/A2 × intermediates B2/C1), which allows us to monitor the development of interlanguage at two distinct stages of phonological acquisition. Native speakers' prominence patterns served as an important reference point. Since a single corpus on Czech-accented speech involving various language levels is not available to the best of our knowledge, we drew the data from two different corpora with a high degree of compatibility [29], [30]. Although we may hypothesize that the more proficient speakers will exploit prominence features in a more native-like manner, we do not know which of the investigated prosodic dimensions will be most amenable to the acquisition process. Thus, the results should be generally informative with respect to prosodic pattern acquisition.

2. Method

2.1. Czech beginners

Czech beginners, 34 pupils of a lower-secondary school in Prague were asked to read aloud a list of 12 words, 8 short sentences and a limerick. The choice of lexicon corresponded to the respondents' language level or was lower. 18 boys and 16 girls at the age of 12 or 13 were recorded individually with a portable professional device Edirol HR-09, sampling frequency of 48 kHz and 16-bit resolution. The subjects reported no or rather limited contact with native English outside the classroom setting. All respondents were perceptually assessed as having a heavy foreign accent.

2.2. Czech intermediate and native British speakers

As for the Czech intermediate and British native speakers, their recordings were taken from the Prague Phonetic Corpus [29]. The subjects were 16 (8 Czech, 8 British) female non-professional speakers ranging in age from 20 to 25 years. They were instructed to read a news bulletin from the BBC in the length of about 500 words. The British subjects were speaking with a Southern Standard British accent, whereas the Czech subjects were selected as having a clear Czech accent, which was afterwards verified by naive listeners in a perception test (see [31]). Nevertheless the speakers were advanced enough to read a relatively complex text without substantial dysfluencies.

The recordings of the Czech intermediate speakers were made in a sound-treated studio with an electret microphone IMG ECM 2000, soundcard SB Audigy 2 ZS, 32-kHz sampling frequency and 16-bit resolution. The English speakers were recorded in the same way as the group of Czech beginners. All the recordings obtained with the Edirol HR-09 were afterwards downsampled to 32 kHz. Word and phone boundaries in both corpora were manually labelled in *Praat* [32] by experienced phoneticians.

2.3. Data selection

For this study, we selected comparable material from both corpora – pairs of vowels in adjacent syllables within a word, where one carries primary stress and the other is unstressed according to the lexical rules [33]. This resulted in two possible patterns – an unstressed vowel following a stressed vowel (S-U pattern, constituting 67.6 % of all pairs) and vice versa (U-S pattern, constituting the remaining 32.4 %).

Tokens containing dysfluencies, external noise or mispronounced elements were discarded. In total, we analyzed 1913 words: 884 uttered by beginners, 493 by intermediates and 536 by native speakers. The following parameters were extracted from or determined for each vowel:

- identity of the vowel (similar vowels clustered into six resulting types: /i/, /e/, /a/, /o/, /u/, /ə/)
- canonical prominence status (stressed, S × unstressed, U)
- position within a prosodic phrase (non-final × final)
- duration (in milliseconds)
- sound pressure level (in dB)
- F0 (in semitones relative to 100 Hz)
- spectral slope quantified with the α measure (difference between spectral energy above and below 1000 Hz, the highest boundary being the Nyquist frequency, that is 16 kHz; expressed in dB)

The last three characteristics were taken in the middle third of each vowel's duration to reduce the effects of consonantal transitions and potential labelling inaccuracies. F0 was not extracted from the beginners' word-list task, since it could be confounded by specific melodies of isolated word reading.

In the analyses, we compared the neighbouring vowels in the selected acoustic parameters. Our variables were:

- duration ratio (U/S)
- F0 difference (S-U)
- SPL difference (S-U)
- α difference (S-U)

The duration variable was expressed as a ratio in order to neutralize the effects of speaking tempo and differences in number of syllables in the examined words. The statistical significance of the results was verified by one- or two-way ANOVAs for independent measures with Tukey HSD post-hoc tests.

3. Results

3.1. Duration

The ratio of duration (unstressed/stressed vowel) in phrase non-final words decreases as the proficiency of the speaker group increases. The BrE speakers exhibit an average ratio of 0.7 (i.e., the unstressed vowel duration constitutes seven tenths of the stressed vowel duration), regardless of the stress pattern type. The CzE speakers differed according to their proficiency level and the investigated stress pattern.

Figure 1 shows how the duration ratio changes in the three speaker groups and in S-U vs. U-S patterns. The interaction of LEVEL*PATTERN is statistically significant (two-way ANOVA: $F(2, 892) = 6.5$, $p = 0.002$). Unlike BrE speakers, the two Czech groups yield different results in their temporal treatment of stress patterns. In U-S, they are more native-like – the difference between the CzE intermediate group and the native speakers is not even significant, whereas in S-U, they treat the temporal ratio quite differently: the beginners prolong the

second, unstressed syllable (the ratio is higher than 1) and the intermediates display roughly the same duration of both.

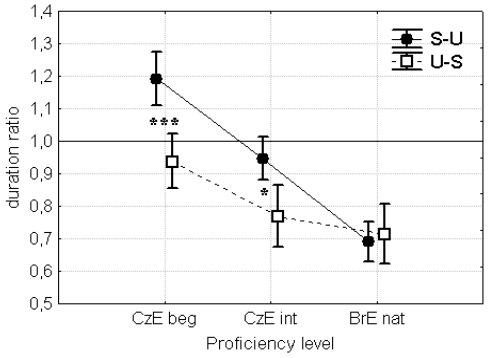


Figure 1: *Average duration ratio of adjacent stressed (S) and unstressed (U) vowels in non-final words of two patterns: S-U and U-S. Whiskers indicate 0.95 conf. intervals. Tukey post-hoc tests: *** → $p < 0.001$, * →$0.01 < p < 0.05$.*

The durations in phrase final words are not easy to interpret, since final lengthening influences the situation in a more arbitrary manner. Nevertheless, our data suggest that phrase-final lengthening may intensify the differences stemming from the stress pattern. This tendency applies to all three groups. In the case of U-S sequence, the effects of both lengthenings combine, resulting in lower ratios than in non-final words. In S-U cases it works in the opposite direction and equalizes the duration values making the ratio around one or higher.

3.2. Sound Pressure Level

In the case of sound pressure level differences, we discovered that Czech speakers of both proficiency levels behave identically – the difference between stressed and unstressed vowels was around 0.7 dB. Conversely, native speakers mark the stressed vowels substantially – the stressed elements have on average 3.3 dB higher sound pressure level than the neighbouring unstressed ones (see Figure 2); one-way ANOVA: $F(2, 1910) = 70.6$, $p < 0.001$.

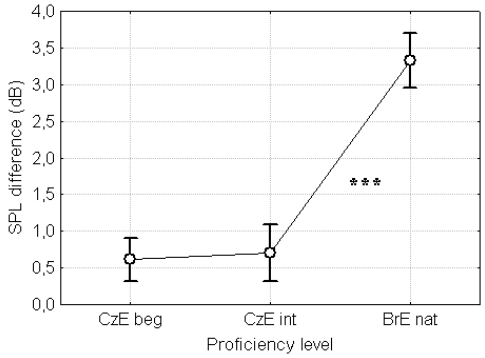


Figure 2: *Average SPL difference between adjacent stressed and unstressed vowels.*

The differences in SPL between words in final and non-final prosodic positions were negligible in all groups. However, when we examined the two stress patterns individually, we discovered a significant difference in the beginners group's final words. In U-S the SPL of the first vowel is 1.5 dB higher, whereas in S-U it is the other way round (the average difference is -1.9 dB). Thus the first syllable of phrase-final words of CzE beginners has always higher SPL regardless of its canonical stress status.

On the other hand, both BrE and CzE intermediate speakers maintain a constant SPL difference between the stressed vowel and its unstressed neighbour, regardless of the position or stress pattern. The former group has an average difference of 3.3 dB and the latter of 0.7 dB.

3.3. F0

Taking into consideration the typical Czech pattern of post-stress pitch rise, it seemed necessary to analyze both S-U and U-S groups separately to avoid its possible influence.

Generally, it is possible to infer from our data that Czech speakers do not use F0 difference to mark the stressed vowel, irrespective of the proficiency level, while the British speakers do, albeit only in the S-U pattern (see Figure 3). The stressed syllable of the native speakers is on average about 1.7 ST higher than the adjacent unstressed one and a Tukey post-hoc test reveals a highly significant difference between S-U and U-S of the British speakers: $p < 0.001$. Both Czech groups in both stress patterns fluctuate around zero and none of the differences between beginners and intermediates is significant.

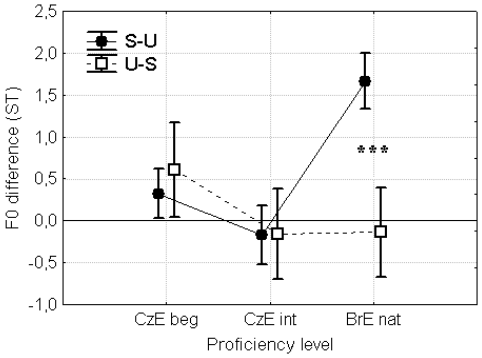


Figure 3: *Average F0 difference between adjacent stressed (S) and unstressed (U) vowels in two orders: S-U and U-S.*

We also investigated prosodically final words on their own as the phrase-final lowering (enhanced declination) could have an effect on the F0 stress pattern. Particularly, it could interact differently with S-U or U-S positions in the same way as in the durational domain.

Due to the exclusion of the word list reading task from F0 measurements, there were not enough U-S final cases in CzE beginners, therefore only CzE intermediates and BrE natives underwent the analysis. Both groups show a similar distinction in their treatment of S-U F0 difference in final words, which coincides with phrase-final declination. In the U-S pattern, the difference is reversed in both groups, i.e., unstressed syllable is higher that the stressed one. In the S-U sequence it is vice versa, the declination and the F0 difference caused by the stress pattern reinforce each other, resulting in higher F0 differences in both groups (Tukey post-hoc: $p \leq 0.01$).

3.4. Spectral slope

The results concerning spectral slope have to be interpreted with caution. As far as we know, all metrics of short-term spectral slope are sensitive to vowel identity (see e.g., [34]), so

the difference in spectral slope between two adjacent vowels can occur due to their different qualities and not due to prominence or vocal effort. Nevertheless, with all vowels pooled, we still found a highly significant dissimilarity between the speaker groups – CzE speakers of both proficiency levels showed α difference around zero, while BrE ranged around 3 dB (one-way ANOVA: $F(2, 1910) = 27.4$; $p < 0.001$). This means that the stressed syllables of the native speakers demonstrate in general a flatter spectral slope than the neighbouring unstressed syllables.

This tendency was confirmed by inspecting vowels with identical qualities – the pairs /i/-/ə/, /e/-/ə/, /a/-/ə/ and /u/-/ə/ (other pairs were not represented sufficiently). The pair /e/-/ə/ with the highest occurrence in our data is shown in Figure 5.

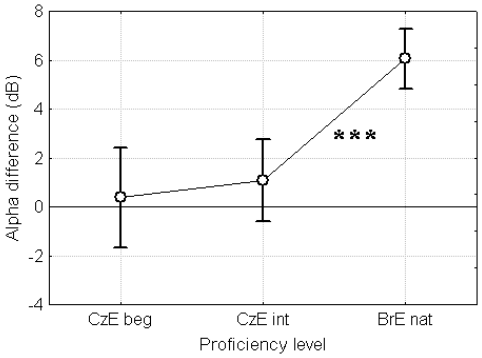


Figure 4: *Average α-measure difference between adjacent stressed /e/ and unstressed /ə/.*

If the vowel quality were the decisive factor, then we would expect no differences across individual groups. However, in this case, the BrE speaker group differs significantly (one-way ANOVA: $F(2, 249) = 16.8$, $p < 0.001$). The same applies to the /u/-/ə/ couple (one-way ANOVA: $F(2, 106) = 7.2$, $p = 0.001$). In the case of /i/-/ə/, the CzE intermediates coincide with the BrE natives, whereas the CzE beginners differ significantly (one-way ANOVA: $F(2, 227) = 8.2$, $p < 0.001$). With /a/-/ə/ the same trend emerges but it was not found significant.

4. Discussion

Addressing our research questions, British native speakers display different behaviour from Czech respondents with regard to all investigated parameters: duration, SPL, F0 and spectral slope of two adjacent vowels. The acoustic characteristics of the BrE native group correspond to the prototypical prominence scheme documented in literature: i.e., the stressed vowels exhibited longer duration, higher SPL, higher F0 and flatter spectral slope than the unstressed ones. On the other hand, Czech speakers fail to create prominence contrasts as systematically and comprehensively as their native counterparts.

Nevertheless, certain variability was detected within the Czech sample. The beginners differ substantially from the intermediates in duration and to a lesser degree in SPL and spectral slope. The character of these findings corroborates the initial hypothesis: more advanced speakers employ the examined features in the direction of native-like realizations.

In the temporal area the results accord with the previous studies exploring Czech English [25], [26], [30]. Moreover, the duration ratio turned out to be the most reliable predictor of the level of L2 phonological acquisition. The ability of Czech speakers to alter the duration of stressed and unstressed vowels seems to improve proportionately to the amount of their linguistic competence in L2. Interestingly, Trofimovich and Baker's data confirmed a similar tendency in English [35], despite differences in their respondents' language background and learning setting (Korean immigrants).

With regard to the SPL, an interesting contrast occurred across the group of CzE speakers. In the phrase-final words the beginners produced the first syllable with a higher SPL consistently regardless of its canonical stress status. This led to a native-like difference in the S-U items and a reversed pattern in the U-S sequences. Since higher SPL does not function as a consistent marker of Czech stress, we may also attribute this outcome to the influence of prosody in the reading task rather than to the mother tongue transfer.

F0 differences showed a notable distinction within the native speaker group. The U-S pattern was treated differently from the S-U, the former exhibiting no change between the two adjacent vowels. This phenomenon could be explained by anticipation: the speakers raise their F0 before the target vowel, so that it manifests itself on the preceding one, too. It has also been shown by [36] that if two subsequent syllables have the same pitch, the second one is perceived as more prominent, because of the expected F0 declination pattern. Examinations of the F0 differences in phrase-final words revealed that phrasal prosody plays a greater role than the F0 stress patterning and tends to override it in both native and CzE intermediate speakers. It would be interesting to see if the F0 measurements in the middle third of the vowels conceal any finer F0 dynamics occurring throughout the whole vowel.

The spectral slope results also suggest that CzE speakers treat the stressed-unstressed contrast differently from the native speakers. The measured contrast is smaller, which implies that Czech speakers articulate both vowels with a similar vocal effort and do not reduce the unstressed elements sufficiently. This finding is in accordance with [27] which found that Czech speakers articulate schwa with a flatter spectral slope than natives.

In conclusion, our results have shed some light on the nature of phonological acquisition of stress by Czech speakers of English. The acquisition process could be viewed as gradual clustering of fragmented acoustic parameters for signalling prominence. While Czech speakers' production illustrates low interconnectedness of the discussed features due to L1 interference, native speakers exploit them in a more cohesive way. Furthermore, SPL, spectral slope and F0 seem to be more compliant with the L1 constraints than duration: the temporal contrast proved to be in stronger relation to learners' L2 experience. How exactly and in which order the remaining aspects develop and interact to create natural prominence patterns poses a question for the future research. In addition, fuller understanding of Czech prominence scheme may help determine the extent of mother tongue transfer.

5. Acknowledgements

The support of the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation is acknowledged. The first author was supported by the project "Acoustic correlates of word stress in Czech, English and Czech English" awarded by the Faculty of Arts, Charles University in Prague in the framework of Specific Academic Research Projects 2014.

6. References

[1] Munro, M. and Derwing, T., "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners", Language Learning, 45: 73–97, 1995.

[2] Lippi-Green, R., "English with an accent", Oxford: Routledge, 2012.

[3] Munro, M. J., "A primer on accent discrimination in the Canadian context", TESL Canada Journal, 20 (2): 38–51, 2003.

[4] Anderson-Hsieh, J., Johnson, R. and Koehler, K., "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure", Language Learning: 42 (4): 529–555, 1992.

[5] Derwing, T. M., Munro, M. J., and Wiebe, G., "Evidence in favour of a broad framework for pronunciation instructions", Language Learning, 48: 393–410, 1998.

[6] Hahn, L. D., "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals", TESOL Quarterly, 38: 201–223, 2004.

[7] Magen, H. S., "The Perception of Foreign-Accented Speech", Journal of Phonetics, 26: 381–400, 1998.

[8] Gut, U., Trouvain, J. and Barry, W. J., "Non-Native Prosody. Phonetic Description and Teaching Practice", Berlin: Mouton de Gruyter, 2007.

[9] Giegrich, H. J., "English Phonology: An Introduction", Cambridge: Cambridge University Press, 1992.

[10] Cruttenden, A., "Gimson's Pronunciation of English", London: Arnold, 2001.

[11] Laver, J., "Principles of Phonetics", Cambridge: Cambridge University Press, 1994.

[12] Fry, D., "Duration and intensity as physical correlates of linguistic stress", Journal of the Acoustical Society of America 27: 765–768, 1955.

[13] Lehiste, I. and Peterson, G. E., "Vowel amplitude and phonemic stress in American English", Journal of the Acoustical Society of America, 31: 428–435, 1959.

[14] Nakatani, L. H. and Aston, C. H., "Perceiving stress patterns of words in sentences", Journal of the Acoustical Society of America, 63, S55, 1978.

[15] Sluijter, A., van Heuven, V. and Pacilly, J., "Spectral balance as a cue in the perception of linguistic stress", Journal of the Acoustical Society of America, 101 (1): 503–513, 1997.

[16] Campbell, N. and Beckman, M., "Stress, prominence, and spectral tilt", ESCA Workshop on Intonation: Theory, Models and Applications, 67–70, Athens: University of Athens, 1997.

[17] Heldner, M., "Spectral emphasis as a perceptual cue to prominence", TMH-QPSR, 42: 51–57, 2001.

[18] Kochanski, G., Grabe, E., Coleman, J. and Rosner, B., "Loudness predicts prominence: fundamental frequency lends little", Journal of the Acoustical Society of America, 118 (2): 1038–1054, 2005.

[19] Ortega-Llebaria, M. and Prieto, P., "Acoustic Correlates of Stress in Central Catalan and Castilian Spanish", Language and Speech, 54 (1): 73–97, 2010.

[20] Beckman, M. E., "Stress and Non-Stress Accent", Dordrecht: Foris, 1986.

[21] Janota, P. and Palková, Z., "Auditory evaluation of stress under the influence of context", AUC Philologica 2/1974, Phonetica Pragensia, 4: 29–59, 1974.

[22] Volín, J., "Z intonace čtených zpravodajství: výška první slabiky v taktu", Čeština doma a ve světě 3–4: 89–96, 2008.

[23] Poesová, K., "Testing the perception of schwa", in I. Kolinská [Ed], Challenges in English Language Teaching III, 72–77, Univerzita J. E. Purkyně – Pedagogická fakulta, 2009.

[24] Skarnitzl, R., "English Word Stress in the Perception of Czech Listeners", in J. Čermák, A. Klégr, M. Malá and P. Šaldová [Eds], Patterns, A Festschrift for Libuše Dušková, 183–194, Praha: Kruh moderních filologů, 2005.

[25] Volín, J. and Poesová, K., "Temporal and spectral reduction of vowels in English weak syllables", in A. Grmelová, L. Dušková, M. Farrell and R. Pípalová [Eds], Plurality and Diversity in English Studies, 18–27, Praha: UK PedF, 2008.

[26] Volín, J., "Rhythmical properties of polysyllabic words in British and Czech English", in J. Čermák, A. Klégr, M. Malá and P. Šaldová [Eds], Patterns, A Festschrift for Libuše Dušková, 183–194, Praha: Kruh moderních filologů, 2005.

[27] Volín, J., Weingartová, L. and Skarnitzl, R., "Spectral Characteristics of Schwa in Czech Accented English", Research in Language, 11 (1): 31–39, 2013.

[28] Volín, J. and Skarnitzl, R., "The strength of foreign accent in Czech English under adverse listening conditions", Speech Communication, 52: 1010–1021, 2010.

[29] Skarnitzl, R., "Prague Phonetic Corpus: status report", AUC Philologica 1/2009, Phonetica Pragensia, XII: 65–67, 2010.

[30] Poesová, K., "Vliv systematického používání vybraných metod výuky výslovnosti anglického jazyka na percepci a produkci hlásky schwa u žáků ZŠ", unpublished PhD thesis. Praha: UK PedF, 2012.

[31] Volín, J. and Weingartová, L., "Acoustic correlates of word stress as a cue to accent strength", in preparation.

[32] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [computer program], version 5.3.41. Retrieved from: http://www.praat.org/.

[33] Wells, J. C., "Longman pronunciation dictionary (3rd ed.)", Harlow: Pearson Education Limited, xxxvii + 922, 2008.

[34] Weingartová, L. and Volín, J., "Spectral Measurements of Vowels for Speaker Identification in Czech", Studies in Applied Linguistics 1: 21–36, 2013.

[35] Trofimovich, P. and Baker, W., "Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech", Studies in Second Language Acquisition, 28: 1–30, 2006.

[36] Terken, J. M. B., "Fundamental frequency and perceived prominence of accented syllables II: Non-final accents", Journal of the Acoustical Society of America, 95 (6): 3662–3665, 1994.

A sketch of an extrinsic timing model of speech production

Alice Turk [1], *Stefanie Shattuck-Hufnagel*[2]

[1] Department of Linguistics & English Language, University of Edinburgh, Edinburgh, Scotland
[2] Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA

turk@ling.ed.ac.uk, sshuf@mit.edu

Abstract

In this paper, we motivate and present a sketch of an extrinsic timing model of speech production. It is a three-stage model, involving 1) phonological planning, where symbolic segmental representations are sequenced and slotted into an appropriate prosodic structure, and where appropriate acoustic cues are selected for each segment in its context, and 2) phonetic planning, where cues are mapped onto sets of articulators, and appropriate values for spatial and temporal parameters of movement are computed, and 3) phonetic implementation, where articulator movements are generated, monitored, and updated. We cite model components from the literature that accomplish many of the required functions.

Index Terms: speech production, extrinsic timing, prosodic structure

1. Introduction

Articulatory Phonology/Task Dynamics (hereafter AP/TD [1,2]) is the model of speech production that currently provides the most comprehensive account of speech timing phenomena. Timing control in this model is intrinsic, that is, surface timing patterns emerge from properties of the system and do not need to be represented, specified, or tracked during an utterance using a system-extrinsic timekeeper. However, several lines of behavioral evidence challenge intrinsic timing as implemented in AP/TD, and support the view that timing control in speech production is extrinsic. In this paper, we first present three types of evidence that support extrinsic timing in speech production, and then discuss a preliminary sketch of an alternative model of speech production that involves symbolic phonological representations and extrinsic timing. We point out model components from the literature that can be used to implement the model.

2. Evidence for extrinsic timing

2.1. Increasing variability with increases in interval duration, as predicted by a "noisy timekeeper" model

Many studies show more variability in interval duration for longer intervals in a variety of motor tasks [3]; for speech production, see e.g. [4]. As explained in [4, p. 422], these findings are expected in extrinsic timing models: "the mechanism that *meters out* intervals of time … is variable, and the amount of variability is directly proportional to the length of the interval of time to be metered out." (This is because time is metered out in smaller units than the total interval, and the variability in each inter-tick interval adds up). The relationship of variability to mean duration follows Weber's law, with an approximately constant coefficient of variation (standard deviation/mean) for a range of intervals in both humans and animals, consistent with an extrinsic timing mechanism [5,6].

2.2. Surface timing constraints and goal specifications suggest extrinsic timing

Within AP/TD, desired surface durations aren't specified as part of the utterance plan, but instead emerge from interacting components within the task dynamical system. For example, gesture durations in phrase-final position reflect the settling-time of their mass-spring system, their gestural activation interval, and an adjustment which lengthens the gestural activation intervals at the boundary [7,8]. In AP/TD, the surface duration emerges from these mechanisms and is not explicitly specified in the original utterance plan. However, [9] suggest that a constraint on surface durations of phonemically short vowels in phrase-final position may be required to preserve the short vs. long phonemic contrast in Northern Finnish. The authors observed that the magnitude of final, accentual, and combined lengthening on phonemically short vowels in word-final syllables was restricted compared to lengthening on phonemically long vowels (17% combined accentual + final lengthening on phonemically short vowels in a word-final syllables vs. 68% on long vowels in the same context). These results are consistent with the view that the surface durations of the phonemically short vowel are restricted in order to avoid endangering the phonemic short vs. long vowel quantity contrast in this language. Although it is possible to *implement* this type of effect in AP/TD, the effect is difficult to *explain* within the theory, since surface durations can't be referred to. Additional support for the representation of surface durations can be found in studies of speech rate effects and durational correlates of prosodic structure and quantity [10-12]; despite considerable variability in the strategies that different speakers use to implement these factors, speakers all achieve a common surface duration pattern of relatively long surface durations e.g. in phrase-final position, at slow speech rates, and for phonemically long vowels. These findings challenge intrinsic timing in AP/TD because they suggest the equivalence of different strategies that result in similar surface duration patterns, and therefore support the specification of surface duration goals.

2.3. Separate control of movement targets vs. onsets challenges intrinsic mass-spring models

In [13], Dave Lee commented "it is frequently not critical when a movement starts—just so long as it does not start too late. For example, an experienced driver who knows the car

and road conditions can start braking safely for an obstacle a bit later than an inexperienced driver…" This type of example suggests that timing variability may be different at target attainment vs. movement onset, difficult to account for in mass-spring models such as AP/TD, but relatively straightforward to account for in extrinsic timing models that allow separate timing specification and prioritization for target attainment vs. other parts of movement [14].

Several studies have confirmed the differential variability in the timing of target attainment compared to the timing of other movement events such as movement onset ([15-18], for non-speech motor activity; [19] for speech). For example, [19] showed differences in of timing variability for onsets vs. target attainment for upper lip protrusion movements during spoken /i_u/ sequences. While AP/TD does provide a mechanism for separately adjusting the timing of the beginning and the end of an activation interval (by applying its prosodic "stretching" mechanism to a proportion of the interval), it doesn't provide a mechanism by which these timings could be differently variable. These findings suggest that target attainment timing is controlled independently of movement onset timing, and that target attainment timing takes higher priority. Similar findings of differential variability at target attainment vs. elsewhere in movement have been observed for spatial characteristics of repeated non-speech movements, where spatial variability is lowest at a movement target and higher elsewhere, e.g. [20]. These findings add further support for the separate control of targets vs. other parts of movement.

3. Key features

The key feature of our proposed model sketch is that it involves extrinsic timing, with a way to assign different priorities to the timing of movement targets vs. other parts of movement such as onsets. Extrinsic timing implies a-temporal representations, and we therefore assume that representations are symbolic, because symbolic representations are a type of a-temporal representation. We favor symbolic representations because they offer a better account of phonological equivalence than alternative a-temporal representations such as spatial paths without timing. Mechanisms of phonetic implementation are required to map these symbolic representations onto their surface phonetic form. We therefore assume a three-stage model, involving 1) phonological planning, and 2) phonetic planning, and 3) phonetic implementation. We assume that timing specification is a part of phonetic planning that is separate from the specification of spectral/spatial information (see [21] for a similar view). Timing information is combined with spatial information to generate movements intended to get to their targets on time. We discuss the planning and implementation stages in more detail below.

4. Phonological planning

We assume that phonological planning involves sequencing symbolic segmental representations and slotting them into a prosodic frame that includes hierarchical constituent and prominence structure [22]. Following [23,24], we hypothesize that prosodic structure is planned with the goal of an even distribution of recognition likelihood by the listener throughout an utterance (called smooth signal redundancy). To this end, predictability information (from language and real-world context) is used to plan prosodic structure so that relatively unpredictable elements are highlighted, either by manipulating relative prosodic prominence, or by manipulating relative prosodic boundary strength (highlighting through edge demarcation). In the planning stage, other task requirements are identified, such as speaking quickly, or in a particular style (e.g. clear speech, periodic speech, etc.), as illustrated in Figure 1. These requirements are assigned relative priorities so that, in the Phonetic Implementation stage, they can be balanced against movement costs to yield optimal movements (see below and Figure 1).

Several aspects of Figure 1 are worthy of comment. First, the effects of predictability on planned phonetic form are assumed to be indirect, where predictability affects plnaned prosodic structure, and prosodic structure in turn affects planned surface phonetics. This view represents our current hypothesis, but we note that it is possible that predictability might have additional direct effects on phonetic form (in addition to those that are mediated by prosodic form). Second, we assume that the effects of non-grammatical factors, like rate and style of speech, on phonetic form have a direct effect on planned surface phonetics. Although these factors have been observed to affect aspects of prosody (e.g. fewer "breaks" at faster rates of speech, cf. [25]), our view is that a speaker would plan the same prosodic structure (i.e. same relative prominence and relative boundary strength structure) for a given utterance at different rates of speech, but that the planned correlates of this structure would be different at different rates because the rate of speech requirement would be balanced against the prosodic structure requirement in determining optimum phonetic characteristics that meet the competing demands. Third, the list of factors mentioned in the "Non-grammatical factors" box is intended to be a preliminary indicator of the many non-grammatical factors that might be at work, and may not be exhaustive.

At the planning stage, each symbolic representation in its context (prosodic, stylistic, etc.) is associated with a set of acoustic cues [26]. For example, in syllable initial position English /t/ might be associated with silence, then a relatively high frequency release burst + aspiration noise, but with a different set of cues in syllable final and ambisyllabic position.

5. Phonetic planning

Phonetic planning involves a) mapping cues onto sets of articulators, and b) assigning values to a set of movement parameters, including timing and spatial parameters (and perhaps accuracy goals).

Selected acoustic cues are mapped onto quantitative acoustic/constriction goals, which are achieved by sets of articulators, or synergies (see [27] for a plausible neural network model of the mapping between acoustics and articulatory goals). We assume that the constriction goals are very similar to the set defined by Saltzman & Munhall [2], and adopted by [27].

We assume that movement parameters include spatial aspects of the constriction target, e.g. lip aperture, tongue tip constriction with the alveolar ridge, etc., a default relative contribution of each articulator in a synergy, as well as a set of timing parameters.

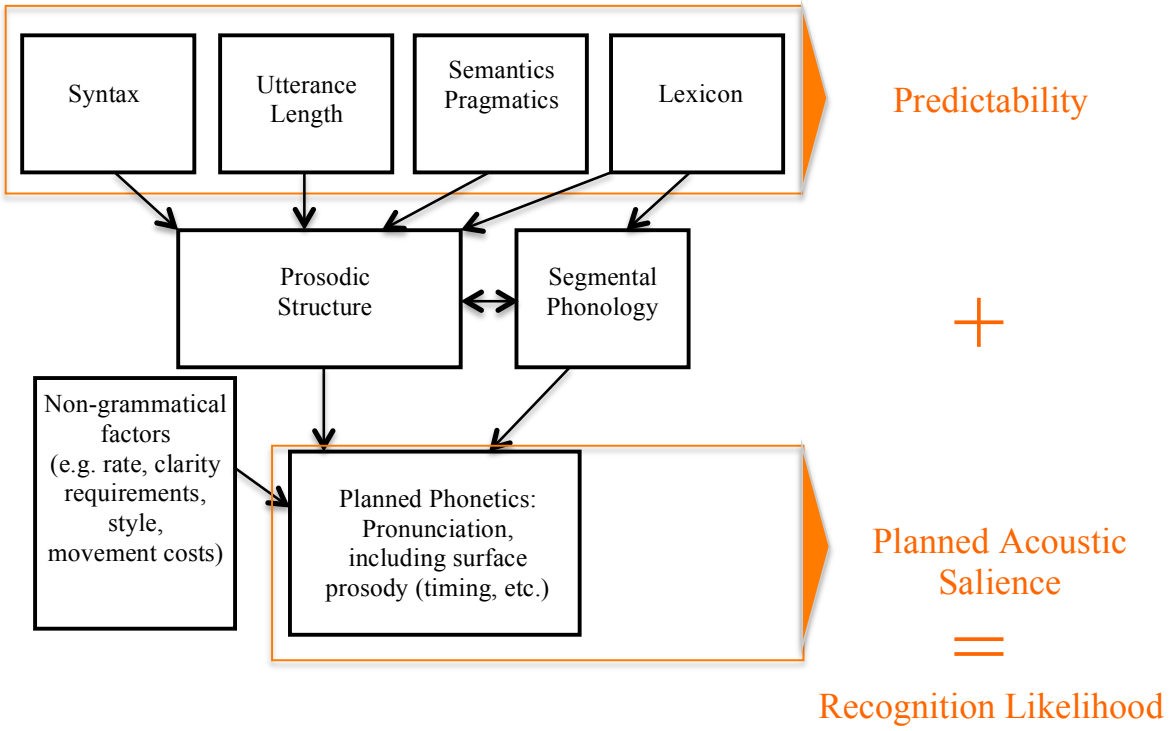


Figure 1: *Factors that shape planned surface phonetics and their relationship **to** predictability, acoustic salience, and recognition likelihood. Based on similar figures in [23,24]*

The timing parameters include:

1. Interval durations of various types, e.g. phrase-final rhyme durations for final lengthening, phrase-initial segment durations for initial lengthening, stressed syllable (or CV) durations for prominence-related lengthening; see [28] for more detail.
2. The timing of acoustic landmarks/constriction targets relative to the preceding one in a sequence.
3. The timing of movement onsets relative to the landmarks/targets
4. The time course of movement (tau, time-to-target achievement at the current movement rate, as a function of time [13]). When combined with spatial information, the tau function determines the velocity profile of movement. Following Lee [13], we assume that tau follows an *intrinsic tau guide*, represented by an equation that describes a family of finite movements (movements from rest that start with an acceleratory component and end after a finite duration): $tauG = k[½(t-T^2/t)]$. The parameters of the equation are T, the duration of movement, and k, which describes the shape of the movement. The variable t is the elapsed time from the start of the movement. Tau-guided movements will have a single-peaked velocity profile if $k < 1$.

5.1. Determining parameter values

It is well-known that surface phonetic characteristics, including timing, vary systematically with prosodic and segmental context, as well as non-grammatical factors such as clarity requirements, rate, and style. Movement timing also varies with movement distance and accuracy requirements (Fitts' law [29]), where longer distance movements take longer in spite of increased movement speeds, and movements with higher spatial accuracy requirements take longer than movements with lower spatial accuracy requirements. All of these factors need to be taken into account in computing timing values. We assume that phonetic characteristics are also constrained by processing demands and movement costs, such as energy, time, and the cost of inaccuracy (see Figure 1). Following Optimal Control Theory [30-32], we propose that movement parameters are determined that represent the optimal balance between prioritized (or weighted) task requirements and movement costs (see [33] for an OCT interpretation of Fitts' law phenomena, and [34] for an example of the use of OCT in a model of speech production).

We acknowledge that computing all of the parameter values for movement is non-trivial, one reason being that parameter values are inter-dependent. For example, the timing between targets in a sequence will depend on timing requirements for supra-segmental intervals such as phrase-final syllable rhymes), and, as we mentioned earlier, movement timing parameter values depend on spatial parameters such as movement distance and accuracy. In the examples which follow, we illustrate the factors involved in determining the values for three of the types of timing parameter specifications defined above.

Example 1: For the timing of at least some intervals that are the sites of durational effects of prosodic context, we hypothesize that requirements for these intervals will have an influence on the timing between movement targets. For example, the timing between targets within the word-final

syllable rhymes will be longer if they are phrase-final than if they are phrase-medial. Because the phrase-final durational requirement is balanced against the cost of time, we expect phrase-final lengthening to be minimized where it can be. Evidence consistent with this view can be found in [37], who observed that it is not the case that e.g. every phrase-final segment has the same duration, rather, all segments of a particular type are longer in phrase-final position than medially, but the amount of absolute lengthening is segment-specific. Nevertheless, the relative durational rank ordering among segments is preserved.

Example 2: For the timing of an acoustic landmark/movement target with respect to a previous landmark, we assume that there is a cost for time that penalizes time between speech landmarks. We hypothesize that the time between targets will additionally depend on prosodic context and other factors, such as rate and style of speech. For example, if the speech rate is slow, the duration between targets will be longer than if the overall speech rate is fast. In cases where the two targets (X,Y) in a sequence involve the same articulators, the duration between targets will additionally depend on the distance between them, and on the target's spatial accuracy requirements, as well as on the energy cost for reaching the second target.

Example 3: For the timing of movement onset with respect to movement target achievement, we assume that costs for time and energy will constrain overall movement time, and that movement time will increase with the spatial accuracy requirement of the target, and will decrease with its timing accuracy requirement, because faster movements are more accurate in terms of their timing [35]. We hypothesize that the relative weighting of these two requirements might vary with speaker and style. Other factors may also affect movement time, such as prosodic position, where some syllable-final movements may be longer than syllable-initial movements (e.g. velum lowering for nasal stops, [36]).

5.2. Coordination for synchronized targets

Lee [13] presents a way of planning movement coordination within an extrinsic timing framework (General Tau theory in this case). On this theory, movements are coordinated through tau coupling, whereby movements whose tau functions are in constant proportion will end at the same time. Coordination can be achieved by coupling one or more movements onto the internal tau guide (mentioned above), or by coupling a movement onto a sensed movement tau. As explained in [40], when two movement tau functions are in constant proportion, e.g. tauA = ktauB, tauA reaches zero as the target is reached, and because tauB is in constant proportion to tauA, it reaches zero at the same time. On this theory, movement coordination involves movement *offset* coordination. It does not require the time course of the movements to be the same, nor is there a strict requirement for the movements to begin at the same time. What this means is that two coordinated movements might have velocity peaks that don't occur at the same time, but as long as their taus are in constant proportion, they will reach their targets at the same time. In addition, if one of two coordinated movements starts later than the other, it is assumed that the later onset movement is accelerated until $tau_{later} = ktau_{earlier}$, and then that the relation is maintained so that the two movements end at the same time [41].

6. Phonetic implementation

Phonetic implementation involves producing movement kinematics to meet planned phonetic goals. Following the VITE (Velocities-into-terminal-endpoints) model of [38], cited in early versions of DIVA [27], we assume that speakers constantly monitor positions relative to a) the planned target (either actual or predicted, depending on the type of predicted and/or sensory information available), as well as b) the time until planned target achievement at each time point (the tau function, [13]). This information is combined to generate appropriate movement velocities to get the articulators to the target on time. Based on evidence in e.g. [39], and following proposals in Optimal Feedback Control Theory [31-32], we further assume that movement corrections and updates can be made during a movement on the basis of sensory feedback and predicted states, and that corrections will be more likely for prioritized parts of movement, e.g. movement offsets, compared to other parts of movement.

7. Coarticulation in sequential movements

One of the key contributions of AP/TD is its account of coarticulation. In our model, coarticulation falls out of the relative timing of movement targets, and of the movement times required to produce the targets on time with required spatial accuracy.

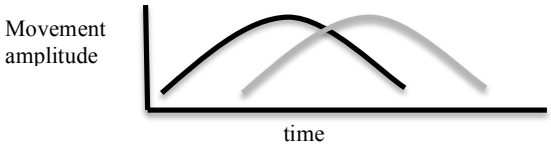


Figure 2. Schematic diagram of coarticulation, where the target of movement A (black) is timed to occur before the target of movement B (grey)

If movement targets are timed closely together, and are produced with different sets of articulators, then for a sequence of two targets AB, the movement onset of B will begin before the movement target A is reached. This is illustrated in Figure 2.

8. Conclusion

The main advantage of our proposal is that it is likely to provide a better fit to existing data in the literature than the AP/TD model, currently the best-worked-out model of speech production. In our experience, careful consideration of a well-motivated alternative to a dominant model can often result in improvements in both competing models. The major drawback to our proposal is that it is still only a model *sketch*. We have not implemented it, and as we note above, implementation will be non-trivial. Attempts to implement it will no doubt bring many deficiencies and oversights to light. However, we hope it will provide a framework for asking fruitful questions about how to model timing in speech production, and for interpreting timing data.

9. References

[1] C. P. Browman and L. Goldstein, "Dynamic modeling of phonetic structure," in *Phonetic Linguistics*, V. A. Fromkin, Ed., ed New York: Academic Press, 1985, pp. 35-53.

[2] E. L. Saltzman and K. Munhall, "A dynamical approach to gestural patterning in speech production," *Ecological Psychology,* vol. 1, pp. 333-382, 1989.

[3] R. B. Ivry and R. E. Hazeltine, "Perception and production of temporal intervals across a range of durations - Evidence for a common timing mechanism," *Journal of Experimental Psychology-Human Perception and Performance,* vol. 21, pp. 3-18, Feb 1995.

[4] D. Byrd and E. Saltzman, "Intragestural dynamics of multiple prosodic boundaries," *Journal of Phonetics,* vol. 26, pp. 173-199, Apr 1998.

[5] M. Treisman, "Temporal Discrimination and the Indifference Interval - Implications for a Model of the Internal Clock," *Psychological Monographs,* vol. 77, pp. 1-31, 1963.

[6] J. Gibbon, "Scalar Expectancy Theory and Weber's law in animal timing," *Psychological Review,* vol. 84, pp. 279-325, 1977.

[7] D. Byrd and E. Saltzman, "The elastic phrase: modeling the dynamics of boundary-adjacent lengthening," *Journal of Phonetics,* vol. 31, pp. 149-180, Apr 2003.

[8] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Speech Prosody 2008*, Campinas, Brazil., 2008.

[9] S. Nakai, A. Turk, K. Suomi, S. Granlund, R. Ylitalo, and S. Kunnari, "Quantity constraints on the temporal implementation of phrasal prosody in Northern Finnish," *Journal of Phonetics,* vol. 40, pp. 796–807, 2012.

[10] J. Berry, "Speaking rate effects on normal aspects of articulation: Outcomes and issues," *Perspectives on Speech Science and Orofacial Disorders,* vol. 21, pp. 15-26, 2011.

[11] J. Edwards, M. E. Beckman, and J. Fletcher, "The articulatory kinematics of final lengthening," *Journal of the Acoustical Society of America,* vol. 89, pp. 369-382, 1991.

[12] I. Hertrich and H. Ackermann, "Articulatory control of phonological vowel length contrasts: Kinematic analysis of labial gestures," *Journal of the Acoustical Society of America,* vol. 102, pp. 523-536, 1997.

[13] D. N. Lee, "Guiding movement by coupling taus," *Ecological Psychology,* vol. 10, pp. 221-250, 1998.

[14] L. H. Shaffer, "Rhythm and timing in skill," *Psychological Review,* vol. 89, pp. 109-122, 1982.

[15] R. M. C. Spencer and H. N. Zelaznik, "Weber (slope) analyses of timing variability in tapping and drawing tasks," *Journal of Motor Behavior,* vol. 35, pp. 371-381, Dec 2003.

[16] M. Billon, A. Semjen, and G. E. Stelmach, "The timing effects of accent production in periodic finger-tapping sequences," *Journal of Motor Behavior,* vol. 28, pp. 198-210, Sep 1996.

[17] R. Bootsma and P. C. van Wieringen, "Timing an attacking forehand drive in table tennis," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 16, pp. 21-29, 1990.

[18] C. Craig, G. J. Pepping, and M. Grealy, "Intercepting beats in predesignated target zones," *Experimental Brain Research,* vol. 165, pp. 490-504, Sep 2005.

[19] J. S. Perkell and M. L. Matthies, "Temporal measures of anticipatory labial coarticulation for the vowel /u/ - within-subject and cross-subject variability," *Journal of the Acoustical Society of America,* vol. 91, pp. 2911-2925, May 1992.

[20] Liu and E. Todorov, "Evidence for the flexible sensorimotor strategies predicted by Optimal Feedback Control," *The Journal of Neuroscience,* vol. 27, pp. 9354 –9368, 2007.

[21] A. Georgopoulos, "Cognitive motor control: spatial and temporal aspects," *Current Opinion in Neurobiology,* vol. 12, pp. 678–683, 2002.

[22] P. Keating and S. Shattuck-Hufnagel, "A prosodic view of word form encoding for speech production," *UCLA Working Papers in Phonetics,* vol. 101, pp. 112-156, 2002.

[23] M. Aylett and A. Turk, "The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration on spontaneous speech," *Language and Speech,* vol. 47, pp. 31-56, 2004.

[24] A. Turk, "Does prosodic constituency signal relative predictability? A Smooth Signal Redundancy hypothesis," *Journal of Laboratory Phonology,* vol. 1, pp. 227-262, 2010.

[25] J. Caspers, Pitch movements under time pressure: Effects of speech rate on the melodic marking of accents and boundaries in Dutch. The Hague: Holland Academic Graphics 1994.

[26] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *Journal of the Acoustical Society of America,* vol. 111, 2002.

[27] F. H. Guenther, "Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural-Network Model of Speech Production," *Psychological Review,* vol. 102, pp. 594-621, Jul 1995.

[28] A. Turk, "The temporal implementation of prosodic structure," in *The Oxford Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, and M. K. Huffman, Eds., ed: Oxford University Press, 2012, pp. 242-253.

[29] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *Journal of Experimental Psychology,* vol. 47, pp. 381-391, 1954.

[30] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[31] E. Todorov and M. I. Jordan, "Optimal feedback control as a theory of motor coordination," *Nature Neuroscience,* vol. 5, pp. 1226-1235, Nov 2002.

[32] R. Shadmehr and S. Mussa-Ivaldi, *Biological learning and control: How the brain builds representations, predicts events, and makes decisions*. Cambridge, MA: The MIT Press, 2012.

[33] C. M. Harris and D. M. Wolpert, "Signal-dependent noise determines motor planning," *Nature,* vol. 394, pp. 780-784, Aug 20 1998.

[34] J. Simko and F. Cummins, "Sequencing and optimization within an embodied Task Dynamic model," *Cognitive Science,* vol. 35, pp. 527–562, 2011.

[35] P. A. Hancock and K. M. Newell, "The movement speed-accuracy relationship in space-time," in *Motor Behavior: Programming, Control, and Acquisition*, H. Heuer, U. Kleinbeck, and K.-H. Schmidt, Eds., Berlin: Springer-Verlag, 1985, pp. 153-185.

[36] R. A. Krakow, "Physiological organization of syllables: a review," *Journal of Phonetics,* vol. 27, pp. 23-54, 1999.

[37] J. van Santen and C. L. Shih, "Suprasegmental and segmental timing models in Mandarin Chinese and American English," *Journal of the Acoustical Society of America,* vol. 107, pp. 1012-1026, 2000.

[38] D. Bullock and S. Grossberg, "Neural Dynamics of Planned Arm Movements - Emergent Invariants and Speed Accuracy Properties during Trajectory Formation," *Psychological Review,* vol. 95, pp. 49-90, Jan 1988.

[39] D. Pélisson, C. Prablanc, M. A. Goodale, and M. Jeannerod, "Visual control of reaching movements without vision of the limb II. Evidence of fast unconscious processes correcting the trajectory of the hand to the final position of a double-step stimulus," *Experimental Brain Research,* vol. 62, pp. 303-311, 1986.

[40] D. N. Lee, A. P. Georgopoulos, M. J. O. Clark, C. M. Craig, and N. L. Port, "Guiding contact by coupling the taus of gaps," *Experimental Brain Research,* vol. 139, pp. 151–159, 2001.

[41] D. N. Lee, personal communication.

SLAM: Automatic Stylization and Labelling of Speech Melody

Nicolas Obin[1], *Julie Beliao*[2], *Christophe Veaux*[3], *Anne Lacheret*[2]

[1] IRCAM - UMR STMS IRCAM-CNRS-UPMC, Paris, France
[2] MoDyCo - UMR 7114, Université de Paris Ouest, Nanterre, France
[3] Centre for Speech Technology Research, Edinburgh, UK

Abstract

This paper presents SLAM : a simple method for the automatic Stylization and LAbelling of speech Melody. The main contributions over existing methods are : the alphabet of melodic contours is fully data-driven, an explicit time-frequency representation is used to derive complex melodic contours, and melodic contours can be determined over arbitrary prosodic/syntactic units. Additionally, the system can handle some specificities of spontaneous speech (e.g., multi speakers, speech turns and speech overlaps). A preliminary experiment conducted on 3 hours of spoken French indicates that a small number of contours is sufficient to explain most of the observed contours. The method can be easily adapted to other stressed languages. The implementation is open-source and freely available †.
Index Terms : intonation, stylization, automatic labelling, prosody, syntax.

1. Introduction

The transcription of speech prosody aims at representing the variations of speech prosody that are considered as relevant with an alphabet of elementary symbols [1, 2, 3, 4], which each instantiates a function in the speech communication process. The inventory of this alphabet is desired to facilitate further studies on the function of speech prosody in speech communication, from prosody/syntax to prosody/discourse and dialogue interfaces, from formal to spontaneous speech.

The first representation of French intonation in terms of global contours [5] focused on modal intonation : a global melodic contour specifies the modality of a sentence (e.g., interrogation, exclamation). More recently, this paradigm was extended to the representation of intonation as the superposition of melodic contours over various syntactic units [6]. The main contribution of this paper tends to the generalization of this paradigm to any linguistic unit – prosodic and syntactic. In other words, we assume that a specific dictionary of elementary contours can be derived for each linguistic unit.

This paper presents a novel method for the automatic labelling of melodic contours over arbitrary prosodic/syntactic units. The main contribution of the method compared to existing methods [2, 3, 7, 8, 9] can be summarized as follows :

- The melodic system (i.e., the alphabet of melodic contours) is fully data-driven (bottom-up processing).
- An explicit time-frequency representation is used to describe complex melodic contours.
- The proposed representation handles a large variety of prosodic/syntactic units : from local (e.g., syllable) to global contours (here, prosodic and syntactic).

Additionally, the representation is normalized with respect to the average range of a speaker, and handles some particularities of spontaneous speech (e.g., multi speakers, speech turns, speech overlaps). Lastly, the implementation is open-source and freely available.

The remainder of this paper presents the main principles of the method used for the transcription of melodic contours over arbitrary prosodic/syntactic units. The compact form of the alphabet of contours proves the efficiency of the proposed method : around 10/20 elementary contours suffice to explain 95% of the observed contours.

2. Intonation Labelling

2.1. Speech Preprocessing

The only external requirements for the automatic labelling are : the estimation of the fundamental frequency of speech (F0), and the segmentation of speech into speech units that are desired for the description of speech prosody (arbitrary prosodic/syntactic units). For the F0 estimation, popular methods are freely available (e.g., STRAIGHT [10], YIN [11], and SWIPE [12]). Also, many refinements to the F0 estimation exist to facilitate further processing - from F0 periodicity estimation (and voiced/unvoiced decision - similarly to [8]), to F0 smoothing and interpolation methods. For the speech segmentation, speech-to-text alignment methods exist (IRCAMALIGN [13], EASYALIGN [14], and SPPAS [15] among others - usually based on the open-source HTK library [16]) for the segmentation of speech into phonemes, syllables, words, and phrases. Also, alternative methods exist for language-independent speech segmentation into syllables and phrases [17].

2.2. Acoustic Representation

The acoustic stylization of speech melody consists in representing the F0 variations that are considered as relevant for the description of speech prosody. In general, F0 stylization methods are based on the representation of F0 variations according

†. This study was supported by the French National Research Agency (ANR) for the RHAPSODIE project : reference prosody corpus of spoken French. The resource is implemented in python and freely available on : `https://github.com/jbeliao/SLAM/`. The current release supports PRAAT TextGrid input/output format for segmentation and labelling. There is no need for preliminary F0 estimation, which is processed automatically with the python implementation of the SWIPE algorithm (see `https://github.com/kylebgorman/swipe/` for details).

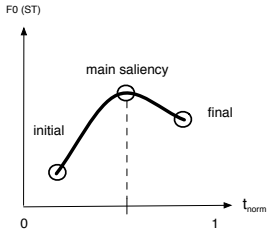


FIGURE 1 – *Acoustic representation of a contour.*

to a small set of parameters that are used to describe slowly time-varying F0 variations [18, 19, 20, 21]. In the proposed method, the F0 contour is represented by a set of 5 acoustic values for each given unit :

1. INITIAL : the initial value of the F0 on the unit. This value corresponds to the first F0 value for which the acoustic frame is considered as voiced ;
2. FINAL : the final value of the F0 on the unit. This value corresponds to the last F0 value for which the acoustic frame is considered as voiced ;
3. MAIN SALIENCY : the value corresponding to the most salient F0 peak – if one exists. The F0 variations over a unit can be decomposed into a main saliency (optional, if present) and a set of secondary salience (optional, if present). The method assumes that only the main saliency contributes to the definition of a global contour, while secondary salience contribute to the internal structure of the global contour, and can be neglected in a first-order approximation.

Finally, the following values are added to the description :

4. MAIN SALIENCY POSITION : the time position of the main saliency ;
5. LOCAL REGISTER : the mean F0 over the unit.

All frequency values are expressed in semi-tones (STs), with respect to the overall mean F0 of the speaker :

$$\text{F0[ST]} = 12 \times \log_2 \frac{\text{F0[Hz]}}{\text{F0}_{\text{mean}}\text{[Hz]}} \quad (1)$$

All time positions are expressed relative to the boundaries of the unit :

$$t_{norm} = \frac{t - t_{start}}{t_{end} - t_{start}} \quad (2)$$

This acoustic representation adapts automatically to the nature of the prosodic/syntactic unit. Here, the notion of micro and macro prosodic variations is assumed to be relative to the linguistic unit considered : for short units (and local contours ; e.g., syllable), the phoneme variations will be considered as micro variations compared to the syllable variations ; for large units (and global contours ; e.g., phrases), the syllable variations will be in turn considered as micro variations compared to the larger unit.

2.3. Symbolic Representation

The acoustic representation presented in the previous section serves as a time-frequency representation for the labelling of contours.

2.3.1. Frequency Quantization

First, frequency values are represented with respect to 5 pitch levels covering the whole F0 range of the speaker (table 1). Each pitch level covers a range of 4 semi-tones centred on the average F0 value of the speaker. For instance, the medium range covers from - 2 STs to + 2 STs around the average range of the speaker ; the high range covers from +2 STs to +6 STs ; and the extreme-high range covers all values that exceed +6 STs.

PITCH LEVELS	DESCRIPTION	RANGE (STs)
H	extreme-high	$>$ +6
h	high	+2/+6
m	medium	-2/+2
l	low	-2/-6
L	extreme-low	$<$ -6

TABLE 1 – *Pitch levels used for the symbolic representation.*

From this representation the sequence of initial/final/saliency values can be converted into a corresponding sequence of pitch levels. Then, this representation can be used to describe static tones, simple contours, and complex contours. An illustration is provided in figure 2.

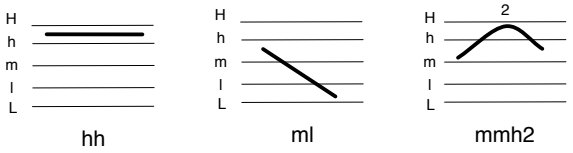


FIGURE 2 – *Types of contours that can be described. From left to right : a static tone (flat contour in the high range of the speaker), a simple contour (a falling contour from the medium to the low range of the speaker), and a complex contour (medium to medium with a saliency observed in the high range of the speaker in the middle part of the unit).*

Additionally, the main saliency is considered as significant only if the corresponding point differs by more than 2 ST from the initial and the final points. If this is not the case, the main saliency is not considered as relevant, and is removed from the symbolic representation of the contour. An illustration is provided in figure 5.

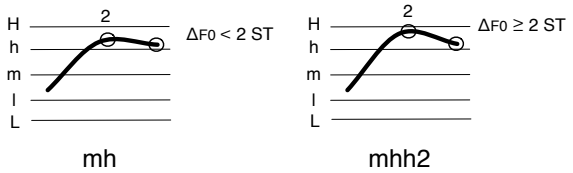


FIGURE 3 – *Determination of the presence of a saliency. On the left, the saliency is not considered as relevant, and the contour is transcribed as "mh" (medium to high) ; on the right, the saliency is relevant, and the contour is transcribed as "mhh2" (medium to high with the presence of a saliency in the second part of the unit).*

2.3.2. Time Quantization

Second, the time position of the main saliency is represented with respect to 3 time positions, which are determined from the relative position of the saliency within the unit and the decomposition of the unit into 3 equal parts 2. An illustration of contours with various positions of the main saliency is provided in figure 5.

TIME POSITION	MAIN SALIENCY
1/3	first part of the unit
2/3	middle part of the unit
3/3	last part of the unit

TABLE 2 – *Time position of the main saliency of a contour used for the symbolic representation.*

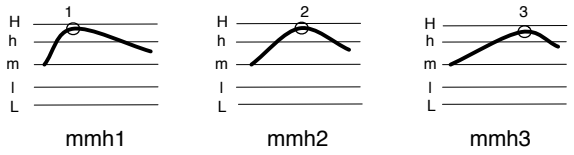


FIGURE 4 – *Contour labelling with regard to the position of the saliency within the unit. From left to right, a saliency is observed in the first/middle/last part of the unit.*

2.3.3. Formal Representation

Finally, the time-frequency representation is formally described as :

Initial Final [Saliency] [Position]

where [.] indicates optional fields dependent on the relevancy of the main saliency. Illustrations of the various contours that can be represented are shown in figure 2.

The alphabet of this formal representation is expressive in the sense that the alphabet can cover a large variety of contours - here, 400. Theoretically, the optimality of a phonological system assumes the smallest number of elements required to account for the largest number of observations. Practically, a compact alphabet is desired to facilitate labelling and linguistic interpretation. For instance, the inventory in the ToBI system (in order to describe phrasal tones and pitch accents) includes two elementary tones (H, L), and around 5 to 8 pitch accents have been proposed for American-English [22]. In order to address the efficiency of the proposed representation, a preliminary experiment will be described in section 3, which proves that a small number of contours (around 10/20) is sufficient to explain most of speech prosody variations in the real-conditions of ordinary speech. An illustration of the acoustic representation and the labelling of contours is provided in figure 5.

3. Experiment

A preliminary experiment was conducted on the RHAPSODIE treebank (prosody/syntax) of spoken French [23], composed of speech recordings of French ordinary speech and orthographic transcription. The transcription and the annotations are all aligned on the speech signal : phonemes, syllables, words, speakers, speech turns, speech overlaps. The RHAPSODIE treebank comprises : 57 speech recordings, 3 hours of ordinary speech, and 33,000 words ; multiple situations : monologue/dialogue, formal/informal ; multiple speakers : male/female.

Firstly, the analysis of contours reveals that a small number of contours suffices to explain most of the observed contours : around 10/20 elementary contours suffice to explain 95% of the observed contours, regardless of the prosodic/syntactic unit (table 3, see [23] for details). This can be interpreted as follows : a small number of elementary contours commonly serves for usual speech communication, and a variety of rare contours may convey specific speaker and/or expressive information (e.g. emotions). This constitutes a first validation concerning the acoustic/symbolic representation : the representation is efficient (a small alphabet explains most of the observations) and expressive (the representation can describe a variety of contours that are not accounted by the standard alphabet).

Secondly, the distribution of the most observed contours is detailed for some prosodic/syntactic units in figure 6 (here, syllable, discourse markers, and illocutionary units). In particular, the alphabet substantially changes depending on the prosodic/syntactic unit : in comparison with the common syllable unit, the discourse marker and illocutionary units present a larger variety of contours (e.g., extreme ranges, complex contours), that potentially instantiates various functions : from modalities, to semantic and pragmatic. This constitutes a second validation concerning the labelling of contours over various prosodic/syntactic units : an alphabet of contours can be derived specifically for each prosodic/syntactic unit.

In conclusion, the representation can derive a small alphabet of contours that is specific to each prosodic/syntactic unit, that can be advantageously used for automatic labelling. This is crucial for further research on the role of prosody in speech communication : the grail search for the mapping of forms and functions.

4. Conclusion

This paper presented a simple method for the automatic labelling of intonation. The proposed method presents various advantages over existing methods : the alphabet of contours is fully data-driven, an explicit time-frequency representation is used to derive complex contours, and contours can be determined over arbitrary prosodic/syntactic units. A preliminary experiment conducted on 3 hours of spoken French indicates that a small number of contours is sufficient to explain most of the observed contours. The method can be easily adapted to other stressed languages. The implementation is open-source and freely available. This representation will be further used to study the role of speech prosody in speech communication, from prosody/syntax and prosody/discourse interfaces [24, 25], to the modelling of speech prosody for text-to-speech synthesis and voice conversion [26].

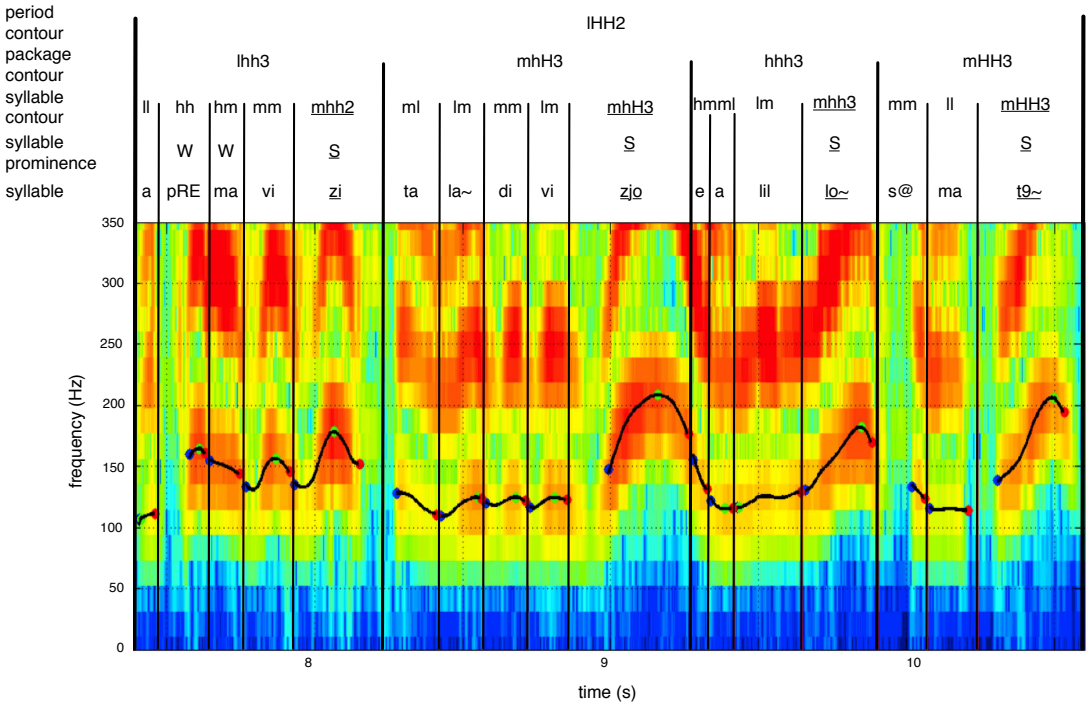


FIGURE 5 – *Acoustic representation and labelling of contours over syllable, prosodic package, and prosodic period for the speech sequence : "Après ma visite à Landivisiau et à l'île Longue ce matin" ("After my visit to Landivisiau and l'île Longue this morning"). Speech sample [Rhap-M2001, corpus C-PROM] : monologue, ordinary speech. Blue and red dots denote initial and final values, respectively ; and green dots intermediate saliencies. For syllable prominence : W indicates a weak prominence, and S a strong prominence. Information about the last syllables of a prosodic package are underlined.*

PROSODIC UNITS		# CONTOURS (> 95%)	SYNTACTIC UNITS		# CONTOURS (> 95%)
syllable	(43192)	8	discourse marker	(1749)	7
word	(32083)	9	pre-kernel	(1159)	15
foot	(22705)	11	post-kernel	(220)	39
group	(18104)	12	in-kernel	(192)	23
package	(14206)	15	illocutionary unit	(3050)	28
period	(2507)	29	...	...	...

TABLE 3 – *Occurrence of the contours observed over the prosodic/syntactic units. From left to right : nature of the prosodic/syntactic unit, total number of units observed, and number of contours that explain 95% of of the observed contours for each unit.*

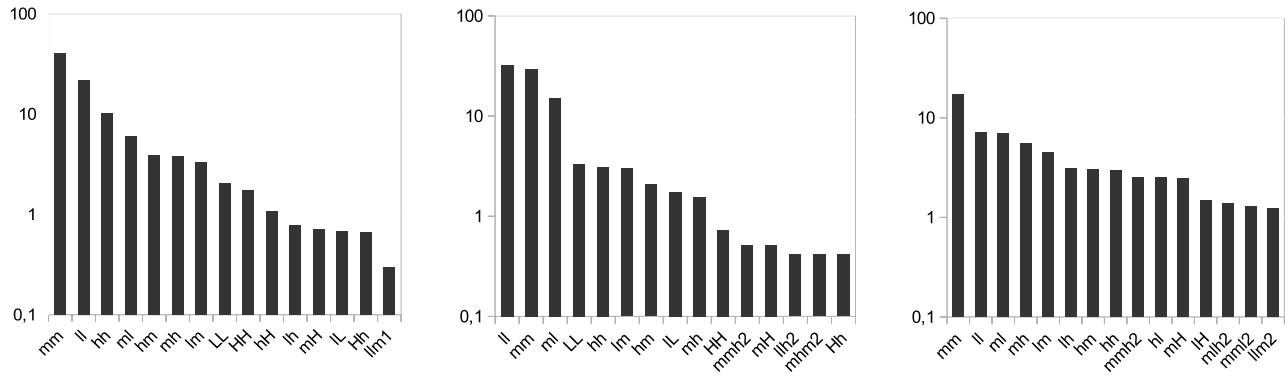


FIGURE 6 – *Proportion of the 15 most frequent contours observed for a set of prosodic/syntactic units (log % of occurrences). From left to right : syllable, discourse marker, and illocutionary units.*

5. References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI : a Standard for Labeling English Prosody," in *International Conference of Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.

[2] P. Taylor, "The Rise/Fall/Connection Model of Intonation," *Speech Communication*, vol. 15, pp. 169–186, 1994.

[3] E. Campione, D. Hirst, and J. Véronis, *Automatic Stylisation and Symbolic Coding of F0 : Implementations of the INTSINT Model*. Dordrecht : Kluwer, 2000, ch. Intonation. Research and Applications.

[4] B. Post, E. Delais-Roussarie, and A.-C. Simon, "IVTS, un Système de Transcription pour la Variation Prosodique," *Bulletin de la Phonologie du Français Contemporain*, vol. 6, pp. 51–68, 2006.

[5] P. Delattre, "Les Dix Intonations de Base du Français," *The French Review*, vol. 40, no. 1, pp. 1–14, 1966.

[6] V. Aubergé, "La Synthèse de la Parole : "Des Règles aux Lexiques"," PhD. Thesis, Université Pierre Mendès-France, Grenoble, France, 1991.

[7] K. Syrdal, A., J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody," *Speech Communication*, vol. 33, no. 1-2, pp. 135–151, 2001.

[8] P. Mertens, "The Prosogram : Semi-Automatic Transcription of Prosody based on a Tonal Perception Model," in *Speech Prosody*, Nara, Japan, 2004, pp. 549–552. [Online]. Available : http://bach.arts.kuleuven.be/pmertens/prosogram/

[9] ——, "Automatic Labelling of Pitch Levels and Pitch Movements in Speech Corpora," in *Tools and Resources for the Analysis of Speech Prosody (TRASP)*, Aix-en-Provence, 2013.

[10] H. Kawahara, H. Katayose, A. De Cheveigné, and R. D. Patterson, "Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity," in *Eurospeech*, Budapest, Hungary, 1999, pp. 2781–2784.

[11] A. De Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," *Journal of the Acoustic Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[12] A. Camacho, "SWIPE : A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music," PhD. Thesis, University of Florida, 2007. [Online]. Available : http://www.cise.ufl.edu/~acamacho

[13] P. Lanchantin, A. Morris, X. Rodet, and C. Veaux, "Automatic Phoneme Segmentation with Relaxed Textual Constraints," in *International Conference on Language Resources and Evaluation*, Marrakech, Morroco, 2008, pp. 2403–2407.

[14] J.-P. Goldman, "EasyAlign : a Semi-Automatic Phonetic Alignment Tool under Praat," in *Interspeech*, Florence, Italy, 2011. [Online]. Available : http://latlntic.unige.ch/phonetique

[15] B. Bigi and D. Hirst, "SPeech Phonetization Alignment and Syllabification (SPPAS) : a Tool for the Automatic analysis of Speech Prosody," in *Speech Prosody*, Shanghai, China, 2012. [Online]. Available : http://aune.lpl.univ-aix.fr/~bigi/sppas

[16] S. Young, "The HTK Hidden Markov Model Toolkit : Design and Philosophy," *Entropic Cambridge Research Laboratory, Ltd*, vol. 2, pp. 2–44, 1994.

[17] N.Obin, F. Lamare, and A. Roebel, "Syll-O-Matic : an Adaptive Time-Frequency Representation for the Automatic Segmentation of Speech into Syllables," in *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013.

[18] D. Hirst and R. Espesser, "Automatic Modelling of Fundamental Frequency using a Quadratic Spline Function," in *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, 1993, pp. 71–85.

[19] P. Taylor, "The TILT Intonation Model," in *International Conference on Spoken Language Processing*, Sydney, Australia, 1998, pp. 1383–1386.

[20] C. D'Alessandro and P. Mertens, "Automatic Pitch Contour Stylization using a Model of Tonal Perception," *Computer Speech and Language*, vol. 9, no. 3, pp. 257–288, 1995.

[21] J. Teutenberg, C. Watson, and P. Riddle, "Modelling and Synthesising F0 contours with the Discrete Cosine Transform," in *International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, U.S.A, 2008, pp. 3973–3976.

[22] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, *Prosodic Typology - The Phonology of Intonation and Phrasing*. Oxford University Press, 2005, ch. The Original ToBI System and the Evolution of the ToBI Framework, pp. 9–54.

[23] A. Lacheret, J. Beliao, A. Dister, K. Gerdes, J.-P. Goldman, S. Kahane, N. Obin, P. Pietrandrea, and A. Tchobanov, "Rhapsodie : a Prosodic-Syntactic Treebank for Spoken French," in *Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland, 2014. [Online]. Available : www.projet-rhapsodie.fr

[24] J. Beliao, "Characterizing Genres through Syntax and Prosody," in *European Summer School in Logic, Language and Information*, Düsseldorf , Germany, 2013, pp. 1–12.

[25] A. Lacheret, S. Kahane, and P. Pietrandrea, *Rhapsodie : a Prosodic and Syntactic Treebank for Spoken French*. Amsterdam, Benjamins, 2015.

[26] N. Obin, "MeLos : Analysis and Modelling of Speech Prosody and Speaking Style," PhD. Thesis, IRCAM - UPMC, 2011.

Lexical Stress in Brazilian Portuguese in Contrast with Spanish

Antônio R.M. Simões

Department of Spanish and Portuguese, University of Kansas, Lawrence, USA

asimoes@ku.edu

Abstract

This study discusses stress assignment in prosodic, non-verbal words in Brazilian Portuguese, in comparison with descriptions of stress assignment for Spanish [9, 13, 15, 16, 17, 18]. Given the conflicting claims regarding stress assignment in Brazilian Portuguese (see [11, 1, 2, 10, 3]) there is still a need to revisit discussions on stress assignment in Portuguese. In general, stress assignment in Spanish has been explained through the interplay between the morphological and phonological domains. Similar descriptions for Portuguese still requires far more abstraction and use of artifacts than in Spanish, which makes Mattoso Câmara Jr.'s [4, 5] claim that lexical stress is unpredictable in Brazilian Portuguese surprisingly unchallenged.

Index Terms: stress assignment, prosodic stress, syllable weight, Spanish and Portuguese

1. Introduction

This is a phonological study in progress developed to provide the author with adequate grounds for future phonetic analyses of stress and intonation systems in Spanish (SP) and Brazilian Portuguese (BP) in contrast. Therefore, this study is not yet based on actual data analysis, but instead on current theoretical descriptions of stress assignment.

In BP and SP there is a strong pressure to stress words on the penultimate syllable, i.e. paroxytones. In both languages, the great majority of words are paroxytones. Despite this coincidence, predicting lexical stress BP is not as governable as in Spanish. Among the differences in phonological and phonetics patterns between both languages, the most significant is perhaps the phonetically weak or unstable surfacing of postonic syllables in BP and the relatively stable surfacing of postonic syllables in Spanish.

This different behavior of postonic syllable in both languages is reflected in versification. Spanish and Portuguese count syllables in verses differently, in ways that reflect their rhythmic patterns. Portuguese counts the number of syllables until the last stressed syllable, whereas in Spanish the number of syllables is computed until the last stressed syllable plus one, regardless of the physical existence of a postonic syllable.

Thus, Martí's verses below have eight syllables each. As an illustration, if we did the counting of Spanish verses in the way Portuguese versification does, these verses would have seven syllables each. But in fact, they have eight syllables each. The dot (.) indicates syllable boundaries after resyllabification, accounted as needed. The last lexically stressed syllables in each verse are in capitals.

Yo . soy . u . (also so . yu) nhom.bre . sin.CE.ro (8 syllables)
De . don.de. cre.ce. la. PAL.ma, (8 syllables)
Y an.tes . de. mo.rir.me . QUIE.ro (8 syllables)
E.char . mis .ver.sos. de.l AL.ma. (8 syllables)
(...)
Oi.go un . sus.pi.ro, a. tra.VÉS (7+1 = 8 syllables)
De . las . tie.rra.s y . la . MAR, (7+1 = 8 syllables)
Y . no e.s un . sus.pi.ro,—ES (7+1 = 8 syllables)
Que . mi hi.jo . va a . des.per.TAR. (7+1 = 8 syllables)
(Poesía de José Martí, Versos Sencillos, 1981)

These differences and similarities of both languages can be further refined. Similar trends to move stress in proparoxytones to paroxytonic position happen in both languages, although the phonological processes are different.

Spanish	**Brazilian Portuguese**
olimpíadas → olimpiAdas	*abóbora → aBObra*
¡Pórtate bien! → ¡PorTAte bien!	*xícara → XIcra*

The preceding examples show similar trends in both languages. There are, however, important differences to take into account when attempting to propose stress assignment algorithms in Portuguese. For example, in SP, and only to a certain extent in Portuguese, a great number of words stressed on the antepenultimate syllable are learned words or *palabras cultas* or *palavras cultas* in both languages, which are sometimes taken from Greek sometimes from Latin, e.g. *Arquímedes* in SP but *ArquiMEdes* in BP, *Demóstenes* in both languages, *hypérbaton* in SP and *hipérbato* in BP, *épsilon* in SP and *ép[i]silon*, *ép[i]silo*, *íp[i]silon* or *ip[i]diLOne* in BP, *máximum* or *máximo* in SP and *máximo* in BP, *régimen* in SP but *reGIme* in BP and many other examples. Thus, while trends in SP are relatively more predictable, BP shows no clear trends, i.e. less predictability. This lack of clear trends permeates BP, contrary to SP.

A comparison of trends to paroxytone patterns in SP hypocoristics with no such trends in BP further reveals the difficulty researchers face to create an algorithm to predict stress assignment in BP. Hypocoristics in both languages are enlightening in this discussion. Whereas SP has a predominant pattern of disyllabic paroxytones for hypocoristics, BP produces disyllabics, monosyllabics, paroxytones and oxytones hypocoristics, without particular trends, as the Table 1 illustrates.

SP does have dialectical variations that use monosyllables in hypocoristics, e.g. Daniel ~ Dan, Cristina ~ Cris, but it happens less frequently than the patterns above, and it usually happens in closed syllable (CVC), while monosyllables in BP have open syllables (CV). In BP the variations are much greater.

Penultimate stress is also more predictable in SP loanwords, acronyms and foreign proper names, but not in BP, as depicted in the comparison below in Table 2. These examples show further the greater tendency in SP to stress penultimate syllable, compared to Portuguese.

The next section will discuss the notion of prosodic words and metrical notions common in the American generative frame of Metrical Theory. This study is only using the generative frame to discuss metrical theory as it has been applied to SP, and to show how difficult if not impossible and unmotivated it is to attempt to predict lexical stress in Portuguese. In other words, although the generative frame is very useful to discuss stress

assignment in any language, this study does not support the claim that it can predict stress assignment in Portuguese.

Table 1. *A comparison of trends in Spanish hypocoristic and the lack of trends in hypocoristic in BP.*

Spanish			Brazilian Portuguese		
Noun	**Hypo coristic**	**Stress Pattern**	**Noun**	**Hypo coristic**	**Stress Pattern**
Adriana	*Adri*	paroxytone	Benedito	Benê	oxytone
Daniel	*Dani Dan*	paroxytone mono-syllable	Fernando	Nando Fê	paroxytone mono-syllable
Francisco	*Pancho*	paroxytone	Francisco	Chico Chicô	paroxytone oxytone
Juan Ramón	*Juanra*	paroxytone	Gustavo	Gugu Gu	paroxytone mono-syllable
José	*JOse, Pepe*	paroxytone	José	Zé	mono-syllable
MiGUEL	*MIguel*	paroxytone	Rodrigo	Ro	mono-syllable
Ignacio	*Nacho*	paroxytone	Pedro	PePEU PEpe	oxytone, paroxytone
Ariel	*Ari*	paroxytone	Maria José	Zezé	oxytone
Antonio	*Toño*	paroxytone	Antônio	Totonho Tonho Tunico	paroxytone trisyllable

Table 2. *A comparison of paroxytone and oxytone trends in SP and BP loanwords. This table was produced with the help of eleven native speakers of SP and five native speakers of BP, who answered to a questionnaire (not here for lack of space), sent to them by e-mail. LW stands for loanwords, AC for acronyms and FN for foreign proper names.*

Spanish		Brazilian Portuguese	
LW, AC, FN	**Stress Pattern**	**LW, AC, FN**	**Stress Pattern**
barman	BARman	barman	barMAN
email	Email	email	eMAIL
cocktail	COCtel	cocktail	coqueTEL
karaoke	karaOke	karaoke	karaoKE
Gorbachev	GorbaCHEV, GorBAchev	Gorbachev	GorbaCHEV[i]
Muhammad Ali	MuhamMAD MuHAMmad	Muhammad Ali	MuhamMAD[i]
PEMEX	PEmex	PEMEX	peMEX
PC	Pc (PEce)	pC (peCE)	pC

Given the three types of syllable prominences in words of the two languages, e.g. the SP triplet *CÉlebre, ceLEbre, celeBRÉ*, the BP triplet *PÁssara, pasSAra*[1], *passaRÁ*, the two languages have proparoxytones, paroxytones and oxytones. In order to discuss stress assignment in the following sections, it would be helpful to firstly review the concepts of prosodic word, prosodic stress and syllable weight, and then discuss the case of the most common pattern, the paroxytones, then the proparoxytones and finally the oxytones.

[1] Postonic vowels in BP change considerable in quality, but in the case of the /a/, an inherently strong vowel and more resistant to significant changes in quality, the stress contrast in this triplet is still a valid one, especially in careful clear speech.

2. Discussion

Although I do not follow the common division of words into a simple classification of two main classes, **content** and **function** words, it is helpful to use them in this discussion. Thus, in Linguistics it is common to say that content words are the only ones that receive lexical stress, while function words are unstressed. By the same token, prosodic words have stress and function words or **clitics** do not. Although I do not see a problem to classify words in terms of content and function words, I do not agree that they correlate 100% with having or not having (lexical) stress. Function words sometimes have one **prominent** syllable if they have more than one syllable, e.g. the word for "while," *enquanto* in Portuguese and *mientras* in SP. But it is helpful to assume the notion of stress used here, because it makes it easier to compare stress assignment in SP and Portuguese within the generative frame of Metrical Theory. I am also using this view to make this discussion manageable. I am using this view to refuse current explanations that generative grammar is capable of predicting word stress in BP.

Taking the preceding remarks into consideration, mora (μ) is the unit that makes the prominent weight of prosodic feet. In SP, its weight is generally uniform, one mora. In SP, contrary to English, even in complex nuclei like diphthongs, all vocalic features of a syllable nucleus fit into one mora, as illustrated with the word *sentimiento*:

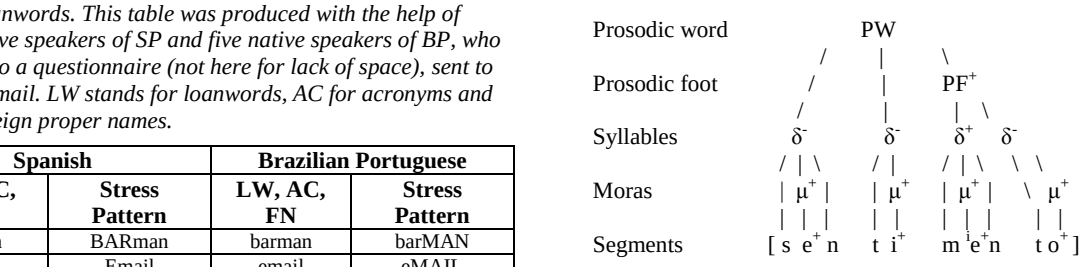


Figure 1. *The prosodic structure of the Spanish word* sentimiento*, to illustrate the concept of mora (μ).*

SP does not contrast the so-called heavy or bimoraic syllables and light or monomoraic syllables. The words "array, key, pie and tear (=rip)" in English, for example, contrasts heavy and light syllables: aᵘ.'rraᵘyᵘ, 'keᵘyᵘ, 'piᵘeᵘ, 'teaᵘrᵘ. Hence, in English one syllable words are binary in terms of the number of moras. The majority of English words have one syllable. SP needs two syllables to have two moras and the majority of SP words have two syllables. BP shows these SP characteristics without the regularity or uniformity found in SP. For example, in Rio, and maybe in some other areas of Brazil, but particularly in Rio, we find bimoraic and monomoraic contrasts, e.g. when answering the phone: "Alô!," "a ᵘ.loᵘaᵘ." It is important to keep in mind that we are thinking of a system found in social middle classes and higher. If we go into the SP spoken in rural areas of the Hispanic world and in the underworld of drug dealings, these will be completely different varieties of SP. The same can be said about Portuguese spoken in similar contexts.

The structure of paroxytonic words contains a nucleus in the two last syllables, identified with the symbols "<" and ">":

cua<derno>, <casa>, carre<tera>, pensa <miento>, <tráiga> melo, desespe<rándo>nos, desafortunada<mente>. These nuclei are *prosodic feet*, as illustrated with the word *final*<*mente*>, in Figure 2.

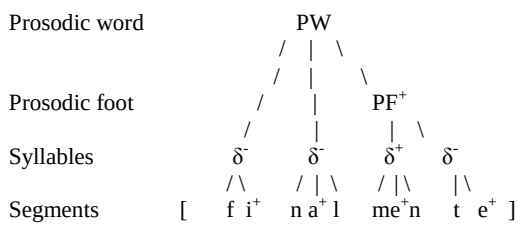


Figure 2. *The prosodic foot <mente> in the Spanish word* finalmente.

Given the framework used in this discussion, the preceding structure helps determining stress in SP paroxytones because words like "finalmente" contain all the basic structural requirements of finality, trocaicity and binarity. In BP we also find similarities in paroxytones. The obstacle one will find in BP is to fit proparoxytones and oxytones into these ideal structural property requirements. According to the generative frame, in SP, although oxytones seem, in a superficial look, to have unary foot (co.li.<brí>), and proparoxytones terciary foot (<sá.ba.do>), this can be solved with a morphological interpretation, as shown later in this discussion.

SP is known to have words with proparoxytonic stress because proparoxytones include vowels without morae, resulting in **extrametrical** elements.

Words like *mínimo, sábana, número* will have a non-moraic vowel in its root (mín<u>i</u>m-o, sáb<u>a</u>n-a, núm<u>e</u>r-o). Consequently, root morphemes such as these, with one syllable without weight, carry the potential to cause **retraction of stress**, as illustrated with the word *espárrag-o* in Figure 3, because their penultimate syllable *(-rra-* in the case of *espárrago*) is invisible to rules, or extrametrical. This characteristic or idiosyncrasy of proparoxytones leads to the conclusion in the generative framework that stress assignment in SP is morphologically conditioned. This morphological condition also applies in a different way to oxytones.

This invisibility, usually considered a relatively "small problem" in generative Metrical Theory, can be said to reduce the transparency of these processes. Other words carry this invisibility: *murciélago, máscara, árabe*. Portuguese for example, commonly deletes the non-moraic vowels like the one in *máscara* as *mascra*.

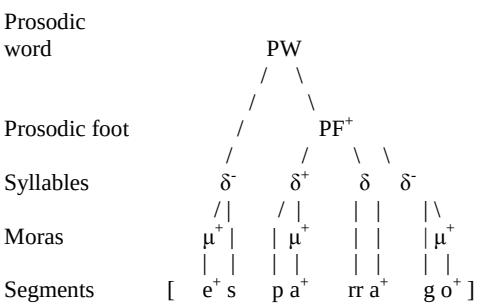


Figure 3. *The morphological interpretation of proparoxytones as underlying paroxytones.*

Therefore, root morphemes can result in stress retraction. Suffixes can also cause retraction. Derivative suffixes also result in retraction, as in words with the suffixes *–ic* as in *metál<u>i</u>co, canón<u>i</u>co.*

Here, as in other seemingly exceptional cases, morphological elements help understanding. The main explanation in cases such as *astronómico* is based on the concept of morphological nuclei. In other words, the morphemes in *astronómico* are *astr-o-nom-ic-o.* The morphemes *–nom* and *–ic* are some of the morphemes that have a non-moraic vowel. Given that *–ic*, and not *–nom*, is the morpheme that characterizes "astronómico" as adjective, then *–ic* is the nuclear morpheme that keeps the vowel invisible. In this exceptional case, *-nom* then is not a nuclear morpheme and consequently it carries one mora.

The same arguments can be attempted in Portuguese and most of the examples in SP are similar in BP: *mínimo, número, canônico, metálico, cronômetro, astrônomo*. Therefore, according to the preceding discussion, proparoxytones have the conditions of syllable foot on the right of the word, just as paroxytones do, as follows. (1) Only morphemes that function as morphological nucleus can retain non-moraic vowels. This also explains why this type of word is not common; (2) Given the condition above, the same principles of paroxytones apply to proparoxytones: binarity, finality and trocaicity.

If we keep the same view we have been using in this discussion, oxytonic words are also morphologically conditioned. Whereas derivative morphemes or morphemes with morpho-syntactic function interferes in the irregularity of stress assignment of proparoxytones in SP, in the case of oxytones the reason of irregularities has to do with morphemes whose function is exclusively morphological. In SP, the morphemes that have an exclusively morphological function are the class markers and markers of person/number, which normally appear at the very end of words. A distributive test helps knowing which ones mark class and the ones that do not. For instance, if we add *–er* morpheme to *joy<u>a</u>*, we obtain *joyero*, and not *joyaero. Therefore this *<u>a</u>* in *joya* is a class marker. Likewise, *guante* obtains *guantera* (not *guanteera).

In the case of *café*, we obtain *cafetera* with the insertion of *t* to preserve the *e*, which indicates that *e* in *café* is not a class marker. Likewise *maní – manicito, sofá – sofacito*, with the insertion of *c* to preserve the vowels *i, a*, in these cases with diminutive *–it*, these vowels are not class marker either. The conclusion is that oxytones ending in vowel behave as if they lacked class marker.

Oxytones ending in consonants also share this trait, *mujer*, *laurel*, *caimán*. Taking into account that in SP the majority of words ending in consonant are oxytones, there is a relation between oxytones and the lack of word markers. Furthermore, this lack of word marker leaves an "empty" slot at the end of words.

All common SP words or *palabras patrimoniales* have class markers. That is because some words have prosodic structure, but not segmental structure. In other words their morphemes are **catalectic**. Basically, SP has regular moraic class markers (all five vowels a, e, i, o, u) and catalectic class markers.

Having prosodic but not segmental structure creates opacity. Such a description of stress assignment in oxytones says that the segment is invisible although there is a prosodic structure.

There is evidence for this claim. For instance, if we take into account the singular and plural forms for *sofá* and *mujer*, this claim may be more easily accepted despite the opacity of the description with respect to the oxytones.

Therefore, plural formation in SP also supports this idea of catalectic class markers. SP forms plurals with the morpheme/suffix *–s*. Since *–s* is inflectional, it will go after derivate morphemes. In words with regular plural formation, the process is transparent: *pal-o-s, car-a-s, cruc-e-s, curs-i-s, trib-u-s*. But in the case of catalectic class markers adding plural morphemes forces the catalectic element to acquire a segmental content: *sofá-e-s, maní-e-s, mujer-e-s, caiman-e-s*.

This additional *e* with the plural morpheme reveals the "emptiness" of the catalectic space, as well as other characteristics of SP such as the preference for CV sequences and the use of epenthetic *e*, which is the only vowel that SP uses in epenthesis. In its plural form, the word becomes paroxytone. Given this argumentation, it is reasonable to say that all oxytones, regardless of its ending in vowels or consonants and contrary to paroxytones and proparoxytones, lack a class marker, that is to say inflectional morphemes.

A way to explain this lack of inflectional morphemes is to assume that the space is there, and it can be said to be a prosodic "space," not segmental. Therefore, the underlying structure of oxytones should have a prosodic mora that is filled out with a physical segment as needed. This leads, among other consequences, to assigning an underlying *e* vowel that will surface in plural forms of oxytones. This description has a number of advantages and the most obvious is the generalization of assigning only *-s* to plural, instead of two rules, one for adding *-s* and another to add *-es*.

In the case of words like *café* and *té* the insertion of *e* still happens, and in actual speech this fusion is common in SP (*mijito* for *mi hijito*) or *alcol* for *alcohol*). Some dialects have variations (e.g. *sofás, rubís*), but the description of stress assignment must be maintained within the same register of a dialect of reference. Otherwise, if one considers how different SP can be throughout the Hispanic world, stress assignment as a general trend would be unnecessarily difficult if not impossible to describe. It is simpler and desirable to stay within the limits of a given register that is considered representative, and then move on into different dialectical variations. For the intended description of stress assignment for this discussion, we used the acrolect as reference. Taking the educated register into consideration, we need to keep in mind that the morphological nucleus changes by adding a derivative suffix, e.g. *verd-e, verd-ur-a* or *verd+or+catalexis*.

As *–ur* and *–or* become the morphological nuclei, they select the segment *a* and a prosodic mora as class markers. Among morphemes that select class markers, the most commons in SP are *-dad*, *-itud*, *-ción*, *-zón*, *-dor*, *-al*, *-il*, and *–r*.

There are obviously serious flaws with the framework just discussed. It requires accepting an *adhoc* description and the assumption that the opacity is an acceptable price to pay, given the other advantages of the description. These flaws are even worse in the case of Portuguese.

In the case of oxytones in BP, for example, one can attempt to propose a process similar to the one in SP. All non-verbal plural would be formed with one rule, by adding *-s* to the singular form. Through this view, in BP, in the case of oxytones, there is no physical segment in the slot where normally an inflectional morpheme is expected. Words need a class marker to indicate its lexical class. The explanation then, seems to be similar to the one for SP, which would be desirable without flaws because it would require only one rule of plural formation. Therefore, if oxytones have an invisible prosodic unit that we can call mora, it materializes as *e*. Portuguese words like *sofá*, *café*, *colibri*, *mulher* and *mar* would then add *-s* in their plural. After the addition of *s*, *e* would be inserted before s, and finally carry on other known phonological processes of BP as needed:

sofá_ → sofá_s → sofáes → sofás
café_ → café_s → cafées → cafés
colibri_ → colibri_s → colibries → colibris
mulher_ → mulher_s → mulheres → mulheres
mar_ → mar_s → mares → mares
azul_ → azul_s → azules → azues → azuis

The processes above can also be viewed as tree diagrams in the singular and plural forms:

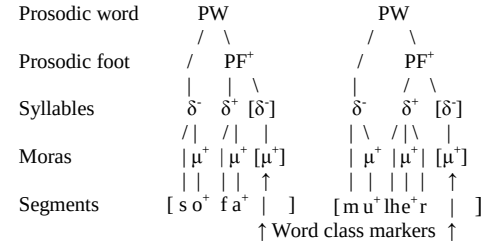


Figure 3. *Brazilian Portuguese words* sofá *and* mulher.

As it can be seen, however, it would be necessary to propose biased underlying forms based on historical forms for other words **brasiles*, **papeles*, **colchones*, which have no synchronic correspondences or motivation in BP. Then, it may be simpler, more efficient and more realistic to maintain Mattoso Câmara Jr.'s [4, 5] claim that lexical stress in Portuguese is unpredictable. One must use memory to learn stress placement in Portuguese.

3. Conclusions

This study examined lexical stress assignment in SP to help understanding why lexical stress is unpredictable in BP. The degree of abstraction, the creation of artifacts and various other flaws to create an algorithm for lexical stress assignment in BP make such an algorithm unnecessarily complicated. In the history of sciences we well know that the more complicated a description is, the far we are from truth.

Unpredictability should not be regarded as undesirable. It is just inherent in some aspects of natural human languages. Stress assignment is most likely unpredictable in Russian [12], to appear) and very likely in English as well [14], in spite of the claims in Chomsky and Halle [6] that it can be predicted.

4. Acknowledgements

I am very thankful to Juliette Blevins, at CUNY, and Seung Hwa Lee, at UFMG, in Brazil, for their helpful discussions with me, about lexical stress. I am also thankful to the anonymous reviewers who made valuabe comments to improve this study. The interpretations in this study are mine.

5. References

[1] BISOL, Leda. 1994. O Acento e o Pé Métrico Binário. In *Letras de Hoje* 98, 25-36. Porto Alegre: ediPUCRS.

[2] BISOL, Leda, org. 2005. *Introdução a estudos de fonologia do português brasileiro*. Porto Alegre: EDIPCURS, 2005.

[3] CAGLIARI, Luiz Carlos. 1999. *O Acento em Português*. Campinas: Editora do Autor.

[4] CÂMARA, Jr, Joaquim Mattoso. 1970. *Estrutura da Língua Portuguesa*. Petrópolis, RJ: Vozes.

[5] ---------- . 1979. *The Portuguese language*. Chicago: University of Chicago Press.

[6] CHOMSKY, Noam and Morris HALLE. 1968. *The Sound Pattern of English*. New York: Harper & Row, 1968.

[7] DOMAHS, U., I. PLAG and R. CARROLL. (To appear). Word stress assignment in German, English and Dutch: Quantity-sensitivity and extrametricality revisited. In *Journal of Comparative Germanic Linguistics*.

[8] LANGACKER, Ronald. *Fundamentals of linguistic analysis*. New York: Harcourt Brace Jovanovich,Inc., 1972.

[9] HUALDE, José Ignacio. 2005. *The Sounds of Spanish*. New York: Cambridge University Press

[10] LEE, Seung-Hwa. 2007. O Acento Primário no Português: Uma Análise Unificada na Teoria da Otimalidade. In *O Acento em Português: Abordagens Fonológicas*. São Paulo: Parábola Editorial, 121-143.

[11] MATEUS, Maria Helena Mira. 1983. O Acento de Palavra em Português: Uma Nova Proposta. Lisboa: *Boletim de Filologia* 27, 211-229.

[12] MOLCZANOW, J., U. DOMAHS, J. KNAUS and R. WIESE. (To appear). The lexical representation of word stress in Russian: Evidence from event-related potentials. In *The Mental Lexicon*.

[13] NÚNEZ CEDEÑO, Rafael A. and Alfonso MORALES-FRONT. *Fonología generativa contemporánea de la lengua española*. Washington, D.C.: Georgetown University Press, 1998.

[14] PLAG, Ingo. 2006. The variability of compound stress in English: structural, semantic, and analogical factors. *English Language and Linguistics* Volume 10 Issue 01 / May 2006, 143-172.

[15] ROCA, Iggy M. 1999. "Stress in the Romance Languages. In *Word Prosodic Systems in the Languages of Europe*, editor van der Hulst, H. Berlin, DEU: Mouton de Gruyter, 659-811.

[16] -------- (1990), "Diachrony and synchrony in Spanish stress", *Journal of Linguistics* 26: 133–164

[17] -------- (2006) 'The Spanish stress window', in F. Martínez-Gil & S. Colina (eds), *Optimality-theoretic advances in Spanish phonology*, Amsterdam: John Benjamins. 239-77

[18] SOSA, Juan Manuel. *La entonación del español: su estructura fónica, variabilidad y dialectología*. Madrid: Cátedra, 1999.

Prosodic Characteristics of Vocalic Hesitations in Comparison with Overlong Vowels in Estonian

Rena Nemoto

Institute of Cybernetics at Tallinn University of Technology, Tallinn, Estonia
rena.nemoto@phon.ioc.ee

Abstract

The goal of this paper is to investigate vocalic hesitations in Estonian and compare them to the related vowels of overlong (Q3) quantity degree. We wonder if there are some language-specific characteristics of hesitations. If yes, which kind of characteristics can be observed in Estonian language? We analyze duration, fundamental frequency (f_0), intensity, and first two formants using 39.5 hours of manually transcribed mono- or dialogue speech from a spontaneous speech corpus. Investigated vocalic hesitations and Q3 vowels are: */ee*, *ää*, *aa*, *õõ*, *öö/*. The characteristics of hesitations as compared to those of Q3 vowels show that hesitations have longer duration range. Hesitations generally include lower f_0 and intensity values. However, the values vary in terms of vowels. First two formants of hesitations tend to be located at more centralized positions in a vocalic triangle than related Q3 vowels.

Index Terms: vocalic hesitation, Estonian, spontaneous speech, prosody

1. Introduction

Hesitations carry language-specific information. The transcription of vocalic hesitations varies across languages, for example uh/um in American English, *euh* in French or *eh* in Spanish suggesting differences in their perception by native speakers [1]. So we wonder what kind of specific can be found in Estonian language. An early Estonian disfluency study by Hennoste [2] described particles and um-s as initiators of repair in Estonian spontaneous speech. One of the most frequent um in spoken Estonian is *ee*. The um *ee* contained 38% of all um-s in the studied corpus. The other frequent um-s were expressed by *õõ*, *ää*, *mm*, etc. But little study has been done to explore these um-s in Estonian at acoustic level. We are then interested in studying these vocalic hesitations at acoustic level. Hesitations are expressed by lengthened vowels (cf. Figure 1 from [3]) or consonant /mm/, or mixing a vowel with a consonant, etc.

Estonian is a language with word-initial lexical stress in general [4], even though some exceptions like loanwords [5] can be found. And Estonian has a three-way quantity system referred as short (Q1), long (Q2), and overlong (Q3) quantity degree. Most studies of the three-way quantity system concentrate on disyllabic words and compare first and second syllables at foot level: for example (Q1) *sada* /sata/, hundred, singular nominative; (Q2) *saada* /saata/, to send, verb: singular imperative; (Q3) *saada* /saa:ta/, to get, verb: infinitive [5]. The quantity system is complex combining durational and tonal components. Lippus [6] explained the f_0 peak as follows: the peak for Q1 and Q2 is located usually in the second half of the stressed first syllable, but for Q3, it usually appears in the beginning of the stressed syllable vowel.

As overlong (Q3) vowels have longest duration among quantity degrees, we compare prosodic and acoustic behaviors between five vocalic hesitations (lengthened vowels) and its related overlong vowels. In this paper, we address the question as follows: how prosodic features of vocalic hesitations behave from the beginning to the end of a segment in comparison with their related Q3 vowels.

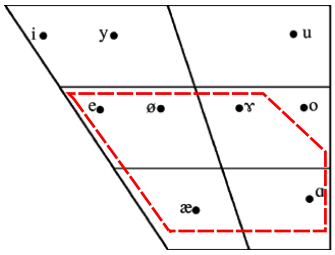


Figure 1: *Estonian vocalic system (vocalic hesitations are in red line).*

2. Corpus and methodology

2.1. Corpus

We used the manually transcribed phonetic corpus of Estonian spontaneous speech of the university of Tartu[1] [6]. Investigated corpus consists of 32 male and 39 female speakers of monologues or dialogues, including 18.5 hours for male and 21 hours for female speakers. The corpus is manually segmented at different levels: phoneme, syllable, word, quantity degree, voice quality, etc.

2.2. Methodology

Fundamental frequency (f_0), intensity, and two first formant (F1 and F2) values were extracted every 5 milliseconds (ms) by using Praat software [7]. Phonemic duration was taken from the manually segmented corpus. f_0, intensity, and first two formant measurements were averaged over segments. Also each phonemic segment was split in three parts (begin, center, end) and measurements of each part were also averaged. Voicing ratio was computed as following: for each segment, the number of voiced (f_0 >0 Hz) frames was divided by the total number of frames. F1 and F2 values were taken only when f_0 values were determined. The target vocalic hesitations and related overlong (Q3) vowels were extracted from the transcribed corpus. Hesitations were annotated by transcriptors. Table 1 gives the

[1] http://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech

number of occurrences of five vocalic hesitations and related Q3 vowels for male and female speakers. Analyzed Q3 vowels were intra-lexical vowels and vocalic hesitations occurred between silent pauses or words, after a word before a silent pause or inversely (cf. Figure 2).

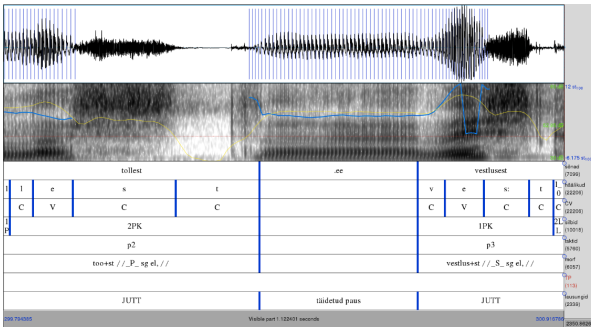


Figure 2: *Example of vocalic hesitation. Vocalic hesitation .ee is located between two words: "tollest (that, singular, elative)#.ee (vocalic hesitation)#vestlusest (conversation, singular, elative)".*

Table 1: *Number of vocalic hesitations and Q3 vowels for male (MS) and female speakers (FS).*

hesit./Q3 vow.	MS		FS	
	hesit.	Q3	hesit.	Q3
.ee/ee: [eeː]	1907	2166	1424	1936
.ää/ää: [ææː]	97	437	102	481
.aa/aa: [ɑɑː]	103	2205	89	2125
.õõ/õõ: [ɤɤː]	608	60	197	41
.öö/öö: [øøː]	68	358	65	359
Total	2,783	5,226	1,877	4,942

3. Acoustic and prosodic analyses

For this study, we compare duration, f_0, intensity, and two first formants (F1 and F2) between five vocalic hesitations and their related overlong (Q3) vowels. We hypothesize that Q3 vowels are intra-lexical vowels, so there are some acoustic and prosodic differences between vocalic hesitations and Q3 vowels.

3.1. Duration

Figure 3 shows duration distribution of five vocalic hesitations (red line) and their related Q3 vowels (black line). Mean hesitation duration reaches 318 milliseconds (ms) for male speakers (standard deviation (SD): 171 ms) and 270 ms for female speakers (SD: 142 ms), whereas mean duration of related Q3 vowels is 154 ms (SD 68 ms) for male and 174 ms (SD 82 ms) for female speakers. We notice that Estonian vocalic hesitations have much longer durations than related Q3 vowels as revealed in the literature like other languages. The profile of vocalic hesitation duration is more enlarged. The difference between hesitations and related Q3 vowels turned out to be statistically significant ($p < .001$) for both male and female speakers by Wilcoxon test using R software [8].

3.2. Fundamental frequency (f_0)

Estonian words in general have a stress with higher f_0 and intensity values on first syllable and these values gradually decrease

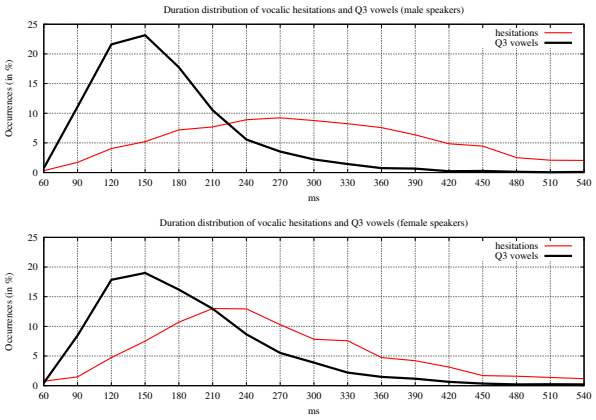


Figure 3: *Duration distribution of vocalic hesitations (in red line) and their related Q3 vowels (in black line) for male (top) and female (bottom) speakers.*

to last syllable. We hypothesize that f_0 contours are more stable for vocalic hesitations while Q3 vowel contours tend to fall within a segment. A segment is divided into three parts (begin, center, end) in order to compare f_0 contour movements of vocalic hesitations from Q3 vowels. Figure 4 presents average f_0 values of begin, center, and end parts over 70% of voicing ratios (to avoid extracting f_0 value errors) for each vocalic hesitation (left) and its related vowel (right). It is noticeable at a glance from Figure 4 that f_0 values are lower for vocalic hesitations than Q3 vowels for both male (top) and female (bottom) speakers. We can also observe that: (i) f_0 values slightly decrease from begin to center parts for both vocalic hesitations and Q3 vowels; (ii) for Q3 vowels, f_0 values continue to fall from center to end parts; (iii) however f_0 contours of vocalic hesitations are stable or they rise from center to end parts (except vocalic hesitation of *.aa* for female speakers where f_0 values continue to fall like Q3 vowels).

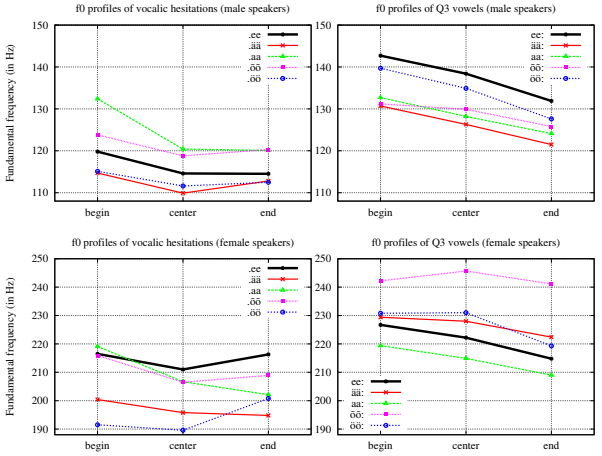


Figure 4: *f_0 contours of vocalic hesitations (left) and their related Q3 vowels (right) for male (top) and female (bottom) speakers.*

To verify whether these f_0 differences between vocalic hesitations and their related Q3 vowels are significant, we conducted statistical tests using Wilcoxon test. First, we compare mean f_0 values of each part (begin, center, end) between vocalic

hesitations and their related Q3 vowels. Three pairs (*ee*, *ää*, *öö*) of each part show statistically significant differences ($p < .05$) for both male and female speakers. Especially begin and center parts of these three pairs are strongly significant ($p < .001$) for both gender. However, for the *õõ* pair, only center part is statistically significant ($p < .05$) for both male and female speakers. No significance can be shown in the *aa* pair for both male and female speakers.

Next we compare the difference of f_0 values between begin and center, and between center and end. With respect to the difference of values between each part, the Wilcoxon test shows significant differences in the *ee*, *ää*, and *õõ* pairs for male speakers ($p < .05$), and the *ee* and *õõ* pairs for female speakers ($p < .001$). The difference between begin and center parts of the *aa* pair reveals significant differences for both male and female speakers ($p < .001$). The difference between center and end parts of the *öö* pair are statistically significant for both male and female speakers ($p < .001$) and of the *ää* pair has significant difference for female speakers ($p < .001$).

The *õõ* pair, showing significant difference only in mean center part, has significant difference between begin and center parts, and between center and end parts. The similar phenomenon is observed in the *aa* pair. These suggest that f_0 movements are important to separate vocalic hesitations and its related Q3 vowels.

3.3. Intensity

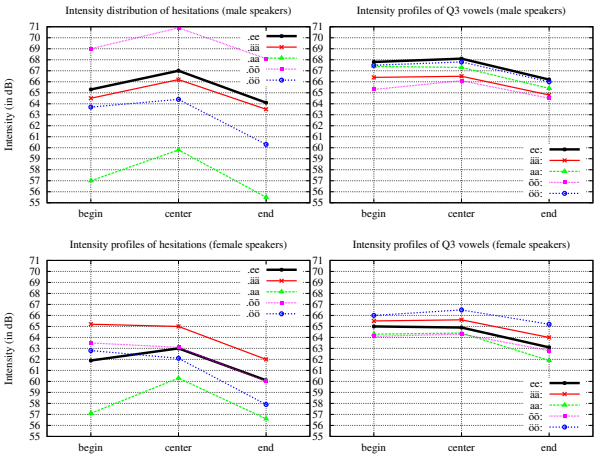


Figure 5: *Intensity distribution of vocalic hesitations (left) and their related Q3 vowels (right) for male (top) and female (bottom) speakers* .

Intensity values are expected to have similar contours as f_0 ones because Estonian words normally have a stress on first syllable. Like f_0, a segment is separated into three parts. Figure 5 (left) shows the mean intensity profiles of begin, center, and end parts of vocalic hesitations for male (top) and female speakers (bottom). For male speakers, intensity values of vocalic hesitations increase from begin to center parts and decrease from center to end parts. Intensity contours of the vocalic hesitations *.ee* and *.aa* for female speakers have also rise-fall contours, whereas the other *.ää*, *.õõ*, and *.öö* hesitation contours decrease from begin to end. However values fall more from center to end than from begin to center. The ranges of each vocalic hesitation values are large for both male and female speakers.

As for the related Q3 vowels (right), mean intensity values are quite stable or values rise slightly from begin to center parts and decrease from center to end parts for both male and female speakers. The range of each Q3 vowel values is not as wide as that of vocalic hesitations for both male and female speakers.

Like f_0, same statistical tests using Wilcoxon are carried out between vocalic hesitations and their related Q3 vowels. The comparison of mean intensity values of each part (begin, center, end) between vocalic hesitations and their related Q3 vowels shows significant differences in the *ee*, *aa*, *õõ*, and *öö* pairs for male speakers ($p < .005$). As for female speakers, the *ee*, *aa*, and *öö* pairs are statistically significant ($p < .001$). Mean intensity values of begin part for the *ää* pair are statistically significant for male speakers, however any statistical significance is not observed in center and end parts. Mean intensity values of begin and center parts do not show any significant difference in the *ää* and *õõ* pairs for female speakers, while significant differences are observed in the end part of the *ää* pair ($p < .001$) and the *õõ* pair ($p < .05$) .

The differences of intensity values between begin and center, and between center and end show significant differences in all pairs for female speakers ($p < .05$). As for male speakers, statistically significant differences are found in the *ee*, *ää*, and *aa* pairs for all pairs ($p < .005$). The difference of intensity values between begin and center part of the *öö* pair are statistically significant ($p < .001$) and also between center and end part of the *õõ* pair ($p < .001$). It is noticeable from statistical tests that not only mean intensity value comparison but also comparison of value difference between three parts are important to separate each characteristics like f_0 contours.

3.4. Formants: F1/F2

As F1 and F2 values of begin and end parts might be influenced by preceding and/or following segments, especially for Q3 vowels, which are in general located within a word, only center part values are taken into account for formant analysis. Center part values seem to give stable values than begin and end parts. Mean F1 and F2 values of center part are illustrated in Figure 6 for male speakers and Figure 7 for female speakers. Vocalic hesitations are presented in red on the top and their related Q3 vowels are in black on the bottom. It is noticeable from two figures that vocalic hesitations (top figures) are superposed each other. However it is observable that vocalic hesitation of *.aa* is less superposed and its ellipse is bigger than other vocalic hesitations for both male and female speakers. On the other hand, Q3 vowels (bottom figures) are distinguishable from each other, although ellipses of *õõ:* and *öö:* are quite superposed. Vocalic hesitation figures show that F1 and F2 positions of these five vocalic hesitations are more centralized to each other than their related Q3 vowels. Mean value of *.ee* is located in more posterior place and that of *.õõ* is more anterior than their related Q3 vowels. Two vocalic hesitations of *.ää* and *.aa* are situated at more closed positions. Concerning the *.öö* hesitation, it seems that mean F1/F2 values are close to those of its related Q3 vowel *öö:*. Vocalic hesitations were manually transcribed, so it might be possible that transcribed vowels and vocalic hesitations were influenced by each transcriptor's perception.

We conducted statistical tests (Wilcoxon test) to verify if these center part values of F1/F2 formants are different between five vocalic hesitations and their related Q3 vowels. Mean center values of F1 for male speakers show significant difference ($p < .001$) except the *öö* pair, whereas mean F2 center values are statistically significant for all pairs ($p < .05$), especially the *ee*, *õõ*, and *öö* pairs ($p < .001$). As for female speakers, statistically

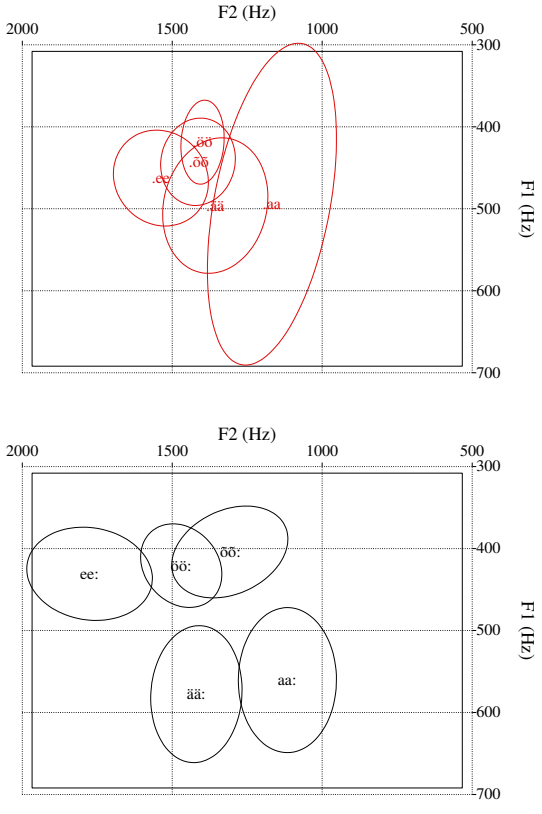


Figure 6: *F1/F2 center position plot of male speaker: hesitations (top, red) and their related Q3 vowels (bottom, black).*

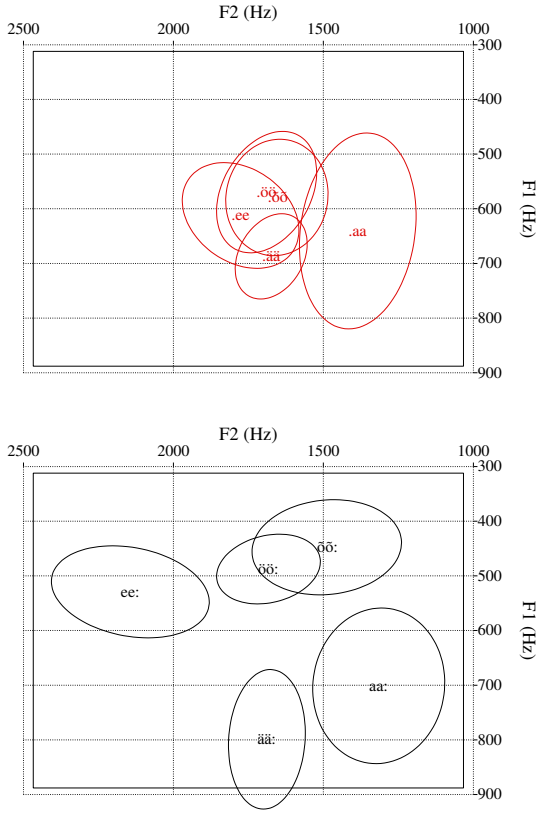


Figure 7: *F1/F2 center position plot of female speaker: hesitations (top, red) and their related Q3 vowels (bottom, black).*

significant differences of F1 values are observed in all pairs ($p < .005$). Concerning F2, no significant difference is found in the *ää* and *öö* pairs, whereas the other pairs show significant difference ($p < .001$). Statistical tests reveal the difference between vocalic hesitations and their related Q3 vowels in all pairs for both F1 and F2 values or for at least one of them.

4. Discussion

This paper described acoustic and prosodic characteristics (duration, fundamental frequency, intensity, and first two formants) of five vocalic hesitations and their related Q3 (overlong) vowels of Estonian language. Investigated five vowels were: *ee*, *ää*, *aa*, *õõ*, and *öö*. Both male and female speakers contained around 5,000 Q3 vowels and 2,800 vocalic hesitations for male speakers and 1,900 for female speakers in 40 hours of the spontaneous speech corpus.

Duration distribution revealed that longer durations were observed in vocalic hesitations than their related Q3 vowels. Duration range was more spread for vocalic hesitations. The f_0 contours of vocalic hesitations behaved differently in comparison with their related Q3 vowels. Lower f_0 values were found in vocalic hesitations. f_0 contours of Q3 vowels tended to decrease from begin to end parts, whereas vocalic hesitation contours decreased begin to center parts and were stable or rising from center to end parts. The results from intensity analysis confirmed different intensity contours between vocalic hesitations and Q3 vowels. The range of vocalic hesitations was bigger than Q3 vowels. Intensity contours of vocalic hesitations raised from begin to center parts and dropped from center to end parts, or decreased from begin to end parts, while Q3 vowel contours showed rise-fall contours. What differentiate vocalic hesitations from Q3 vowels was that intensity values of vocalic hesitations between center and end parts were bigger than those of Q3 vowels. Two first formant results showed that vocalic hesitations were superposed each other and they were more centralized than their related Q3 vowels.

This study concerned general tendencies of vocalic hesitations in comparison with their related Q3 vowels from averaged prosodic and acoustic values using large corpus. We wondered if vocalic hesitations depend on speaker specificities; one speaker might utter hesitations more opened, or other prefers more closed. Hesitations may be influenced by before and/or after phoneme or words. Further studies will include individual speaker aspects and take into account surrounding phonemes or words.

5. Acknowledgements

This research was supported by the European Regional Development Fund (ERDF) through the Estonian Center of Excellence in Computer Science (EXCS) and the Estonian Ministry of Education and Research target-financed research theme No. 0140007s12 to the author. We also thank Einar Meister and Lya Meister for their help and suggestions.

6. References

[1] Vasilescu, I., Nemoto, R. and Adda-Decker, M., "Vocalic hesitations vs vocalic systems: a cross-language comparison", Proc. of ICPhS, Saarbrücken, 2007.

[2] Hennoste, T., "Repair-initiating particles and um-s in Estonian spontaneous speech", Proc. of DiSS'05, Aix-en-Provence, pp. 83–88, 2005.

[3] Asu, E. L. and Teras, P., "Estonian", Journal of the International Phonetic Association, 39, pp. 367–372, 2009.

[4] Asu, E. L. and Nolan, F., "Estonian and English rhythm: a two-dimensional quantification based on syllables and feet", Speech Prosody, 2006.

[5] Meister, L. and Meister, E., "Perception of the short vs. long phonological category in Estonian by native and non-native listeners", Journal of Phonetic 39, 212–224, 2012.

[6] Lippus, P., "The acoustic features and perception of the Estonian quantity system, Ph.D. dissertation", Tartu University, 2011.

[7] Boersma, P. and Weenink, D., Praat: doing phonetics by computer, [Computer program], http://www.praat.org/.

[8] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, âĂŃhttp://www.R-project.or, 2012.

Articulatory Reorganizations of Speech Rhythm due to Speech Rate Increase in Brazilian Portuguese

Alexsandro R. Meireles [1], *Plínio A. Barbosa* [2]

[1] Phonetics Laboratory, Federal University of Espirito Santo, FAPES, Brazil
[2] Speech Prosody Studies Group, State University of Campinas, Brazil
meirelesalex@gmail.com, pabarbosa.unicampbr@gmail.com

Abstract

This paper examines how speech rate increase acts to change speech rhythm at the articulatory level. Main results show that speech rate increase worked to change articulatory parameters in the following way: a) decrease of acceleration duration; b) decrease of y-extremum; c) decrease of constriction displacement; d) decrease in modulus of peak and/or valley velocity; e) decrease of gestural duration; and f) constant proportional time-to-peak (or valley) velocity. Besides, results have shown that speech rate tends to affect all gestures in an utterance independently of their phrasal position. Nevertheless, there was evidence that some articulatory parameters could, if properly manipulated, provide cues for rhythmic restructurings in speech. Finally, results show that the dynamical speech rhythm model (Barbosa, 2007) is more appropriate to deal with Brazilian Portuguese acoustical data than the pi-gesture model (Byrd & Saltzman, 2003), and that both models could explain articulatory reorganizations due to speech rate increase.

Index Terms: speech prosody, speech rhythm, speech rate, EMA, articulatory study.

1. Introduction

The speech rate influence on speech rhythmic reorganization can be explained by two recently proposed prosodic models: (i) the speech rhythm model [1, 2, 3] (henceforth SRM); and (ii) the π-gesture model [5] (henceforth PG). A rhythmic reorganization is considered here as a change in the temporal prosodic structure of articulatory gestures [9]. Both models account for prosodic structuring using a dynamical systems approach (cf. [7, 8]).

Byrd and Saltzman [5] state the main features of prosodic gestures: (i) prosodic gestures have a temporal extension and overlap with constriction gestures; (ii) prosodic gestures' gestural scores represent the activity of a set of abstract point attractors, in order to make the model as abstract as possible; (iii) prosodic gestures do not have an independent articulatory realization. Therefore, they are only indirectly realized by its effect on the articulatory dynamics.

Barbosa [2, 3] highlights the linguistic characteristics of SRM, as follows: (i) the linguistic rhythm is a consequence of the way the phrase-stress oscillator pulses align with the onset of lexically stressed vowels; (ii) the hierarchy between the magnitude of phrase-stress oscillator pulses generates a dynamical metrics; (iii) the relative coupling force is language dependent; (iv) the prosody-segments interaction is different among languages; (v) syntactic information is crucial, though not alone, to explain the variability of placement and magnitude of phrase stresses; (vi) phrasal stresses are a superficial consequence of the prominences expressed by peaks of abstract duration of the syllabic oscillator; (vii) lexical stress is defined in the abstract gestural score.

These two prosodic models (SRM and PG) predict similar phonetic consequences to syllable-sized gestures, namely: (i) gestures adjacent to prosodic boundaries will be lengthened; (ii) degree of slowing will be greatest as the gesture approaches a prosodic boundary; (iii) boundaries of different strengths are only expected to be distinct in degree of effect. Nevertheless, the scope of action in the SRM model is global, i.e., the phrase stress oscillator acts throughout the whole utterance. The longer vowel-to-vowel (from the beginning of one vowel up to the beginning of the next vowel, henceforth VV) duration at the end of a stress group (Brazilian Portuguese is a right-headed language at this level) is a cumulative function of previous VV durations from the beginning of this group, while the scope of action in the PG model is local: "the π-gesture locally slows the clock that controls the timeflow of an utterance" [5, p. 160]. Another difference between them regards their units of action. While in the SRM these units are articulatory gestures on the lexical level whose stiffness is modified by the effect of the phrase stress oscillator pulses at the properly rhythmic level, in the PG model, these units are gestures spanning a phrasal boundary. Finally, π-gesture action is limited to phrase edges, as can be noticed in the passage: "effects will be limited to gestures near the domain edge and will not occur at gestures quite remote from it" [5, p. 162].

Despite these differences, it is worthwhile to highlight that both models fall into the category of the so-called intrinsic timing, for in the π-gesture model "the activation level dynamics of the clock and the constriction level dynamics of the gestural units are bidirectionally coupled and hence form a single higher-order dynamical collective" [5, p. 156]. Barbosa's dynamical model of speech production [2, 3] works likewise, for the two coupled oscillators, through prosody-segments interaction, is bidirectionally coupled with the gestural score and thereby also form a single higher-order dynamical collective. Furthermore, prosodic timing is explicitly controlled for utterance production in the SRM, that is, the coupled oscillators in interaction with the other levels of grammar control speech rhythm, so that a separate abstract executor is not needed to do this control.

Based on the assumptions of the speech rhythm model, this paper intends (i) to study how speech rate acts to change the temporal structure of articulatory gestures, as related to linguistic rhythm, and (ii) to compare the predictions of both dynamical prosodic models (SRM and PG). To study speech rhythm at the articulatory side, we used the jaw as the basis for articulatory rhythm as proposed by Erickson [6].

Pursuing Erickson's idea of the jaw as a rhythm articulator, we conducted an experiment to examine how speech rate acts to change the phrasal prominences throughout the utterance. The novelty of the present study is to make a fine articulatory description of how speech rate variation works to change the durational patterns of articulatory gestures (defined in [4]).

Our main hypothesis is that phrasal prominences along the utterance are restructured with speech rate increase, and, as a consequence, stressed vowels under a phrasal boundary are realized with lesser jaw opening at fast rates. Besides, it is hypothesized that at fast rates there is no jaw movement reset after some minor prosodic boundaries. This hypothesis is based on Barbosa's studies [2, 3], which show that Brazilian Portuguese (henceforth BP) VV durations exponentially increases up to a phrasal boundary and then a reset of VV duration values occurs, i.e., after reaching its maximal duration, VV duration decreases and starts increasing all over again up to the next phrasal boundary.

2. Methods

A female native speaker of BP (age 28-30) was recorded acoustically (sampling rate: 22.5 kHz) and articulatorily at the USC Phonetics Laboratory. The speaker was paid for the participation in the experiment and signed an approved informed consent form explaining the purpose of the experiment. A 2-D Articulograph AG-200 (www.articulograph.de) (cf. [12], EMA magnetometer system) was used for tracking jaw movement. The movement data was sampled at 200 Hz, head-corrected, rotated to the occlusal plane, and low-pass filtered at 25 Hz. Pellets were attached to the following articulators: tongue (close to the palatal region), lower lips, jaw (at the lower incisors). Two other pellets were used as reference for the signal acquisition system: one at the nose bridge, and one at the center anterior surface of the maxillary incisors. Only the y-movement of the jaw was measured. As in Erickson [6], jaw opening was measured "in terms of the lowest vertical position of the mandibular pellet in the syllable from the maxillary occlusal plane" [6, p. c-134].

The recorded sentences are displayed in Table 1. Ten repetitions of each sentence (randomized within blocks) at three speech rates were recorded. This results in a total of 120 utterances for analysis (4 sentences x 10 repetitions x 3 speech rates).

To obtain three distinct speech rates, the subject was asked to read the sentences according to the following instructions and order: (1) normal: speak in a comfortable way; (2) slow: speak as slow as you can preserving the sentence's meaning and without introducing pauses between words; (3) fast: speak as fast as you can without introducing distortions in speech.

Table 1. *Sentences used in the experiment with their respective translation (TR) and VV phonetic transcription (PT). Bolded words represent where phrasal prominence is expected to fall (no such markings appeared in the stimuli for reading.*

Sentence 1	Ela diz mão de **máfia** no carro da moça do **papai**.
PT	[ɛl.ɐdʒ.izm.ɐ̃ʊdʒ.ɪm.af.ɪɐn.ʊk.aɤ.ʊd.am.os.ɐd.ʊp.ap.aɪ]
TR	She says mafia's hand in the car of my father's girl.
Sentence 2	Foi bão gostar demais de **papai**, mas de **máfia**!?
PT	[f.ɔɪb.ɐʊg.ost.aɤdʒ.ɪm.aɪzdʒ.ɪp.ap.aɪm.azdʒ.ɪm.af.ɪɐ]
TR	It were good to love too much my dad, but not mafia... ["were" instead of "was" implies a very informal style]
Sentence 3	Mamãe não quer mais qu'eu **babe**, mas qu'eu **pape**.
PT	[m.ɐ̃m.ɐ̃ɪn.ɐ̃ʊk.ɛɤm.aɪsk.ɪeʊb.ab.ɪm.ask.ɪeʊp.ap.ɪ]
TR	My mother doesn't want that I dribble anymore, but that I eat [child language].
Sentence 4	Vou lá levar o **pavê** pra filha do **papai**.
PT	[v.oʊl.al.ev.aɾ.ʊp.av.epɾ.af.iʎ.ɐd.ʊp.ap.aɪ]
TR	I am going there to deliver the "pavê [kind of dessert]" to dad's daughter.

For transcription and labeling, MAVIS software [14], modified at University of Southern California, was employed to measure the jaw sensor movement in the vertical dimension (y-axis). The following articulatory variables were measured (see [9, p. 139] for details): (i) jaw maximum extension: measured at y-velocity zero-crossing at maximum opening; (ii) jaw constriction displacement: measured as the difference between the zero crossings at constriction onset and extremum; (iii) jaw gesture (related to acoustic) duration: measured as the interval between maximum peak velocity[1] and maximum (modulus) valley velocity; (iv) jaw gesture duration: measured between y-velocity zero-crossings at constriction onset and maximum; (v) constriction jaw gesture peak/valley velocity (y); (vi) jaw acceleration duration: measured from the time of zero-crossing at constriction onset up to the time of constriction peak (or valley) velocity; (vii) proportional time-to-peak (or valley) velocity: measured from the ratio of y acceleration duration to total constriction formation duration.

3. Results

As objective means of automatically detecting stress group boundaries is not yet available for articulatory data, acoustic analyses were run to detect the stress groups' boundaries following the procedures presented in Meireles' papers [9, 10, 11]. Also, subsequent statistical analyses have shown that the duration of the articulatory VVs were not significantly different from the duration of the acoustic VVs.

Our results suggest that speech rate increase tends to strengthen the right-headness characteristic of BP, i.e., the greatest phrasal prominences occur to the right of the sentence at fast rates. These greatest phrasal prominences are considered here as the greater duration of a VV unit in comparison with the other VV unit. For sentence 1, with VV duration as a function of the articulatory VVs (related to the acoustic one) [afɪ] and [aɪ], the greatest phrasal prominence occurred at [afɪ] for slow ($F(1,4) = 15.858$, $p < 0.017$) and normal rates ($F(1,10) = 72.681$, $p < 0.0002$), and at [aɪ] for fast rate (n.s.). For sentence 2, with VV duration as a function of the articulatory VVs [ap] and [afɪ], the greatest phrasal prominence occurred at [ap] for slow ($F(1,8) = 85.689$, $p < 0.00003$) and normal rates ($F(1,6) = 22.611$, $p < 0.0032$), and at [afɪ] for fast rate (n.s.). For sentence 3, with VV duration as a function of the articulatory VVs [ab] and [ap], no statistical difference was found for slow rate, and for normal ($F(1,6) = 6.0367$, $p < 0.0494$) and fast rates ($F(1,12) = 15.042$, $p < 0.003$) the greatest phrasal prominence occurred at [ap]. For sentence 4, with VV duration as a function of the articulatory VVs [epɾ] and [aɪ], there was no statistical difference among rates. Nevertheless, as these one-way ANOVAs confirmed this tendency only for sentence 3, further studies are necessary to corroborate this hypothesis.

A statistical comparison of the sentences' vowels (cf. table 2) as a function of rate indicates a significant general tendency for smaller displacements (y-extremum and constriction

[1] As we are using velocity in modulus, all general references to peak velocity also applies for valley velocity. So, the hypothesis in fig. 1 for peak velocity can also be extended to valley velocity.

[2] VVs with 2 vowels mean that the second vowel was acoustically produced with a voiceless pattern and, thus, included in this syllable-like unit.

displacement) with speech rate increase. High and mid-high vowels tend to be less high, and low vowels tend to be less low from slow to fast rates (see table 3 and [9, p. 148-149] for details).

A comparison of the peak velocity (consonants) or valley velocity (vowels) as a function of rate for all gestures suggests a decreasing of maximum velocity (modulus) from slow to fast rate. Although this decreasing pattern was statistically found for some gestures within the sentences (SENTENCE 1: 2 out of 15, SENTENCE 2: 5 out of 12, SENTENCE 4: 7 out of 12), only a general pattern of decreasing maximum velocity was found for the gestures in sentence 3 ($F(2,191) = 4.7334$, $p < .0099$, using maximum velocities with their absolute values). Therefore, future studies are needed in order to support this hypothesis of gesture's maximum velocity decrease with speech rate increase.

Table 2. *Vowels (bold type) used in a one-way ANOVA with y-extremum and constriction displacement as a function of rate. Some vowels were not analyzed, because they occurred at the same jaw movement of the preceding consonant/vowel.*

Sentence 1	Ela diz mão de m**á**f**i**a no c**a**rro d**a** m**o**ç**a** do p**a**p**ai**.
Sentence 2	Foi bão gostar dem**ai**s de p**a**p**ai**, m**a**s d**e** m**á**f**i**a**!**?
Sentence 3	Mamãe não quer mais qu'**eu** b**a**b**e**, m**a**s qu'**eu** p**a**p**e**.
Sentence 4	Vou lá lev**a**r o p**a**v**ê** pr**a** f**i**lh**a** do p**a**p**ai**.

Rate effects on both articulatory and acceleration duration revealed a decreasing duration pattern from slow to fast rate (ARTICULATORY DURATION: SENTENCE 1, $F(2,183) = 13.080$, $p < 10^{-5}$, SENTENCE 2, $F(2,141) = 9.1624$, $p < .0002$, SENTENCE 3, $F(2, 177) = 52.992$, $p < 10^{-4}$, SENTENCE 4, $F(2,273) = 76.072$, $p < 10^{-4}$; ACCELERATION DURATION: SENTENCE 1, $F(2,183) = 12.549$, $p < .00002$, SENTENCE 2, $(2,141) = 12.382$, $p < .00002$, SENTENCE 3, $F(2,176) = 34.077$, $p < 10^{-5}$, SENTENCE 4, $F(2,273) = 30.905$, $p < 10^{-5}$).

Analysis of the proportional time-to-peak/valley-velocity as a function of rate indicates no common pattern for the gestures. Some gestures had a downward pattern and others an upward pattern from slow to fast rate. Thereby, stressed and unstressed vowels were grouped together. This grouping still indicates no rate effects on proportional time-to-peak/valley-velocity, but it indicates that stressed vowels have smaller proportional time-to-peak/valley velocity than unstressed vowels (SENTENCE 1, $F(1,91) = 9.6706$, $p < .003$; SENTENCE 2, $F(1,70) = 5.4489$, $p < .023$; SENTENCE 3, n.s.; SENTENCE 4, $F(1,136) = 10.051$, $p < .002$).

In summary, the data indicate a stiffness increase from slow to fast rate, and, consequently, a shrinking of the gestures and smaller spatial movements. Therefore, rhythmic restructurings are more likely to occur at fast rates, since stress groups (henceforth SG) ending at weak boundaries in slower rates may delete due to stiffness increase in fast rates. It is worth to recall that longer durations are necessary to delimit phrasal boundaries in BP.

3.1. Kinematic sources of rhythmic restructurings

Based on the articulatory results outlined above, we proposed the following kinematic sources of rhythmic restructurings with speech rate increase (cf. figure 1): a) decrease of acceleration duration; b) decrease of y-extremum; c) decrease of constriction gesture displacement; d) decrease of peak/valley velocity (modulus); e) decrease of articulatory duration; and f) constant time-to-peak/valley-velocity.

Another kinematic source of rhythmic restructuring not related to speech rate change is the diminishment of the proportional time-to-peak/valley-velocity on stressed vowels compared to unstressed ones.

According to these sources of articulatory patterning, a general decrease of articulatory parameters from slow to fast rate is expected. Figure 1a represents an articulatory shortening caused by a stiffness change (cf. [13]), which indicates a shorter acceleration duration. Figure 1b exhibits a smaller y-extremum caused by a less open jaw position (low vowels) or less closed jaw position (high mid-high vowels), which can be understood as a gesture undershoot. Closely related to figure 1b is figure 1c that represents a smaller difference of y displacement at the initial and final positions of the gesture. If a smaller y-extremum is expected, a smaller constriction displacement at fast rates is also expected. Figure 1d shows a peak velocity decrease from slow to fast rate. This hypothesis can be explained by the fact that to reach a greater peak velocity followed by a change in movement direction at zero velocity (jaw goes up and down), a greater distance is needed, which can be found at slow rates, not fast ones. Figure 1e is merely a consequence of the previous kinematic sources, i.e., diminishment of articulatory duration. Finally, figure 1f shows that the time-to-peak/valley-velocity keeps constant with speech rate increase.

Table 3. *Significance of y-extremum as a function of speech rate. This table represents vowels with their respective word, ANOVA and significance for all sentences in the corpus. *means marginally significant.*

y-extremum			
vowel	word	Anova	p <
Sentence 1			
[a]	['kafiʊ]	$F(2,9) = 6{,}2189$	0.021
[a]	[da]	$F(2,9) = 7{,}4352$	0.013
[aj]	[pa'paj]	$F(2,9) = 3{,}9786$	0.06*
Sentence 2			
[a]	[pa'paj]	$F(2,9) = 6{,}1360$	0.021
[a]	[mas]	$F(2,9) = 12{,}116$	0.003
[a]	['mafɐ]	$F(2,9) = 18{,}273$	0.0007
[ɐ]	['mafɐ]	$F(2,9) = 9{,}1656$	0.007
Sentence 3			
[a]	['babɪ]	$F(2,12) = 4{,}1435$	0.043
[ɪ]	['babɪ]	$F(2,12) = 4{,}0615$	0.045
[a]	[mas]	$F(2,12) = 22{,}454$	0.0001
[a]	['papɪ]	$F(2,12) = 4{,}1749$	0.043
[ɪ]	['papɪ]	$F(2,12) = 5{,}6012$	0.02
Sentence 4			
[a]	[le'var]	$F(2,20) = 6{,}0484$	0.009

4. Discussion

As seen in the last session, main results have shown that speech rate increase seems to change the articulatory parameters in a uniform way, disregarding their phrasal position. Despite that, there is evidence that some factors may be able to explain the rhythmic restructurings found at the articulatory level in BP. More data and speakers are needed to

confirm these trends, however. Normalization in the articulatory signals is needed in order to be able to obtain a consistent methodology to work with them. Recall that acoustic stress group boundaries were found based on normalization procedures applied to the acoustical signal only (cf. [3, 9]). Furthermore, a comparison of the PG and SRM models have led us to different conclusions on the acoustic and the articulatory side

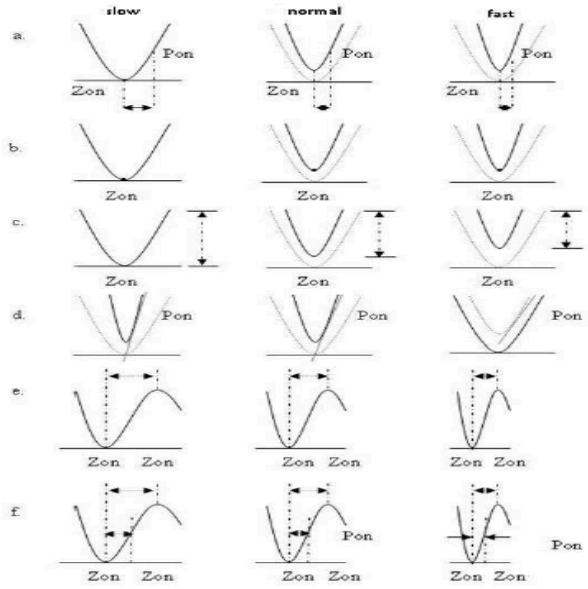


Figure 1. *Kinematic sources of rhythmic restructurings with speech rate increase: a. diminishment of acceleration duration (stiffness increase); b. y-extremum decrease; c. constriction gesture displacement decrease; d. peak velocity decrease; e. articulatory duration decrease; f. constant proportional time-to-peak velocity.*

On the acoustic side, according to Barbosa [1, 2, 3] and Meireles's acoustical data [9, 10, 11], since in BP the greater duration at the end of a SG is a consequence of a exponential increase from the beginning of this SG, SRM model better explains the rhythmic variations found in this paper. Recall that PG model only explains longer durations at (or near) a phrasal boundary.

Yet, on the articulatory side, since we only found some possible influence of articulatory parameters exactly at the places where rhythmic restructurings occurred, both models would work to explain the results found up to this point. However, we remind that because PG researchers have specifically worked with articulatory data, they are more advanced at the methodological side to work with articulatory gestures influenced by prosodic structure. Despite that, we have showed here some evidence that SRM model could perfectly deal with such data once more information about BP gestures is provided and new methods applied to articulatory data are designed.

5. Conclusions

The results of this articulatory study suggest that speech rate increase affects all gestures in a sentence disregarding their phrasal position. Nevertheless, future work is needed since acceleration duration, constriction displacement and articulatory duration seem, though not conclusive, to reflect the rhythmic restructurings found on the articulatory side. Also, results showed that the SRM model is more appropriate to deal with BP acoustical data than the PG model, and that both models could explain the articulatory variations due to speech rate increase.

Finally, this paper's main results have shown that rhythmic structure variation is modified gradually with speech rate increase, i.e., quantitative aspects of speech are acting to modify speech rhythm, showing how a dynamical systems approach to language perfectly suits to linguistic descriptions

6. Acknowledgements

This work was supported by FAPES (54652499/2011), and NIH (DC03172, D. Byrd). The authors thank Elliot Saltzman, Louis Goldstein, and Hosung Nam for their insightful comments on this research. Many thanks are due to USC Phonetics Laboratory, Haskins Laboratories, Dani Byrd and Sungbok Lee, for help with the articulatory design and support, and James Mah, for help in the articulatory recordings. Also, we would like to thank the anonymous reviewers for the remarkable comments.

7. References

[1] Barbosa, P. A., "Explaining Cross-Linguistic Rhythmic Variability via a Coupled-Oscillator Model of Rhythm Production", Proc. Speech Prosody 2002 Conf. [CD], Aix-en-Provence, 163–166, 2002.

[2] Barbosa, P. A., Incursões em torno do ritmo da fala, Campinas, Brazil: Pontes/Fapesp, 2006.

[3] Barbosa, P. A., "From syntax to acoustic duration: a dynamical model of speech rhythm production", Speech Communication, 49:725–742, 2007.

[4] Browman, C. and Goldstein, L., Articulatory gestures as phonological units. Phonology, v. 6, p. 201–251, 1989.

[5] Byrd, D. and Saltzman, E., The elastic phrase: modeling the dynamics of boundary-adjacent lengthening. Journal of Phonetics, v. 31, p. 149-180, 2003.

[6] Erickson, D., On phrasal organization and jaw opening. In: From Sound to Sense: 50+ Years of Discoveries in Speech Communication, MIT, June 11-13. [S.l.: s.n.], 2004. p. S15.

[7] Kelso, J. A. S., Dynamic patterns: the self-organization of brain and behavior. Cambridge, USA: MIT Press, 1995.

[8] Kelso, J. A. S., Saltzman, E. L. and Tuller, B., The dynamical perspective on speech production: data and theory. Journal of Phonetics, v. 14, p. 29–59, 1986.

[9] Meireles, A. R., Self-organizing rhythms in Brazilian Portuguese: speech rate as a system perturbation. Germany: VDM Verlag, 2009.

[10] Meireles, A. R. and Barbosa, P. A., Speech rate effects on speech rhythm. In: Speech Prosody 2008 Conference, 2008, Campinas. Proceedings of the Speech Prosody 2008 Conference. Campinas: RG. v.1. p.327 – 330, 2008.

[11] Meireles, A. R., Tozetti, J. P. and Borges, R. R., Speech rate and rhythmic variation in Brazilian Portuguese. In: Speech Prosody 2010 Conference, 2010, Chicago. Proceedings of the Speech Prosody 2010 Conference. Chicago: RG. v.1. p.1 – 4, 2010.

[12] Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I. and Jackson, M., Eletromagnetic midsagittal articulometer (emma) systems for transducing speech articulatory movements. J. Acoust. Soc. Am., p. 3078–3096, 1992.

[13] Saltzman, E. and Munhall, K. G., A dynamical approach to gestural patterningin speech production. Ecological Psychology, v. 1, n. 4, p. 333–382, 1989.

[14] Tiede, M. K., Vatikiotis-Bateson, E., Hoole, P. and Yehia, H., Magnetometer data acquisition and analysis software for speech production research. ATR Technical Report TR-H 1999. Kyoto, Japan:ATR Human Information Processing Labs, 1999.

Prosody in Turkish learners of German as a Foreign Language

Sabine Zerbian [1], *Jane Kühn* [2], *Christoph Schroeder* [2], *Svenja Schuermann* [2]

[1] Institute of Linguistics: English, University of Stuttgart, Germany
[2] SFB 632 "Information Structure", University of Potsdam, Germany
sabine.zerbian@ifla.uni-stuttgart.de, {jkuehn,schroedc}@uni-potsdam.de

Abstract

Results of a pilot study are presented which investigates the prosodic realization of information structure by six learners of German as a Foreign Language (GFL) with Turkish as first language. Question-answer pairs were read out loud, which systematically varied the position of narrow focus in the response by means of a preceding *wh*-question.
A qualitative analysis of the results shows deaccentuation of postfocal constituents in the case of subject focus for 4/6 GFL-speakers but no consistent pitch increase on focused constituents. Two speakers did not change prosody due to information structure.
The results are discussed in connection with the acquisition of prosody as a marker of information structure. Deaccentuation has been reported to cause problems in L2 prosody. In Turkish, deaccentuation occurs postfocally. The claim will be motivated that the occurrence of deaccentuation in the L1 is a necessary but not sufficient condition for early acquisition of deaccentuation in a foreign language.

Index Terms: L2 prosody, German, Turkish, focus, deaccentuation, production

1. Introduction

The acquisition of L2 prosody, i.e. linguistically relevant changes of fundamental frequency (F0), intensity and/or duration over the course of an utterance, has received comparatively limited attention in the literature on Second or Foreign Language Acquisition. According to [1:57] and referring back to the intonational typology in [2], phonological influences in L2 prosody must be differentiated from phonetic influences. Phonological influences on L2 prosody result from differences in the inventory of phonological tunes, their form, and in the meanings assigned to the tunes. Phonetic influences result from a difference in the phonetic realization of an identical phonological tune.
Previous studies on L2 of non-related languages have shown that the target-like placement of sentence accent (phonological level) as well as the target-like realization of pitch accents (phonetic level) cause difficulties in foreign language acquisition. One of the particular difficult features reported in studies like e.g. [3-7] lies in the function of prosody in languages like English and German, where sentence-level prosody indicates information structural notions such as topic, focus and givenness. For L2 prosody, a non-target-like reaccentuation of constituents that are given through the preceding discourse has repeatedly been reported in the studies cited above.
The article reports the results of a pilot study into the prosody of Turkish learners of German as a foreign language. The functional use of prosody as marking focused and given constituents will be central. The German-Turkish language pairing is interesting, because German, just like English, shows both prosodic focus marking through an increase in fundamental frequency (F0) and duration, as well as deaccentuation of given constituents. For English, prosodic focus marking and deaccentuation have been considered "opposite sides of the same coin" [8:67]. Turkish is not related to German. However, it can be considered to have a similar prosodic representation, being a stress language which also uses prosody at the sentence level. However, it shows less systematic prosodic focus marking (see discussion in section 2.2) but has postfocal deaccentuation[1].
The article is structured as follows: Section 2 gives the relevant background on the prosody of the two languages involved. Section 3 presents the experiment with a summary of the results in section 3.5. Section 4 provides a discussion.

2. Focus prosody in German and Turkish

2.1. German

German has been classified as an intonation-only language (cf. [9]) which has lexical stress. Postlexically, it uses different types of pitch accents together with boundary tones in order to express pragmatic contrasts (see [10]). The interaction of information structure and intonation is uncontroversial for German.
Default sentence accent (in all-new sentences) is assigned on the basis of syntactic structure: every argument of the verb is accented, and the verb might also be accented if there is phrasal integration of the verb and its immediately preceding argument. Perceptually, the last accent in a sentence is perceived to be the most prominent. In terms of phonetic implementation of the accents, downstepping of high-toned accents has been reported to occur, i.e. realization of a high tone at a lower phonetic level relative to a preceding high tone [11].
Under narrow focus, the most prominent accent occurs on the focused constituent. As experimental work by [11] for German confirms, pitch is raised under focus when compared to a baseline of all-new sentences. Givenness leads to a lowering of pitch, whereby the position of the given constituent with respect to the focus is relevant. If the given constituent occurs preceding the focus of the sentence (prenuclear position), the constituent will have a pitch accent with a comparatively low pitch. If the given constituent occurs following the focus of the sentence (postfocal position), it has been said to be deaccented. This overall pattern was confirmed by German native speaker controls, participating in the same experimental task as reported in the current article.

[1] The difference between deaccentuation and postfocal compression is not always clear-cut in actual data. The former refers to the deletion of a pitch accent whereas the latter refers to the compression of the pitch range of a pitch accent. Here, the term deaccentuation is used, bearing in mind that this issue requires further investigation.

Next to F0, increased duration on the focused constituent is a further cue to focus in German with given constituents only being shortened in prefocal position [12].

2.2. Turkish

Turkish is an SOV-language. Though differing in details of analysis, scholars agree that the default sentence accent occurs on the last argument before the verb, which is normally the object [13, 14]. In addition, an accent can occur on every argument.

As for the expression of focus, [15] differentiates between a syntactic strategy in which the preverbal position is used for information focus (e.g. question-answer pairs) where the answers is not accessible from context. In cases of e.g. correction focus where answer can be chosen from an explicit set of alternatives, a constituent is marked as focused in situ by means of a pitch accent. Along the same lines, [16] describes that for both contrastive and non-contrastive emphasis prosodic means such as strong stress and high pitch are used. However, in addition to these prosodic means, non-contrastive emphasis is further marked by placing the focused constituent in the immediately preverbal position. [17], on the other hand, do not assign the immediately preverbal position a primary role in focus assignment. In their examples, focused constituents always bear stress and can occur in any preverbal position.

To our knowledge, only one acoustic study on focus in Turkish exists. [18], in using a comparable methodology as outlined in the present paper, finds that there is no increase in F0 on the focused constituent in the Turkish data she collected. She found rather unsystematic significant differences for the other acoustic correlates: A sentence-initial focused subject differs acoustically from its neutral counterpart in terms of longer duration, a sentence-final verb differs in being higher in intensity. An object did not show differences between the focused condition and the neutral counterpart in any of the three acoustic measures F0, intensity and duration. Unexpectedly, an F0 increase on the preceding object is found when the final verb is in focus. For givenness marking, [18] found that in cases of subject focus there is a significant F0 drop in the post-focus domain. The same pattern does not hold for object focus though. As tested by means of a perception experiment, subject focus has the highest recognition rate, probably due to the salient F0 drop.

It is uncontroversial that in Turkish sentence accent never occurs in the postverbal domain. The prosody in the postverbal domain has been described as being without "any indicator of prosodic structure" [19:144]. [19: fig. 3] provide a pitch track showing that we can expect to find deaccentuation postfocally, not only postverbally. In their example, the pitch contour remains entirely flat following the constituent in focus, which in this specific case is a sentence-initial subject. Their observation is thus in line with the results on the F0 drop following a focused subject in [18].

2.3. Hypotheses

The research question is whether Turkish learners of German change prosody depending on information structure and if so, by means of which acoustic cues. More precisely:

- Do Turkish learners of German use increased F0 on the focused constituent? Both languages seem to differ on the acoustic cues used for prosodic focus marking: Whereas German relies on F0 as well as on duration, Turkish does not seem to rely on F0 as an acoustic cue to prosodic focus.
- Do Turkish learners of German use deaccentuation to signal givenness? Although both languages use deaccentuation for postfocal constituents, research on the acquisition of L2 prosody has reported that learners have difficulties with this feature ([3-7]).

3. Experiment

3.1. Task

An elicited-production study was conducted using the same methodology as in [20] and [18]. German target sentences were elicited that differed in the focused constituent only. The target sentences consisted of simple SVO-structures, and the focused constituent was either the subject, the verb or the object. As a baseline condition, a broad focus rendering was elicited first.

The target sentences were presented in writing, together with a preceding question unambiguously eliciting the desired focus structure, and accompanied by a picture illustrating the action. In addition, the focused target words were underlined in order to reduce errors (cf. [20]). This experimental setup has proven in various studies to be successful at eliciting the predicted focus structures (cf. [20] and subsequent studies).

3.2. Target sentences

Target sentences were constructed in such a way as to systematically control a number of factors and to be comparable across the five different target sentences. To this end, all sentences displayed the same number of words. To minimize segmental variation, the target words showed the same stress pattern (initial stress), the same phonological length in the stressed vowel, and the same segmental make-up (CV with sonorant consonants wherever possible). The object nouns were disyllabic to allow for pitch accents and boundary tones to be realized without tonal crowding.

In constructing five different target sentences as shown in (1), we opted for type repetitions instead of token repetitions in order to increase the diversity of the targets.

(1) Target sentences
   a. Lena malt ein Lama.
      PROP.NAME paint ART llama
      'Lena is painting a llama.'
   b. Nele wohnt in Meißen
      PROP.NAME live PREP PROP.NAME
      'Nele lives in Meißen.'
   c. Nina webt das Leinen.
      PROP.NAME weave ART linen
      'Nina is weaving the linen.'
   d. Heiner baut die Mühle.
      PROP.NAME build ART mill
      'Heiner builds the mill.'
   e. Maya liest die Bücher.
      PROP.NAME read ART books
      'Maya is reading the books.'

3.3. Participants

Nine Turkish learners of German participated in the study. Participants were students of Translation Studies in German at Ege University in Izmir, where the recording took place in

September 2012. The data were collected by the second author, communicating with the participants in German. The data of one speaker (speaker 1) was not recorded properly due to technical problems, the data of two further speakers (sapeakers 2 and 7) had to be excluded because only two target sentences were recorded.
Of the remaining 6 speakers, 5 were female and one was male. The participants were between 20 and 25 years old. They all spoke Turkish as an L1, had good to very good knowledge of English (self-reported) and had studied German as a Foreign Language (GFL) for 3 to 6 years, including a preparatory language course at the University which has been designed to lead to level B1/B2 (Common European Framework of Reference for Language Teaching (CEFR)).

3.4. Analysis

The analysis concentrates on fundamental frequency (F0) as the main correlate of sentence prosody, and more specifically prosodic focus marking and deaccentuation in the target language German. All target sentences were segmented into syllables and 10 measurements of F0 were taken for each syllable, using ProsodyPro [21]. Data were checked manually for spurious pitch values. The datapoints were averaged across the five utterances of the same focus condition for each speaker, and are presented in figure 1 (see next page) for each speaker individually (in the same pitch range of 50-250 Hz, maintaining individual differences in F0 expansion).
In addition, a ToBI annotation of the averaged pitch contours is provided in table 1. Tone labels were assigned following these conventions: H* for each discernible pitch peak, H*+L for a pitch peak with a clear fall on the same constituent, !H* if the pitch peak is lower than the preceding pitch peak, ↑H* for a pitch peak that is higher than the preceding pitch peak. In some cases (e.g. speaker 9), labels had to be assigned based on deviations from expected pitch transitions if there had been no tone target. No evidence was found for boundary tones at intermediate phrases. However, the ToBI annotation should be seen as a phonetic rather than phonemic annotation of the produced intonation, given that further research needs to be done for the phonemic status of the pitch accents transcribed.

Table 1: ToBI annotation based on the averaged pitch tracks

speaker	focus	S	V	O
3	broad	H*+L	!H*+L	↑H*+L L%
	subject	H*+L		L%
	verb		H*	↑H*+L L%
	object	H*+L	!H*	↑H*+L L%
4	broad	H*+L	↑H*	L%
	subject	H*+L		L%
	verb	H*+L	↑H*	L%
	object	H*+L	↑H*	L%
5	broad		H*+L	!H* L%
	subject	H*+L		L%
	verb		H*+L	L%
	object		H*+L	!H*+L L%
6	broad		H*+L	!H* L%
	subject		H*+L	!H* L%
	verb		H*+L	H* L%
	object		H*	H* L%
8	broad		L+H*	!H* L%
	subject		L+H*	!H* L%
	verb		L+H*	!H* L%
	object		L+H*	!H* L%
9	broad	L*H	!H*	!H*+L H%
	subject	L*H		!H*+L H%
	verb	L*H	L+H*	LH%
	object	L*H	H*	!H*+L H%

3.5. Results

The observations are given for each speaker individually first, before they are summarized (number refers to speaker):

- 3: on-focus increase in F0, deaccentuation only in case of subject focus
- 4: no on-focus increase in F0, deaccentuation only in subject focus
- 5: on-focus increase in F0, deaccentuation in subject and verb focus
- 6: no on-focus increase in F0, despite a lower F0 in subject focus, the F0 patterns are qualitatively the same across all focus conditions
- 8: identical F0 contours across all focus conditions
- 9: on-focus increase in F0 in verb focus only, deaccentuation only in subject focus

All speakers realize intonation in broad focus context similar to object focus. Beyond that, as can be expected in a group of learners, we find considerable heterogeneity in the data. The speakers can be described to fall into roughly two groups: speakers 6 and possibly 8 do not show any change in the F0 pattern dependent on the focus structures of the sentences uttered. The other speakers do show a change in prosody related to information structure, as detailed above.
To sum up, of the speakers investigated, most (though not all) use some kind of prosody in connection with information structure (cf. discussion section). Deaccentuation is the most reliable cue used by four out of six speakers, but only in the case of subject focus. Half the speakers also use an increase in F0 on focused constituents. It needs to be noted though, that only speaker 3 links the overall F0 pattern to information structure in a way comparable to the target language.
Given that [18] reported duration and intensity as cues to focus in Turkish rather than F0, these parameters were checked quantitatively for those two speakers who did not show any F0 changes due to information structure (speakers 6 and 8). Only for speaker 6 we find higher duration and intensity on a focused constituent, when compared to the broad focus condition, although the differences are not significant.

4. Discussion

The production data show that some Turkish learners of German do not change prosody according to information structure, neither using phonetic cues of the L1 like intensity and duration [18], nor phonetic cues of the L2 such as F0 increase on focused constituents nor phonological features shared by L1 and L2 such as deaccentuation on given constituents.
Others learners, however, do show a change of prosody due to information structure. Interestingly, these speakers have in common that they show deaccentuation in the case of subject focus. Postfocal deaccentuation in the case of subject focus is a prosodic feature shared by the L1 Turkish and the L2 German and it is a very prominent cue to a shift of focus to the subject ([18] for perception results, auditory impression suggests the same for our data). Some of those speakers also show an increase of F0 on the focused constituent when compared to the broad focus condition and/or deaccentuation on given constituents other than the subject.
Two results thus need further discussion: First, the observation that some learners produce postfocal deaccentuation in cases

of subject focus, although deaccentuation has been reported as a difficult feature to be acquired in foreign language acquisition. Second, the observation that some learners do not change prosody at all under the influence of information structure.

The studies on L2 prosody which constitute the claim that deaccentuation is a difficult feature, investigate language pairings with English as L2 and L1s which either show equivalents to deaccentuation in English but whose prosodic systems are typologically very different (e.g. English, Korean and Japanese in [4] or English, Mandarin and Korean in [7]) or which do not have deaccentuation themselves [5, 22, 23]. [24] have argued that phonological representation is crucial in foreign language acquisition in that representational differences between two languages cause difficulties. If transferred to the suprasegmental domain, Korean and Japanese learners of English are predicted to have difficulties with English prosody due to the representational differences between the two prosodic systems involved, independent of any specific function of the target prosody.

[5, 22, 23] investigate the use of prosody by Spanish learners of English. Spanish and English can both be classified as stress and intonation-only languages and thus share a similar prosodic representation. Postfocal deaccentuation was still found to be difficult for these learners. Here we need to note that L1 Spanish is said to not use deaccentuation (e.g. [25]).

Prosody used for information structure is a typologically marked feature [26]. Following the Markedness Differential Hypothesis [27], this makes it a difficult feature to be acquired, which has been confirmed by the studies on Spanish learners of English. However, in the case of Turkish learners of German, the L1 shares the marked feature as discussed in section 2.2. This might be the reason why even learners at the level of B1/B2 show the use of this feature. The lack of prosodic marking of focused constituent by a gradient increase of F0 on these constituents can be interpreted along the same lines. Prosody used for information structure is a marked feature whose acquisition is difficult, especially if the feature is not shared with the L1. As Turkish does not use F0 increase on focused constituents, this new feature has been acquired by very few learners only.

It can thus be argued that the presence of a marked feature in the L1, such as postfocal deaccentuation, is a necessary condition for the acquisition of that feature in the L2. The results of our production study equally show that it is by no means a sufficient condition. Two speakers do not show any changes in prosody due to information structure, not even in those acoustic correlates which are allegedly used in Turkish (intensity and duration according to [18]). How can the lack of any prosody in speech production be accounted for? We follow [28] and [7], who observe that under certain circumstances learners seem to fall back on some default intonation patterns. [28] investigated the prosody of advanced German learners of English and speculated that observable errors in the placement of pitch accents give "room for the assumption that non-native speakers develop a consistently simplified assignment of tonal categories, a kind of default accent" [28: 87]. This would be in line with [29]'s "basic variety" and a corresponding reduced prosodic realization of the target language. In some of her experimental tasks, [7] observes that Mandarin and Korean learners of English rely on common prosodic patterns used for broad focus.

We thus interpret the results of our pilot study as evidence for the following developmental path in the acquisition of IS-related intonation by Turkish learners of German:

1. Default intonation: One all-purpose default pattern, not differentiating e.g. different kinds of narrow focus.
2. Transfer of L1 feature: Transfer leads to postfocal deaccentuation with subject focus. This is possible because L1 Turkish shares this marked feature with German.
3. Acquisition of a new phonetic feature, namely the gradient increase of F0 on the focused constituent, shown by only very few learners of our group.

Further research is clearly needed though in order to confirm the results.

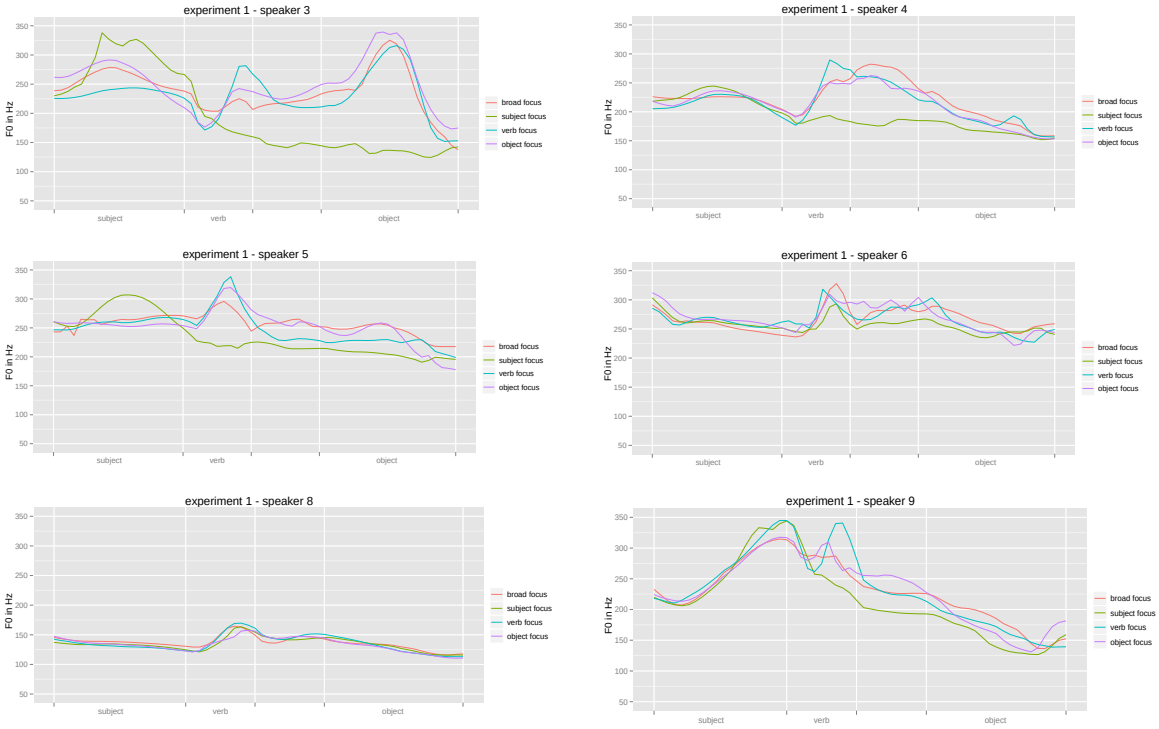


Fig. 1: averaged F0 contours for the four focus conditions for each speaker individually (pitch range of 50-250 Hz)

5. References

[1] Mennen, I., "Phonological and phonetic influences in non-native intonation", in J. Trouvain & U. Gut [Eds], Non-native Prosody: Phonetic Descriptions and Teaching Practice, 53-76, Mouton de Gruyter, 2007.

[2] Ladd, D.R., "Intonational Phonology", Cambridge University Press, 1996.

[3] McGory, J., "The Acquisition of Intonation Patterns in English by Native Speakers of Korean and Mandarin". PhD dissertation, Ohio State University, Columbus, 1997.

[4] Ueyama, M., Jun, S.-A., "Focus Realization in Japanese English and Korean English Intonation", Japanese and Korean Linguistics 7:629–645, 1998.

[5] Verdugo, D.R., "Prosodic realization of focus in the discourse of Spanish learners and English native speakers", World Englishes 14, 9–32, 2006.

[6] Gut, U., "Non-native Speech. A Corpus-based Analysis of Phonological and Phonetic Properties of L2 English and German", Peter Lang, 2009.

[7] Baker, R.E., "The Acquisition of English Focus Marking by Non-Native Speakers", PhD dissertation, Northwestern University, Evanston, Illinois, 2010.

[8] Ladd, D.R., "The Structure of Intonational Meaning - Evidence from English", Indiana University Press, 1980.

[9] Gussenhoven, C., "The Phonology of Tone and Intonation", Cambridge University Press, 2004.

[10] Féry, C., "German Intonational Patterns", Max Niemeyer, 1993.

[11] Fery, C., Kügler, F., "Pitch accent scaling on given, new and focused constituents in German", Journal of Phonetics 36:680–703, 2008.

[12] Kügler, F., "The role of duration as a phonetic correlate of focus", in P. Barbosa, S. Madureira & C. Reis [Eds], Proceedings of Speech Prosody. Editora RG/CNPq, 591–594, 2008.

[13] Kan, S., "Prosodic Domains and the Syntax-Prosody Mapping in Turkish", Master's Thesis, Bogazici University, Bogazici, 2009.

[14] Kamali, B., "Topics at the PF Interface of Turkish", PhD dissertation, Harvard University, Cambridge, Massachusetts, 2011.

[15] Issever, S., "Information structure in Turkish - the word order-prosody interface", Lingua 113:1025–1053, 2003.

[16] Kornfilt, J., "Turkish Grammar", Routledge, 1997.

[17] Göksel, A., Özsoy, A.S., "Is there a focus position in Turkish?", in A. Göksel & C. Kerslake [Eds], Studies on Turkish and Turkic Languages. Proceedings of the Ninth International Conference on Turkish Linguistics, Harrassowitz, 219–228, 2000.

[18] Ipek, C., "Phonetic realization of focus with no on-focus pitch range expansion in Turkish", in Proceedings of the 17th International Phonetic Congress (ICPhS), 140–143, 2011.

[19] Özge, U, Bozsahin, C., "Intonation in the grammar of Turkish", Lingua 120:132–175, 2010.

[20] Xu, Y., "Effects of tone and focus on the formation and alignment of f0 contours", Journal of Phonetics 27:55–105, 1999.

[21] Xu, Y., "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis", in Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013). Aix-en-Provence, 7–10, 2013.

[22] Nava, E., "Prosody in L2 acquisition", in 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007). Cascadilla Proceedings Project, 2007.

[23] Nava, E.; Zubizaretta, M.L, "Deconstructing the Nuclear Stress Algorithm: Evidence from second language speech", in N. Erteschik-Shir & L. Rochman [Eds], The Sound Patterns of Syntax. Oxford University Press, 2010.

[24] Eckman, F.R., Elreyes, A., Iverson, G.K. "Some principles of second language phonology", Second Language Research 19 (3): 169–208, 2003.

[25] Cruttenden, A. "The de-accenting of given information. A cognitive universal?", in G. Bernini & M.L. Schwartz [Eds], Pragmatic Organization of Discourse in the Languages of Europe. Mouton de Gruyter, 311–355, 2006.

[26] Rasier, L., Hiligsmann, P. "Prosodic transfer from L1 to L2. Theoretical and methodological issues", Nouveaux cahiers de linguistique francaise 28: 41–66, 2007.

[27] Eckman, F. "Markedness and the Contrastive Analysis Hypothesis", Language Learning 27 (2): 315–330, 1977.

[28] Jilka, M., "The contribution of intonation to the perception of foreign accent. Identifying intonational deviations by means of F0 generation and resynthesis", PhD thesis, University of Stuttgart, 2000.

[29] Klein, W., Perdue, C. "The Basic Variety (or: Couldn't natural languages be much simpler?)", Second Language Research 13 (4): 301–347, 1997.

Synthesizing sports commentaries: One or several emphatic stresses?

Sandrine Brognaux[1,2,3] *Thomas Drugman*[3]*, Marco Saerens*[2]

[1]Cental, [2] ICTEAM, Université catholique de Louvain, Belgium
[3]TCTS Lab, Université de Mons, Belgium

sandrine.brognaux@uclouvain.be, thomas.drugman@umons.ac.be, marco.saerens@uclouvain.be

Abstract

Emphatic stresses are known to fulfill essential functions in expressive speech. Their integration in speech synthesis usually relies on a prosodic annotation of the training corpus. Emphasized syllables are then assigned a single label or can receive several labels according to their acoustic realization. While it is more complex to predict those various labels for a new text to synthesize, it might allow for a better rendering of the stress in the synthesized speech. This paper examines whether the use of more than one emphatic label improves the perceived expressivity of the synthesized speech. It relies on a manually-annotated expressive corpus of sports commentaries. Statistical acoustic analyses show that four distinct realizations of emphatic stresses can be distinguished. However, perceptual tests indicate that the integration of this distinction in HMM-based speech synthesis does not lead to a significant improvement in expressivity. This seems to imply that the different acoustic realizations of the stress are not required to be explicitly annotated in the training corpus.

Index Terms: Emphasis, Emphatic stress, Expressive speech, HMM-based speech synthesis, Prosody.

1. Introduction

Recent research in speech synthesis has been targeting the generation of expressive speech [1, 2, 3]. Strategies have been proposed to modify both voice quality and prosody so as to produce the most natural quality of expressivity. Albeit rather unfrequent in neutral speech, emphasis plays a crucial role in the prosody of expressive speech [4, 5]. This phenomenon relates to highly prominent syllables which are sometimes seen as (particularly prominent) 'pitch accents' [6, 7]. Emphatic stresses are known to fulfill various functions like contrasting or highlighting elements and contribute to the liveliness of the message [4]. Their generation in expressive speech is therefore essential. It is even more the case when synthesizing sports commentaries which have been shown to display high rates of emphatic stresses, falling on specific positions like numbers in scores [8, 9].

Several attempts to integrate emphasis have been proposed, both in unit-selection [10, 4, 5, 11] and HMM-based speech synthesis [7, 12]. They usually rely on a prosodic annotation of the corpus in terms of emphatic stresses. The acoustic characteristics of the corresponding syllables are then learned to be reproduced at synthesis stage. While some annotations present various labels associated with different acoustic realizations of the emphatic stress (like ToBI [13]), most studies only use a single label for emphasis [12, 14].

The obvious advantage of using one single label is that it makes it easier to predict it for a new sentence. Most studies investigating the automatic prediction of emphasis from text have, for that matter, considered only one single emphatic label [5, 6, 14]. Conversely, predicting several emphatic labels from text requires a correlation between the labels and specific distinct functions, which is rarely the case.

However, emphatic stress is often regarded as a gathering of different kinds of stresses with various functions, positions, and, importantly, different acoustic realizations (as proposed in ToBI [13]). While some studies have mentioned the potential existence of different *levels* of emphasis [4, 15], we rather believe that different *kinds* of emphasis may co-exist, with no specific order relation between them. If one label is associated with each type of acoustic realization, it allows the training of more acoustically-consistent models, more inclined to generate suitable emphatic stresses in speech synthesis.

The question that still remains is whether the use of a single emphatic label in HMM-based speech synthesis still allows for an appropriate rendering of the various acoustic realizations. In other words, it should be assessed whether the acoustically-different emphatic stresses are learned by the models, based on the linguistic context, or if the distinction requires to be made explicit by annotating with distinct emphatic labels.

This study investigates the benefit of annotating emphatic stresses for expressive HMM-based speech synthesis with several labels instead of one. For that matter, it relies on a large corpus of sports commentaries. The corpus is spontaneous while containing a natural variety of prosody, conversely to studies relying on artificially-produced emphatic stresses by actors (see e.g. [5]). Besides, sports commentaries are characterized by a high density of emphatic stresses with strong acoustic correlates [8], which makes them much more suitable for the study of emphasis than rather neutral read speech as used in [12]. Emphatic stresses were manually annotated in the corpus and were statistically analyzed in order to define several sets of labels, corresponding to stresses with distinct acoustic realizations. The manual annotation having been realized on a functional basis, our study partly answers Hirst's critics [16], i.e. the fact that prosodic function and form tend to be merged in prosody annotation. The objective is here to distinguish between various forms of a single emphatic function and assess the resulting improvement reached in the expressivity of the synthesized speech.

The paper is organized as follows. Section 2 presents the corpus and its emphatic annotation. The statistical acoustic analysis of the emphatic stresses is further described in Section 3. The integration of different sets of emphatic labels in HMM-based synthesis is investigated in Section 4 through a perceptual evaluation. Finally, Section 5 concludes the paper.

2. Corpus design

This study is based on a corpus of live commentaries of two basketball games, uttered by a professional French commentator and recorded in sound-proof conditions. The speaker watched the game and commented it without any prompt. The issue with sports commentaries corpora is usually the high level of background noise which precludes their precise acoustic analysis [17]. Conversely, our corpus exhibits the advantage of being spontaneous and of high acoustic quality, being therefore suited for speech synthesis. Both matches star the Spirou Belgian team with very tight final scores, which induces a high level of excitation. The corpus lasts 162 minutes, silences included.

The corpus was orthographically and phonetically transcribed with [18], with manual check. The phonetic transcription was aligned with the sound using Train&Align [19] and other linguistic information (syllables, parts of speech, etc.) was generated by Elite [20]. Manual annotation of emphatic stresses was realized by assigning a 'F' label to syllables for which an emphatic function was perceived. The annotation results from two or three listenings of each sequence of 4-5 words and was submitted to a second check by the same annotator. In total, 803 syllables were annotated 'F'. Twenty percents of the corpus were annotated by a second expert, and rather high kappa scores were reached (see [8] for the complete analysis of the prosodic annotation).

3. Statistical analysis of emphatic stresses

The statistical acoustic analysis of the emphatic stresses in the corpus consists in four steps. First, a set of acoustic features is extracted for each emphasized syllable (Section 3.1). Dimensionality reduction techniques are then used to minimize the set of features and delete potential redundancy (Section 3.2). The reduced feature set is then used to cluster the emphasized syllables, as an attempt to find the more suitable number of distinct emphatic stresses (Section 3.3). These new sets of stresses are then investigated for potential correlations with specific linguistic contexts (Section 3.4). For further information about the exploited statistical methods, see [21, 22].

3.1. Extraction of acoustic features

For each emphasized syllable, 65 acoustic values are extracted. The first features consist in prosodic measurements: F0 extracted with SRH [23] (mean, max, etc.), energy (mean, max, etc.) and duration (both of syllable and nucleus). A prominence value is added by PromGrad [15] which assigns a prominence score from 0 to 4 to each syllable, on an acoustic basis. Two additional features indicate the presence of a preceding or following silence and its duration. Finally, contextual information, i.e. comparisons with the acoustic values of the two previous and the next syllable, are also computed as they were shown to be efficient for prominence detection in French [24, 15]. These latter measurements are only extracted if both syllables are not separated by a silence, as it is known that silences tend to be associated with a resetting of the prosodic parameters.

It should be noted that, as in [25], duration values are normalized with respect to the average and standard deviation of the duration of the corresponding phonemes. This choice relies on the fact that the nature of the phoneme clearly affects its duration [26, 27]. Missing values (for contextual information) are replaced by the average value of the feature. Finally, all variables are normalized into standard scores.

3.2. Dimensionality reduction

The second stage of our analysis aims at reducing the number of features. For that matter, a principal component analysis (PCA) is carried out on the data. The scree plot (see Figure 1) shows the contribution of the components to the global variance. Since there is no universal technique for selecting the natural number of dimensions, we relied on two popular rules of thumb: (i) keeping the dimensionality accounting for 70% of total variance and (ii) removing the dimensions for which the contribution to the global variance remains stationary. Based on these considerations, we chose to keep ten and four dimensions.

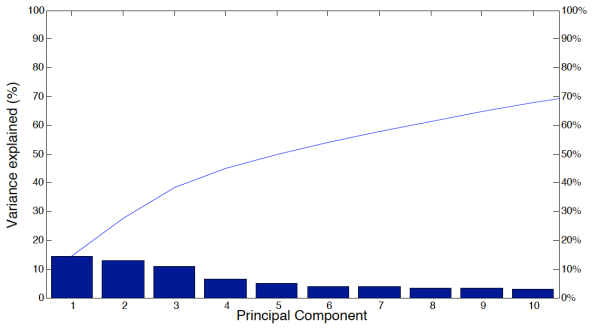


Figure 1: *Scree plot displaying the proportions of global variance carried out by the ten first components of the PCA.*

An insightful analysis regards the weights assigned to the different variables for each component. Interestingly, the first component is clearly related to energy, all 20 higher weights being assigned to energy values. The second component is linked to F0 with the 18 first weights corresponding to F0-based features. Finally, duration-based measurements have most of the heaviest weights in the third PCA component. The fourth dimension is a mix of different types of variables (esp. F0 and energy). The three higher weights for the first three components are assigned respectively to mean energy, mean F0 and syllable duration.

3.3. Clustering

A clustering is now carried out on the reduced data obtained by PCA. The main objective is to define different sets of emphatic stresses characterized by distinct acoustic values.

We first apply a Ward dendrogram [28]. The advantage of this clustering technique it that it visually shows the gathering of the various clusters, which helps in determining the natural number of clusters. Figure 2 shows the dendrogram obtained on the first four PCA components of our data. The algorithm assigns a unique color to each group of nodes where the linkage is less than a specific fixed threshold. This dendrogram clearly shows 4 distinct groups of syllables. Interestingly, when applied to the first 10 PCA components, the dendrogram also points at four distinct clusters.

A second advantage of first applying a dendrogram algorithm is that the centroids of the generated clusters can then be exploited for the initialization of a K-means clustering, which is done in this second stage of our analysis. To assess the quality of the clustering obtained when using various numbers of clusters, we also compute K-means clusterings with 2 to 10 clusters. For that matter, initialization points are selected randomly and the algorithm is run 50 times, the best clustering being kept for

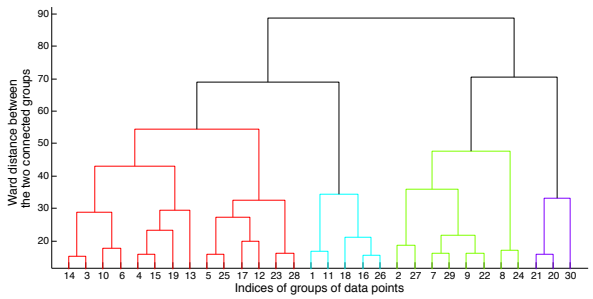


Figure 2: *Ward dendrogram applied to the first four PCA components of our data (the leaves in the plot correspond to more than one data point).*

analysis.

The silhouette value provides an evaluation of the clustering quality, the higher the value, the better the distinction between the various clusters [29]. Figure 3 interestingly shows that, even when launching a K-means with random initialization on the 4 PCA components of our data, the algorithm achieves the best clustering quality with four clusters, which confirms what was shown by the dendrogram. It should be noted that a similar curve, with a peak for 4 clusters, is also observed when launching the K-means on 10 PCA components instead of four. Another interesting finding is that the silhouette value reached with 4 clusters and random initialization is identical to that obtained when initializing the K-means on the centroids of the dendrogram, which might indicate a potential similarity between both clusterings. It should be highlighted, however, that the obtained silhouette values are rather low, indicating a rather uncertain distinction between the clusters.

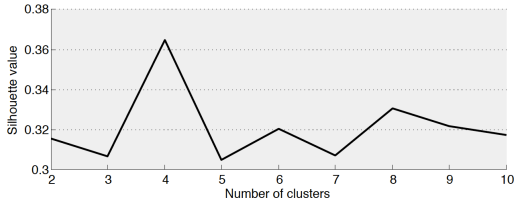


Figure 3: *Value of the silhouette of the K-means on four PCA components according to the number of clusters, with random initialization.*

The rand-index [30] allows comparing two clusterings. Its value ranges between 0 and 1, 1 corresponding to two identical clusterings. Paired comparisons were performed for four K-means clusterings in 4 clusters: with initializations on centroids of dendrogram on 4 (i) and 10 PCA components (ii) and with random initialization on 4 (iii) and 10 PCA components (iv). All rand-index values reach a level above 0.93 which indicates a certain stability, all clusterings converging towards the same solution.

The first of the four clusterings is used in the remainder of this study. A prosodic analysis of the syllables contained in the four clusters indicates that each cluster can be associated to a specific realization in terms of the three main prosodic features (i.e. energy, F0 and duration), as shown in Table 1. To investigate whether the increase in the number of clusters goes in line with an increase in the naturalness of the expressivity, the version with 10 clusters, as obtained with random initialization on the 4 PCA components, is also assessed in the perceptual evaluation. This clustering provides a vectorial quantification of the acoustic space, each region being assigned to a different cluster. It should be noted, however, that it is obviously more complex to predict such a high number of tags from a text to synthesize.

Table 1: *Acoustic characteristics of the clusters, compared to average acoustic values of emphatic stresses.*

Cluster	Energy	F0	Duration
Cluster 1	-	+	-
Cluster 2	-	-	-
Cluster 3	+	+	-
Cluster 4	+	+	+

3.4. Correlation between clusters and linguistic contexts

Potential correlations between the four defined clusters and linguistic information (syllable position, structure, etc.) are investigated. HMM-based speech synthesizers relying on such contextual criteria to cluster the models, correlations would indicate a possible automatic distinction between the various acoustic realizations. This would imply that the clusters would not need to be explicitly distinguished in the annotation.

We analyzed 13 linguistic variables used as contextual information for synthesis: position of the syllable or word in the word or rhythmic group (RG), amount of syllables in word and RG, amount of (content) words in RG, structure of the syllable, nature of the nucleus and part of speech of the word. Seeing the high amount of samples, Chi-square tests tend to be significant for most variables. Cramer's V [31] allows interpreting chi-squares for high effectives. Table 2 shows that only weak associations (i.e. $V < 0.2$ [32]) can be seen between the acoustic clusters and contextual linguistic information.

Table 2: *Correlation between the four clusters and linguistic contextual information (first five variables).*

Variable	Cramer's V
Syllable position in word	0.1379
Nature of the vowel	0.1165
Word position in rhythmic group (forward)	0.1164
Syllable position in word (forward)	0.1141
Syllable position in word (backward)	0.1137

Interestingly, the highest value (i.e. 0.14) is assigned to 'syllable position in word', mainly informative about whether a syllable is initial or final. The omnipresence of both syllable and word position in the ranking drove us to investigate whether some acoustic differences may be due to final syllables at the end of the RG. In that case, they might coincide with what is commonly referred to as boundary tone [13], which could influence their realization. Table 3 shows that the acoustic values of those syllables are significantly higher compared to the other emphatic stresses (respectively p=4.1e-05, p=1.2e-04 and p=1.6e-08 for a bilateral ranksum test performed on 82 final emphatic syllables and 721 other emphatic syllables). This indicates that the distinction between the clusters can partly be explained by linguistic contextual information. Associations

are however rather weak, suggesting that other factors probably play a role in the acoustic realization of emphatic stresses.

Table 3: *Acoustic realizations of final emphatic stresses (end of word at the end of RG) and other emphatic stresses, together with their 95% confidence intervals.*

Emphatic stress	Mean F0 (Hz)	Syllable Dur (z-score)	Mean Energy (dB)
Final	262.9 ±7.4	1.9 ±0.6	49.4 ±1.8
Other	242.9 ±3	0.84 ±0.1	44.2 ±0.4

4. Speech synthesis: A perceptual study

4.1. Evaluation protocol

In order to assess the quality of the expressivity produced when integrating various types of emphatic stresses, several HMM-based speech synthesizers [33] were built, relying on the implementation of the HTS toolkit (version 2.1) publicly available in [34]. For each synthesizer, 90% of the corresponding database was used for the training (called the *training set*), leaving around 10% for the synthesis (called the *synthesis set*). As filter parameterization, we extracted the Mel Generalized Cepstral (MGC) coefficients traditionally used in parametric synthesis. As excitation modeling, the Deterministic plus Stochastic Model (DSM [35]) of the residual signal was used to improve naturalness. Emphatic annotation was used as contextual information, in the same way as linguistic information.

Three models are compared: the baseline model (*Baseline*), using only one emphatic stress, and the models with 4 (*4 Stresses*) and 10 (*10 Stresses*) emphatic stresses, as obtained by annotating the emphatic syllables with the 4 and 10 clusters defined in the previous section. Test sentences were automatically selected from the synthesis set, as being shorter than 5 seconds and displaying at least two emphatic stresses in their annotation. It is indeed much easier to compare short sentences in which more than one difference appears. The test consisted in 18 pairs of sentences, 6 from each comparison, randomly selected from 63 pairs (21 for each comparison).

30 native French-speaking testers, mainly naive listeners, participated in the evaluation. During the test, they could listen to the pair of sentences as many times as wanted. For each comparison, they were first asked whether they heard any difference between both versions of the sentence. If so, they were asked to compare them in terms of naturalness of the expressivity. The scale ranged from -3 (much less natural) to +3 (much more natural). A score of 0 was given if both versions were found to be different but with equivalent naturalness of the expressivity.

4.2. Results

A first interesting finding is that the testers did not hear any difference between both versions for around 20 % of the pairs. This percentage is even higher (i.e. 28%) for models with 4 and 10 clusters which tend to display rather similar intonational patterns. Figure 4 shows the preference percentages for the remaining pairs. Middle sections correspond to pairs which were considered as similarly natural in terms of expressivity. We can observe that the model with 4 emphatic stresses slightly outperforms both other models. This might be explained by the fact that it more accurately synthesizes the various acoustic realizations of the stress. In the10-cluster model, we notice a degradation which may be due to the reduced number of occurrences for each stress, which is partly alleviated with the 4-cluster model. However, the preference for 4 clusters rather than one single stress (i.e. the baseline) is quite weak and is not statistically significant (p=0.11 with a unilateral ranksum test comparing the average percentage of preferences on the 30 testers).

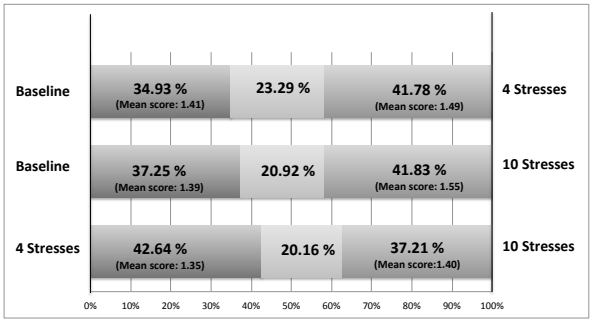


Figure 4: *Percentage of preferences for each model in the three comparison pairs.*

Figure 4 also shows, in parenthesis, the mean score obtained by the three models when preferred in the comparison. These scores are barely higher than 1 because testers mostly assigned a score of '1', reflecting only a 'slight' preference.

5. Conclusion

Emphasis is known to play a crucial role in expressive speech. Its generation in speech synthesis usually relies on a prosodic annotation of the training corpus. For that matter, emphatic stresses can be assigned a single label or be divided into distinct labels according to their acoustic realization. While the prediction of a single label from text is easier, the use of different tags might allow for the generation of more suitable stresses. The question is then whether the use of several emphatic labels effectively improves the naturalness of the expressivity.

The objective of this paper was precisely to answer this latter question by investigating HMM-based speech synthesis using one or several emphatic labels. Statistical acoustic analyses allowed determining 4 and 10 distinct emphatic labels based on a set of extracted acoustic features at the syllable level. The definition of 4 clusters was shown to achieve the best clustering quality. The model with 10 labels was computed to propose a vectorial quantification of the acoustic space. Both models were compared to a baseline model using a single emphatic label.

Perceptual tests showed that the model with four emphatic stresses is slightly preferred over both other models. However, the differences are not significant. While participants did not perceive any difference in 20% of cases, more than 20% of the remaining pairs were scored as 'similar' regarding the naturalness of expressivity. For pairs for which a preference was given, the score was usually low, denoting a weak degree of preference. These results tend to indicate that it might not be required to explicitly annotate different kinds of emphatic stresses in the corpus, when using HMM-based speech synthesis. This is a clear advantage for the annotation of the text to synthesize. However, this finding should be confirmed with further investigations on data with other speaking styles and languages.

6. Acknowledgements

The two first authors are supported by FNRS. The project is partly funded by the Walloon Region Wist 3 SPORTIC.

7. References

[1] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *ICSLP*, 2004, pp. 1185–1188.

[2] J. Yamagishi, K. Onishi, T. Musuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IECE Transactions on Information and Systems*, vol. E88-D(3), pp. 502–509, 2005.

[3] L. Qin, Z.-H. Ling, Y.-J. Wu, B.-F. Zhang, and R.-H. Wang, "HMM-based emotional speech synthesis using average emotion models," in *ISCSLP*, 2006, p. 233240.

[4] R. Fernandez and B. Ramabhadran, "Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis," in *SSW6*, 2007.

[5] V. Strom, R. Clark, and S. King, "Expressive prosody for unit-selection speech synthesis," in *Interspeech*, 2006.

[6] J. Hirschberg, "Accent and discourse in context: Assigning pitch accent in synthetic speech," in *AAAI*, 1990.

[7] L. Badino, J. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in HMM-based speech synthesis," in *Interspeech*, 2009.

[8] S. Brognaux, B. Picart, and T. Drugman, "A new prosody annotation protocol for live sports commentaries," in *Interspeech*, 2013.

[9] B. Picart, S. Brognaux, and T. Drugman, "HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation." in *8th ISCA Speech Synthesis Workshop (SSW8)*, 2013.

[10] A. Raux and A. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *ASRU*, 2003.

[11] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alavrez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Interspeech*, 2007.

[12] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in *ICASSP*, 2010.

[13] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[14] D. Hovy, G. Krishna Anumanchipalli, A. Parlikar, C. Vaughn, A. Lammert, E. Hovy, and A. Black, "Analysis and modeling of "focus" in context," in *Interspeech*, 2013.

[15] J.-P. Goldman, M. Avanzi, A. Auchlin, and A. C. Simon, "A continuous prominence score based on acoustic features," in *Interspeech*, 2012.

[16] D. Hirst, "Form and function in the representation of speech prosody," *Speech Communication*, vol. 46, pp. 334–347, 2005.

[17] J. Trouvain, "Between excitement and triumph - live football commentaries in radio vs. TV," in *17th International Congress of Phonetic Sciences (ICPhS XVII)*, 2011.

[18] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under Praat," in *Interspeech*, 2011, pp. 3233–3236.

[19] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignments," in *IEEE Workshop on Spoken Language Technologies*, 2012, pp. 410–415. [Online]. Available: http://cental.fltr.ucl.ac.be/train_and_align/

[20] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Interspeech*, 2005, pp. 2549–2552.

[21] R. Johnson and D. Wichern, *Applied multivariate statistical analysis, 5th Ed.* Prentice Hall, 2002.

[22] A. Izenman, *Modern multivariate statistical techniques*. Springer, 2008.

[23] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech*, 2011.

[24] J.-P. Goldman, M. Avanzi, A. Lacheret-Dujour, A. C. Simon, and A. Auchlin, "A methodology for the automatic detection of perceived prominent syllables in spoken French," in *Interspeech*, 2007, pp. 98–101.

[25] S. Brognaux, T. Drugman, and R. Beaufort, "Automatic detection of syntax-based prosody annotation errors," in *IEEE Spoken Language Technology Workshop (SLT)*, 2012.

[26] H. A. Rositzke, "Vowel-length in general American speech," *Language*, vol. 15, pp. 99–109, 1939.

[27] A. Di Cristo, "De la microprosodie à l'intonosyntaxe," Ph.D. dissertation, Université de Provence, Aix-en-Provence, 1985.

[28] J. Ward, "Hierarchical grouping to optimise an objective function," *Journal of the American Statistical Association*, vol. 58, p. 236244, 1963.

[29] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[30] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American statistical Association*, vol. 66, pp. 846–850, 1971.

[31] H. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

[32] L. M. Rea and R. A. Parker, *Designing and Conducting Survey Research*. Jossey-Bass, 1992.

[33] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51(11), pp. 1039–1064, 2009.

[34] Hmm-based speech synthesis system (hts). [Online]. Available: http://hts.sp.nitech.ac.jp

[35] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20(3), pp. 968–981, 2012.

Sources of variation of articulation rate in native and non-native speech: comparisons of French and German

Jürgen Trouvain, Bernd Möbius

Department of Computational Linguistics and Phonetics, Saarland University, Germany
trouvain|moebius@coli.uni-saarland.de

Abstract

Speech tempo including articulation rate is often considered as a good predictor in the diagnosis of foreign language proficiency and its comprehension. In this study we investigate various sources of variation of articulation rate such as the L2 proficiency level, individual tempo habits in L1 and L2, and more extensive exposure to native speech. In addition, we also discuss the difficulty of defining the most informative unit for rate metrics which allows comparisons between French and German. The materials used are French and German read sentences, produced as L1 and L2 speech. In contrast to other studies individual habits of articulation rate in the L1 was only partially observed in the corresponding L2 data (a slow L1 speaker does not necessarily articulate slowly in the L2). The convergence of most French learners to the German model speakers shows the advantage of having additional input for phonetic exercises. The fastest German learners also converge to the rather slow French model speaker.

Index Terms: articulation rate, L1, L2, convergence, individuality, French, German

1. Introduction

Speech tempo is often considered as a good predictor for various important concepts in the diagnosis of second and foreign language (henceforth L2) including the level of

- L2 proficiency [1],
- intelligibility and comprehension [2], and
- perceived foreign accent [2].

Articulation rate as the key component of speech tempo would thus be an optimal and easy-to-handle indicator of the level of the *spoken* language learning process.

It can be seen as established that *on average* speech in the first language (L1) is articulated faster than L2 speech, be it that the same speakers are faster in their L1 than in their L2 [3, 4, 5] or that L1 speakers are faster than L2 speakers in a given language [4, 6, 5]. Therefore the quantified tempo, whatever be the metrics, should reflect this difference between L1 and L2 speech. In addition, L2 speakers with a lower level of L2 proficiency will in general be slower than L2 speakers with a higher level of L2 proficiency [7].

Another factor of variability is the fact that languages show different measured rates. When we compare German with French, French usually shows 'faster' syllabic rates than German. One could argue that articulation in French is actually *produced* faster because it is *perceived* as faster than German – by German listeners. However, this argument can be applied in the other direction as well: articulation of German is also perceived as faster than 'usual' – by French listeners.

Cross-language studies on speech tempo either compare L1 speakers of one language with L1 speakers of another language (ideally with comparable text material), or L1 speakers of a given language with L2 speakers of the same language (ideally with identical text material).

A further source of variation lies in the fact that individual speakers differ in their tempo [10]. Some speakers articulate faster than others, be it in their L1 or in an L2.

From a learning perspective it can be assumed that L2 speakers perform better when material to be read is not only presented in its orthographical form but also provided in spoken form by a native speaker. In the latter case the learner's articulation should be faster and thus converging to the native-like articulation rate [8]. Reading only would then be slower than reading after listening to L1 speech.

In total we can thus identify four main sources of variation due to L2 proficiency, rate metrics, individuality and convergence. These sources of variation with expected patterns are illustrated schematically in Fig. 1 and described in more detail in the following subsections. Other sources of variation of articulation rate like for instance sex and age [9] are not considered here because they are not of primary relevance.

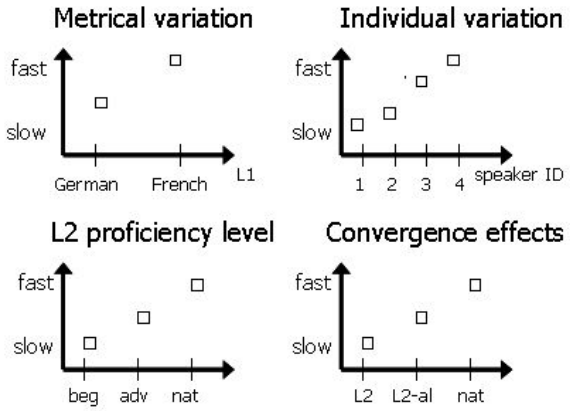


Figure 1: *Sources of variation of articulation rate with expected patterns (beg=L2 beginners, adv=advanced L2 learners, nat=native speakers, L2-al=L2 speech after listening).*

1.1. Speech rate metrics

When considering speech tempo or speech rate we distinguish between the *speaking rate* as the gross rate *in*cluding all pauses and *articulation rate* as the net rate of articulated speech, i.e. *ex*cluding all pauses.

In the concept of *fluency* (which is beyond the scope of this paper) articulation rate would be part of speed fluency which can be distinguished from repair fluency and breakdown fluency (cf. e.g. [10]).

Another issue in speech rate metrics is the choice of the basic linguistic unit. Possible candidates are the word, the syllable and the segment. The most popular unit seems to be the syllable [11, 9], thus having syllables per second or conversely the mean syllable duration as the preferred metrics. The advantage of the syllable in comparison to the word is that it is less variable in length within and across languages (compare for instance the mean length of words in Finnish or Turkish with those in French or English). The disadvantage of segments and syllables as basic units is that counting is not as easy as words. Another disadvantage of the segment is that they are more frequently elided than syllables or even words.

The aspect of elision leads to the next challenge selecting the optimal speech rate metric. Do we operate with phonological units which can be assumed as underlying representations and used as planned units, or should we focus on the actual realisation of these units (cf. [9, 12])? An argument in favour of the phonological units would be that the omitted segments or syllables or words are likely to be accounted for in speech planning and probably in speech perception, too. Thus, one strategy of speaking faster is to omit more units. An argument against the phonological units would be that speech with a moderate speed of articulation but with many omissions would reflect sloppiness but not necessarily fast articulation. For pragmatic reasons the use of phonological units often seems to be the preferred option, because for scripted material the number of units would be the same and the working load of analysing all speech samples with regard to omissions does not apply (apart from the fact that omission of segments is not always as clear as wished).

1.2. Research questions

1.2.1. Rate metrics

What is the most informative unit for rate metrics which allows comparisons between French and German? We are looking for a metric that facilitates the comparison of the speaking tempo between various speech modes such as L1 speech of different languages, L2 speech of different languages, different levels of proficiency in L2 speech, and further individual differences, in L1 *and* L2 speech. Probably there is not a single metric reflecting all these levels of variation.

However, we expect that syll/s is not the best metric to reflect articulation rate in L1 and L2 speech for the French-German language pair. Comparisons of syllable rates between German and French show substantially higher values for French, e.g. 7.3 syll/s (Fr.) vs. 5.6 syll/s (G.) in [13] or 7.18 syll/s (Fr.) vs. 5.97 syll/s (G.) in [14]. This phenomenon can be explained by a higher complexity of syllables in German where more time is required to articulate more segments in a syllable.

1.2.2. Native speech and the level of L2 proficiency

Regarding native speech and the level of L2 proficiency we ask: Do beginners articulate slower than advanced learners, and do native speakers articulate faster than advanced learners? For both questions we would expect an affirmative answer for both languages.

1.2.3. Individual habits of articulation rate

Are individual articulation rate habits in L1 visible in L2 speech? Following [15, 10], who found partial evidence for rate variation on the learner's personal speaking style irrespective of L2 proficiency, we can expect an affiramtive answer to this question, too.

1.2.4. Convergence to L1 speech

Can we observe any convergence of articulation rate to native speakers when learners get more extensive exposure to native speech, e.g. reading a sentence after listening to that sentence read by a L1 speaker? The use of more than one sensory channel is generally assumed to be helpful in teaching and learning foreign language phonetics [16, 17]. We would expect an affirmative answer to this question as well.

1.3. Outline

The remainder of the paper is structured as follows: in section 2 we describe the materials and subjects used for this study, followed by an analysis of various metrics of articulation rate. In section 3 we present the results with regard to rate metrics and the research questions, which will be discussed in section 4 which concludes the paper with an outlook on future work.

2. Method

2.1. Material

We used a part of a phonetic and phonological learner corpus for the language pairs French-German and German-French [18]. The data are from 7 speakers with French as their first language and 7 speakers with German as L1. Each language group consisted of five speakers at a beginner level (A1 or A2 according to the European reference frame CERF) and two at an advanced level (C1 or C2 at CERF). All speakers were recorded in their L2 *and* their L1.

Samples from two conditions were selected, viz. (i) read sentences ("read"), (ii) repeated sentences ("repeated"). In the latter condition only L2 speech was recorded: the sentences were prompted orthographically (as in the "read" condition) but synchronously with an audio file spoken by a native speaker. These native model speakers were not subjects in this corpus. The task was to repeat the written sentence just heard. Note that the repetition was produced a few seconds after listening in order to avoid a direct imitation.

For each of the 14 speakers we analysed 10 sentences in the "read L1", 10 in the "read L2", and 10 in the "repeated" condition in the respective L2. As a consequence the sentences "Read French" and the sentences "Read German" were produced by all 14 speakers, either in their L1 or in their L2. The sentences "Repeat French" were produced only by the German speakers, and the sentences "Repeat German" only by the French speakers. In total the analysis included 30 (sentences) x 14 (speakers) = 420 sentences.

2.2. Analysis

Start and end of the articulation of the sentences as well as silent and breathing pauses (if present) were first determined by means of an automatic speech recognition procedure using forced alignment. In a second step all labels were auditorily and visually checked with the speech editor Praat and manually corrected, if needed. We noted the appearance of possible addi-

tional syllables due to repetition, reflecting a disfluency that is not uncommon in L2 speech.

For each sentence the speaking rate (as a gross rate) and the articulation rate (as a net rate) were calculated for phones per second (phon/s), syllables per second (syll/s), and words per minute (wds/min).

The unit for syll/s was a *phonological* syllable, i.e. as produced in a canonical form. For phon/s the basic unit was the segment as predicted from the canonical form. The unit for wds/min was the word as counted in a text processor, e.g. the French "c'est" or "s'est" was counted as one word.

3. Results

3.1. Rate metrics

In about half of the sentences of L2 speech a pause was inserted. Pause insertion in L1 speech happened infrequently (less than 5 percent).

Fig. 3 shows that the syllable rates for French L1 and German L1 differ substantially whereas this is not the case for the phone rate (Fig. 2) and the word rate (Fig. 4).

The individual values of syll/s for the German speakers (S8-S14) show for the majority of speakers higher rates in their L2 than in their L1.

The distinction between L1 and L2 for each speaker group is achieved best for the metric phon/s. Thus, we present the data in phon/s in the following sections.

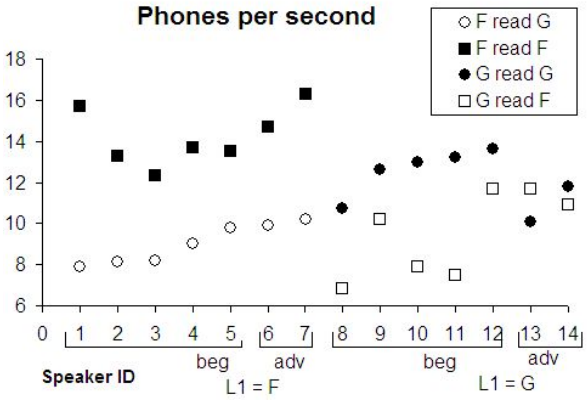


Figure 2: *Articulation rate of speakers in phon/s for the "read" condition. Speakers 1-7 with L1 = French (F), 8-14 with L1 = German (G). Speakers 6-7 and 13-14 are more advanced L2 speakers.*

3.2. L2 proficiency

In Fig. 2 the differences between native and non-native speech are clearly visible. For the French sentences the slowest L1 speaker is faster than the fastest L2 speaker which clearly marks a distinctive line between L1 French and L2 French speech. Similarly, in German the slowest L1 speaker (S13) articulates as fast as the fastest L2 speaker (S7). Both groups are clearly distinguishable.

For the level of L2 proficiency we see that the fastest L2 speakers (S7 for French, S13 for German) belong to the advanced learners. Analogously the slowest L2 speakers are L2 beginners (S1 for French, S8 for German). However, selected

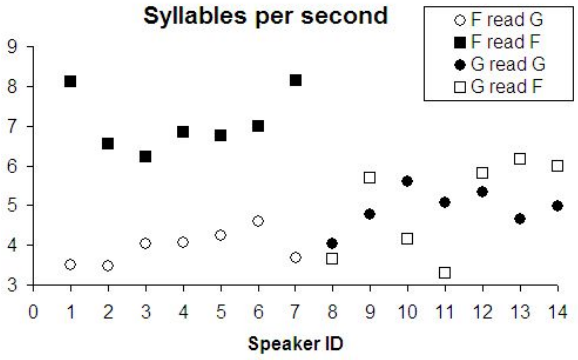


Figure 3: *Articulation rate in syll/s for the "read" condition. Speaker grouping as in Fig. 2.*

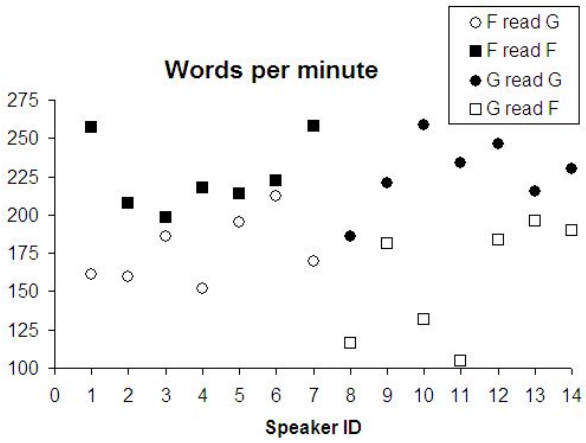


Figure 4: *Articulation rate in wds/min for the "read" condition. Speaker grouping as in Fig. 2.*

beginners such as S5 (for French) and S12 (for German) are also very fast.

3.3. Individuality

Fig. 5 shows that the range of articulation rate for French L1 speech lies between 12 and 17 phon/s, whereas German L1 speech shows values between 10 and 14 phon/s. Thus, the difference between the slowest and the fastest speaker in each L1 group is between 4 phon/s (G.) and 5 phon/s (Fr.).

When looking at the L2 groups these differences change: the French speakers with L2 German articulate with a rate between 8 and 11 phon/s, thus reducing the range to 3 phon/s. The articulation rate of the German speakers with L2 French lie between 7 and 12 phon/s, thus expanding the range to 5 phon/s.

Individual tempo habits also lead to the rather unexpected outcome that one speaker (S13) shows 'faster' values in L2 than in L1 speech. This example and also S14 show that slow talkers in their L1 can be fast talkers in their L2. Likewise, fast L1 speakers are not necessarily fast L2 speakers: the fast L1 speaker S1 is in his L2 speech as slow as S3 who is the slowest L1 speaker in French. Counterexamples are the German speakers S8, who is slow in his L1 and L2, and S12 as a fast counterpart.

3.4. Convergence

To compare the "read" and the "repeat" conditions for the two L2 groups Fig. 5 shows that all French beginners articulate faster when they have additional auditive input of L1 speech. Among the German beginners of French only three out of five were faster in the "repeat" condition.

The convergence effect is diminished for the advanced learners where only one out of four speakers was faster than in the "read only" condition. In general it can be seen that the learners with the slowest rates in "read" speeded up in "repeat".

The mean rate for the sentences by the two German model speakers in the "repeat" condition was 11.9 phon/s (11.7 phon/s for the one speaker, 12.1 phon/s for the other speaker). The speakers S5 (beginner) and S7 (advanced) nearly approached this articulation rate after listening but not for reading only.

The French model speaker was a rather slow speaker with 10.3 phon/s. All L2 speakers were slower than him, although three speakers (S12-S14) were faster in the "read" condtion.

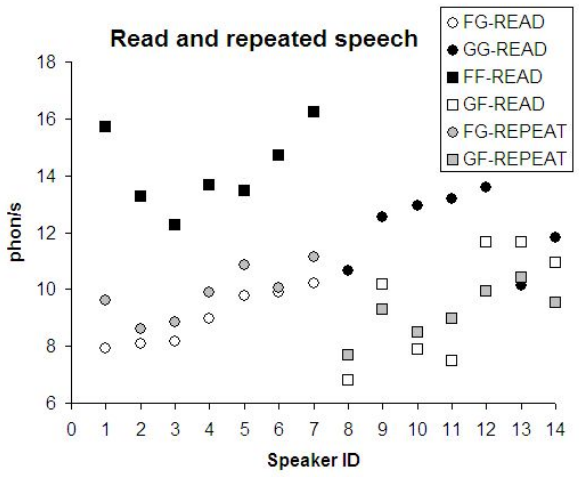


Figure 5: *Articulation rate for the "read" and "repeat" conditions. Speaker grouping as in Fig. 2.*

4. Discussion and conclusion

The rather low number of subjects per language and proficiency level allows only a limited interpretation of the results. Thus we need to be cautious with respect to generalisations. However, the explorative approach taken here has yielded interesting observations worth reporting and they represent good starting points for follow-up studies.

The best reflection of the L1-L2 difference for each speaker in our data seems to be phon/s. However, the phone rate for French is still considerably higher than for German (by 2.1 phon/s). A correction by this difference for the L1 German speech would result in a balanced picture with comparable ranges for L1 French and L1 German on the one hand, and comparable differences between L1 and L2 for the French speakers and the German speakers on the other hand. However, such a correction would ignore a correction for L2 German which would be a theoretical flaw. Although studies on speech rhythm often apply a normalisation of articulation rate between different languages (e.g. [13]), this usually concerns L1 speech and not L2 speech of the same speakers.

L2 proficiency is only to a limited extent reflected by articulation rate in our data. We deliberately left out learners at a medium level (B1 or B2 at CERF), but the distinction between beginners and advanced learners is not as clear-cut as expected.

The expectation regarding the transfer of the individual articulation rate habits from L1 to L2 is not entirely met. There is a huge variation among L2 speakers that *only partially* reflects L1 rate habits – in contrast to other studies [15, 10].

Our hypothesis about converging articulation rates of an L2 speaker after listening to an L1 speaker was confirmed. There is a positive effect of additional listening on speeding up for beginners with slow articulation rates. The extremely slow articulation rate of the French model speaker had a slowing-down effect on the fastest German speakers after listening. In both cases convergence in terms of articulation rate has evidently occurred. Since speaking slower may also indicate a more careful speaking style the slowed down L2 speakers possibly felt motivated to speak more clearly as well. Future studies are needed to clarify what kind of speaking style and what articulation rate can help which type of learner to improve the intelligibility and the fluency in L2 speech production.

Further topics for future research include the choice of rate metric. First, the metrics syll/s and phon/s are calculated based on the phonological structure rather than its actual realisation. Second, it is currently unclear how far the metrics applied here reflect the *perceptual* tempo [12, 19]. Another question in the context of speech perception is how comprehension is correlated with articulation rate. A last but not least point concerns the speech materials. In this study we used read sentences. It remains unclear how we can generalise from this sort of data to scripted and non-scripted styles, particularly when the discourse units are larger than a sentence. Nevertheless, read sentences, also with additional listening, are quite useful for exercises in learning and teaching environments.

The basic idea we started with was that articulation rate as the key component of speech tempo would be a good and easy-to-handle indicator of the level of the *spoken* language learning process. We have shown that for a comparison of different languages a correction for the variation caused by different phonological complexities should be considered.

As expected, more proficient L2 learners show remarkably high articulation rates, although not as high as L1 speakers. It is important to note that among the beginners there are speakers with similarly high articulation rates. One possible explanation for this finding is that different phonetic talents [20] also influence the speed of articulation.

An interesting observation is that there is generally an improvement in the temporal control of articulation when the L2 speakers had an additional auditory input of the sentence to be read. Although this is not the case for all L2 speakers, most of the *beginners* speeded up. It can be expected that this group benefit most from the input in multiple modalities when learning an L2. This point is important for the selection of exercises where speed of articulation should be part of a test and evaluation scheme which also includes computer-assisted pronunciation training.

5. Acknowledgements

The authors would like to thank Anjana Vakil for her help with data processing and Frank Zimmerer for useful discussions. This work was made possible by the project IFCASL funded by Deutsche Forschungsgemeinschaft (PIs: B. Möbius and J. Trouvain) and Agence National de Recherche (PI: Y. Laprie).

6. References

[1] P. Lennon, "Investigating fluency in EFL: A quantitative approach," *Language Learning*, vol. 40, pp. 387–417, 1990.

[2] M. Munro and T. Derwing, "Modelling perceptions of the comprehensibility and accentedness of L2 speech: The role of speaking rate," *Studies in Second Language Acquisition*, vol. 23, pp. 451–468, 2001.

[3] H. Pürschel, *Pause und Kadenz. Interferenzerscheinungen bei der englischen Intonation deutscher Sprecher*. Tübingen: Max Niemeyer Verlag, 1975.

[4] M. Raupach, "Temporal variables in first and second language speech production," in *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, H. Dechert and M. Raupach, Eds. The Hague: Mouton, 1980, pp. 263–270.

[5] U. Gut, *Non-native Prosody. A corpus-based analysis of the phonetic and phonological properties of L2 English and L2 German*. Frankfurt: Peter Lang, 2009.

[6] R. Wiese, *Psycholinguistische Aspekte der Sprachproduktion*. Hamburg: Buske, 1983.

[7] U. Gut, "Prosody in second language speech production: the role of the native language." *Fremdsprachen Lehren und Lernen*, vol. 32, pp. 133–152, 2003.

[8] R. L. Street and H. Giles, "Speech accommodation theory: A social cognitive approach to language and speech behavior," in *Social Cognition and Communication*, M. E. Roloff and C. R. Berger, Eds. Beverly Hills, CA: Sage, 1982, pp. 193–226.

[9] J. Trouvain, *Tempo Variation in Speech Production. Implications for Speech Synthesis*. PhD Dissertation, Saarland University, Saarbrücken, 2004.

[10] N. H. De Jong, R. Groenhout, R. Schoonen, and J. H. Hulstijn, "Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior," *Applied Psycholinguistics*, vol. 34, p. http://dx.doi.org/10.1017/S0142716413000210, 2013.

[11] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of the Acoustical Society of America*, vol. 88, pp. 101–112, 1990.

[12] J. Koreman, "Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech," *Journal of the Acoustical Society of America*, vol. 119, pp. 582–596, 2006.

[13] V. Dellwo and P. Wagner, "Relations between language rhythm and speech rate," in *Proc. ICPhS, Barcelona*, 2003, pp. 471–474.

[14] F. Pellegrino, C. Coupé, and E. Marsico, "A cross-language perspective on speech information rate," *Language*, vol. 87, pp. 539–558, 2011.

[15] T. M. Derwing, M. J. Munro, R. I. Thomson, and M. J. Rossiter, "The relationship between L1 fluency and L2 fluency development." *Studies in Second Language Acquisition*, vol. 31, pp. 533–557, 2009.

[16] U. Hirschfeld and J. Trouvain, "Teaching prosody in German as a foreign language," in *Non-Native Prosody. Phonetic Description and Teaching Practice.*, J. Trouvain and U. Gut, Eds. Berlin: Mouton De Gruyter, 2007, pp. 171–187.

[17] H. Mitterer and J. McQueen, " Foreign subtitles help but native-language subtitles harm foreign speech perception," *PLoS One*, vol. 4, pp. A146–A150, 2009.

[18] J. Trouvain, Y. Laprie, B. Möbius, B. Andreeva, A. Bonneau, V. Colotte, C. Fauth, D. Fohr, D. Jouvet, O. Mella, J. Jügler, and F. Zimmerer, "Designing a bilingual speech corpus for French and German language learners," in *Proc. Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages, Strasbourg*, 2013, pp. 32–34.

[19] H. R. Pfitzinger, "Local speech rate perception in german speech," in *Proc. ICPhS, San Francisco*, vol. 2, 1999, pp. 893–896.

[20] M. Jilka, "Talent and proficiency in language," in *Language Talent and Brain Activity.*, G. Dogil and S. Reiterer, Eds. Berlin: Mouton De Gruyter, 2009, pp. 1–16.

Extending AuToBI to prominence detection in European Portuguese

Helena Moniz[1,2]*, Ana Isabel Mata*[2]*, Julia Hirschberg*[3]*,*
Fernando Batista[1,4]*, Andrew Rosenberg*[5]*, Isabel Trancoso*[1,6]

[1]Spoken Language Systems Lab - INESC-ID, Lisbon, Portugal
[2]FLUL/CLUL, Universidade de Lisboa, Portugal
[3]Department of Computer Science, Columbia University, United States
[4]ISCTE - Instituto Universitário de Lisboa, Portugal
[5]Computer Science Department, Queens College (CUNY), United States
[6]IST, Universidade de Lisboa, Portugal
{helenam;fmmb;imt}@l2f.inesc-id.pt, aim.@fl.ul.pt, julia@cs.columbia.edu

Abstract

This paper describes our exploratory work in applying the Automatic ToBI annotation system (AuToBI), originally developed for Standard American English, to European Portuguese. This work is motivated by the current availability of large amounts of (highly spontaneous) transcribed data and the need to further enrich those transcripts with prosodic information. Manual prosodic annotation, however, is almost impractical for extensive data sets. For that reason, automatic systems such as AuToBi stand as an alternate solution. We have started by applying the AuToBI prosodic event detection system using the existing English models to the prediction of prominent prosodic events (accents) in European Portuguese. This approach achieved an overall accuracy of 74% for prominence detection, similar to state-of-the-art results for other languages. Later, we have trained new models using prepared and spontaneous Portuguese data, achieving a considerable improvement of about 6% accuracy (absolute) over the existing English models. The achieved results are quite encouraging and provide a starting point for automatically predicting prominent events in European Portuguese.

Index Terms: prosody, automatic prosodic labeling system, spontaneous speech.

1. Introduction

The role of detecting prosodic events is becoming more and more pervasive in different automatic speech processing tasks. The detection of prosodic events has proved to be useful for, *e.g.*, improving speech summarization [1], in ASR models [2, 3, 4], in predicting ASR recognized turns in dialogue systems [5], in predicting structural metadata events [6, 7, 8, 9], in improving unit-selection synthesis [10] or in detecting phonological units for expressive speech synthesis [11], and in identifying paralinguistic events [12].

The literature has documented a set of acoustic and visual correlates of prominence - duration, intensity, pitch, voice quality, and visual cues [13, 14, 15, 16, 17, 18, 19, 20, 21]. Systems built to predict prominence based on acoustic correlates are being used for cross-language studies. One such example is the **Au**tomatic **ToBI** annotation system (AuToBI) for Standard American English (SAE) by [17, 22]. AuToBI is a publicly available tool[1], which detects and classifies prosodic events following SAE intonational patterns. AuToBI relies in the fundamentals of the ToBI system, meaning it predicts and classifies tones and break indices using the acoustic correlates - pitch, intensity, spectral balance and pause/duration. AuToBI uses a modular architecture, which allows to perform six tasks separately and provides English trained models for spontaneous and read speech (for further details, *vide* [17] and references therein). The six tasks correspond to: i) detection of pitch accents; ii) classification of pitch accent types; iii) detection of intonational phrase boundaries; iv) detection of intermediate phrase boundaries; v) classification of intonational phrase boundary tones; and vi) classification of intermediate phrase boundary tones. Previous work on prosodic event detection using AuToBI [17, 22, 23, 24] have shown that prominence and phrase boundaries can be predicted in a cross-language (American English, German, Mandarin Chinese, Italian and French) context, albeit with language specific properties. Those studies also found little support for the hypothesis that language families are useful for cross-language prosodic event identification. Taking into account the described results, the aim of this work is to extend the AuToBI prosodic event detection system from English to Portuguese in two stages. First, English models are used to predict prosodic events in European Portuguese (detection and classification of pitch accents). Second, AuToBI capabilities are adapted using a relatively small amount of annotated data to train Portuguese models.

This paper is organized as follows. Section 2 comprises the description of the corpus used in the experiments. Section 3 presents the process of converting of tone inventories from P-ToBI to SAE, a pre-requirement to apply AuToBI to European Portuguese. Section 4 describes the experiments conducted and the main results achieved. Section 5 presents our conclusions and future work.

2. Corpus

This study uses a subset of CPE-FACES [25, 26], an European Portuguese Corpus Spoken by Adolescents in School Context. The corpus consists of spontaneous and prepared unscripted speech from 25 students (14-15 years old) and 3 teachers, all

[1]http://eniac.cs.qs.cuny.edu/andrew/autobi/

Corpus subset →	train	test	total
total time (minutes)	33.4	10.9	44.2
useful time (minutes)	20.6	6.6	27.2
number of pitch accents	2061	717	2778
number of phrasing units	1382	489	1871
number of words	4361	1456	5817

Table 1: Properties of the CPE-FACES subsets.

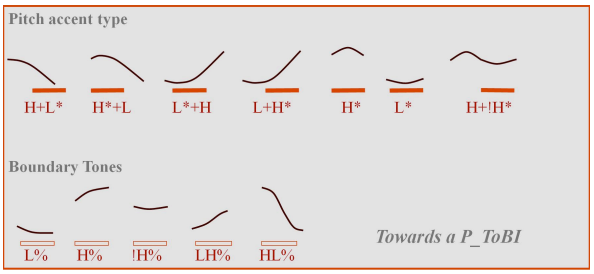


Figure 1: Schematic F0 contours. Thick red lines indicate the stressed syllable.

speakers of standard European Portuguese (Lisbon region), totaling approximately 16h. It was designed to represent some of the speech tasks that are common in school context and it was collected in the last year of compulsory education (9th grade), in three Lisbon public high schools. It was recorded in a natural setting – the speakers classroom of Portuguese as L1 – in different communication situations: two dialogues (both spontaneous) and two oral class presentations (one spontaneous and another one prepared, but unscripted). In the spontaneous presentation, students and teachers were unexpectedly asked to relate an (un)pleasant personal experience. It was assumed that the involvement of speakers on topics related to their personal interests and day-to-day life would manifest in the naturalness and spontaneity of their talks [27]. The prepared situation corresponds to typical school presentations; the presentation was about a book the students must read following specific programmatic guidelines. For students, a variety of presentations on Ernest Hemingway's "The Old Man and the Sea" and on Gil Vicente's "Auto da Índia" was recorded. As for the teachers, all prepared presentations were related to the study of "Os Lusíadas" by Luís de Camões, and two address the same episode - the lyric-tragic episode of Inês de Castro.

Basically, spontaneous and prepared presentations differ in the degree of planning involved, the type of information communicated, the speakers' attention to the speech task and effort to speak clearly. In spontaneous presentations, the speakers can talk freely about any topic of their choice; they can change topic and move on to another topic whenever they feel like it. As far as typical (prepared) oral presentations at school are concerned, it was argued before that "more than talking about a pre-determined theme, an oral presentation presupposes the capacity to individually produce a greater amount of utterances, organizing the information that is given to the public in a clear structured form" [25].

The recordings of the two female teachers and all the students were done with an UHER 400 Report Monitor recorder with a BASF LPR 35 magnetic tape and a SENNHEISER MD 214 V-3 worn suspended from the neck microphone. These recordings were latter digitized at 44.1 kHz, using 16 bits/sample and afterwards downsampled to 16 kHz. CPE-FACES was recently extended with the recordings of a male teacher, using a TASCAM HD-P2, a Portable High Definition Stereo Audio Recorder, and a head-mounted microphone Shure, a Sub-miniature Condenser Head-worn Microphones, model Beta 54. The sound was recorded in mono, with 16-bit at samples rates of 44.1kHz, and afterwards downsampled to 16 kHz.

The subset of the corpus used in this study comprises 9 spontaneous presentations and 9 prepared unscripted presentations, from 6 teenage students (balanced by gender) and 3 teachers (2 female and 1 male). The data was split into train and test subsets, where the training part corresponds to about 75% of each speech file and the test part corresponds to the remaining 25%. Both subsets include portions of each speech file, which reduces the bias of training and testing using different conditions. Even so, one should note that the test set of our data is harder to automatically process, mostly due to the fact that by the end of a presentation interlocutors tend to interrupt more, leading to more overlapping speech. These rich interactions result in harder stretches of speech to process. The overall statistics about the subset of the corpus are presented in Table 1, where "total time" correspond to useful time and silences. The useful time is similar to other corpora subsets used by [23] for cross-language detection of prosodic events with AuToBI, namely the Italian subset (25 minutes of read speech by a single speaker). Having an equivalent amount of data (at this point a small sample of 27 minutes) allow us to make more direct comparisons with previous studies applying AuToBI.

The data was annotated by two expert linguists using the ToBI prosodic system adapted to European Portuguese (Towards a P_ToBI by [28]), in order to conduct experiments on automatic ToBI-labeling in European Portuguese, as part of the ongoing project[2] that funded this research. All the pitch accents (H+L*, H*+L, L*+H, L+H*, H*, L*, H+!H*) and the final boundary tones (L%, H%, !H%, LH%, HL%) that are covered in the Towards a P_ToBI proposal were used in this subset (see the schematic F0 contours in Figure 1). For further details on the annotation of the corpus, *vide* [26, 29].

3. Converting tone inventories

As previously said, AuToBI relies on the fundamentals of SAE. Therefore, as a pre-requirement to apply AuToBI to European Portuguese, the tonal inventory of P_ToBI, displayed in Figure 1, had to be converted converted to the SAE inventory. Since this work is our first step towards fully automatic prosodic labeling, at this stage no feature adaptation (pitch, intensity, spectral balance and pause/duration) was done. The conversion process was jointly carried out with one of the American co-authors of this paper. This task involved the analysis of a considerable amount of examples for each pitch accent and boundary tone in the P_ToBI inventory. Table 2 summarizes the conversions that have been made.

The main differences concern falling pitch accents and single boundary tones at intonational phrases (IPs). First, H+L* - a very frequent nuclear accent in European Portuguese (associated with declaratives and wh- question), which is absent in the SAE inventory. Second, H*+L - relatively uncommon in the subset analyzed, which is also absent in the SAE inventory. Third, H% and L% - single boundary tones at intonational

[2]PTDC/CLE-LIN/120017/2010.

P_ToBI	SAE
H+L*	H+!H*
H*+L	H*
L%	L-L%
H%	H-H%
	H-L%
HL%	!H-L%
!H%	H-L%
H*	H*
L*	L*
L+H*	L+H*
L*+H	L*+H
H+!H*	H+!H*
LH%	L-H%
L-	L-
H-	H-
!H-	!H-

Table 2: Inventory conversion from P_ToBI to SAE. In the first part the events changed and in the second part the ones kept.

phrases, as proposed by [30, 28]. In [30, 28], two different levels of phrasing are equated both to the intonational phrase: the major IP and the minor IP, in line with [31]. Minor and major IPs are marked with breaks 3 and 4, respectively, and the diacritics "-" and "%" represent the different strengths of the IP boundaries. Although [32] proposes a reanalysis marking both IP levels as "%", the labels "-" and "%" were kept in the subsets of CPE-FACES. Table 2 also displays the tones shared by both intonational systems (H*, L*, L+H*, L*+H, H+!H*, LH%, L-, H-, !H-).

4. Results

This section presents our first efforts towards the automatic detection of prominent prosodic events in European Portuguese. We have started by identifying prominent events using exclusively English models. The evaluation of such models was initially performed using the Portuguese data as a whole and later using the test set exclusively. Finally, we have trained new Portuguese models and verified their impact on prominence detection. Portuguese models were trained firstly using the whole training set (which includes students and teachers data) and later using exclusively teachers data. In all steps, results were evaluated using standard performance metrics [33]: precision, recall, f-measure, and accuracy, which can be expressed in terms of true positives (*tp*), true negatives (*tn*), false positives (*fp*), and false negatives (*fn*) as follows:

$$precision = \frac{tp}{tp+fp},$$

$$recall = \frac{tp}{tp+fn},$$

$$accuracy = \frac{tp+tn}{tp+fp+fn+tn},$$

$$Fmeasure = \frac{2 \times precision \times recall}{precision+recall}$$

4.1. Prominence detection using English models

After the conversion process described in Section 3, existing English (EN) models were applied to the Portuguese (PT) data. The English models used correspond to the 1.3 version, which includes training material from three corpora: Boston Directions Corpus, Boston University Radio News Corpus, Columbia Games Corpus ([17] and references therein). Building upon previous studies using AuToBI [23, 24], we hypothesized that prominence could be fairly detected by using English models on Portuguese data, with no further adaptations of the system rather than the conversions of tonal inventories.

EN models	Accented			Unaccented			
PT.all	Prec	Rec	F	Prec	Rec	F	Acc
Overall	80.8	64.9	72.0	70.0	84.1	76.4	**74.4**
Students	73.7	47.6	57.8	64.3	84.7	73.1	67.1
Teachers	83.4	73.8	78.3	74.3	83.8	78.8	78.5

Table 3: Prominence prediction, using EN models and PT data.

EN models	Accented			Unaccented			
PT.test	Prec	Rec	F	Prec	Rec	F	Acc
Overall	74.0	70.8	72.3	67.9	71.3	69.6	**71.0**
Students	61.1	63.1	62.1	55.8	53.8	54.8	58.8
Teachers	82.4	75.3	78.7	74.0	81.5	77.6	78.2

Table 4: Prominence prediction, using EN models and PT test set.

Table 3 presents the results of prominence prediction after applying the English models to the Portuguese data, discriminating between students and teachers. Results show that overall prominent events are detected with 72.0% of f-measure (80.8% of precision and 64.9% of recall). The accuracy of prominence prediction using English models is 74.4%. Thus, the results provide evidence that a considerable percentage of prominent events may be predicted from English to Portuguese, supporting previous research using AuToBI [23] for West-Germanic and Romance Languages, Portuguese not included. Table 3 displays a striking difference of around 11% accuracy between speakers (67.1 for students *vs.* 78.5% for teachers). Results are clearly better for teachers than for students, evidencing an age/status dependent effect on the prominence detection tasks.

English models were also applied to the test set of the Portuguese data exclusively. It is important to emphasize that, as previously stated, the test set of our data is harder to automatically process due mostly to very lively interactions between interlocutors. We believe that this is the core reason for the poorer results presented in Table 4 when compared with the ones in Table 3: the overall accuracy decreases by about 3% in the Portuguese test set. This result is mostly related to the overall prediction for students. For teachers the results are quite similar, with a slight improvement in prominence prediction (from 78.3% to 78.7% f-measure). To sum up, using exclusively the Portuguese test set, the overall accuracy decreases from 74.4% to 71% and the main differences between speakers still stand.

4.2. Prominence detection using Portuguese models

Previous results suggest that the prediction of prominence in Portuguese is fairly accounted for with English trained models. In this section, we target the training of Portuguese models and their impact on prominence detection. We expect that Portuguese trained models improve the overall accuracy of

PT models	Accented			Unaccented			
PT.test	Prec	Rec	F	Prec	Rec	F	Acc
Overall	78.8	78.6	78.7	75.4	75.6	75.5	**77.2**
Students	73.8	66.9	70.2	65.6	72.7	69.0	69.6
Teachers	81.3	85.4	83.3	82.1	77.3	79.6	81.6

Table 5: Prominence prediction, using PT models and PT test set.

Classified as →	Accented	Unaccented
Accented	613	165
Unaccented	167	511

Table 6: Confusion matrix for prominence prediction.

prominence detection, as language-dependent models usually enhance a system performance.

At this point, we have extended AuToBI capabilities to Portuguese by training Portuguese models with the Portuguese train set. No feature adaptation was performed, since we aim at evaluating the impact of language specific data integration in the overall accuracy. Table 5 summarizes the results achieved, showing an overall accuracy increase of around 6% in prominence prediction (77.2%), when compared with previous results obtained with English models (71.0%, see Table 4). Moreover, better performances are achieved for both students and teachers, with accuracy improvements of 11% for students and 3% for teachers. Comparatively, prominence detection show a higher gain for students than teachers, however, there are still striking inter-speaker differences in terms of age/status. Results show that training with language specific data improves prominence detection, in line with our expectations.

The confusion matrix in Table 6 shows prominence (mis)detection, displaying a higher percentage of non-prominent events classified as prominent (24.6% *vs.* 21.2%).

Along this work we have been pointing out striking differences between speakers, showing that teenage data is clearly more difficult to deal with than teachers data. Taking such differences into account, we have trained a new model using exclusively teachers training data. Such model allows us to perform more direct comparisons with previous experiments for other languages, which are commonly based on adult speech, either read or spontaneous speech (task-oriented). The training material for this final evaluation corresponds to 18.8 minutes (silences included) of teachers data only, aiming at producing a more homogeneous training data. In this final evaluation, the overall accuracy decreases to 59.7% (53.7% for students and 63.2% for teachers), which clearly demonstrates that the amount of training material is insufficient. Moreover, these results also show that is preferable to have large amounts of training material from English than having small language dependent samples (see Table 4). This also suggests that increasing the amount of Portuguese training data will likely lead to even better results than the ones already achieved (see Table 5).

The comparison between language dependent or cross-language models applied to the same test set is on its own very informative. Namely, we could verify that using models trained with large amounts of English data captures a considerable amount of prominent events. Our results for European Portuguese are closer to the ones presented in [23], which reported an accuracy ranging from 62.9% to 82% for West-Germanic and Romance languages, as French and Italian. This may suggest that a considerable amount of acoustic information is shared amongst different typological languages.

5. Conclusions and future work

This paper presented our first steps towards the extension of the AuToBI prosodic event detection system to European Portuguese. The first step concerned the prediction of prominent prosodic events based on models trained for English. In the second step the AuToBI models were retrained with a moderate sample of Portuguese ToBI annotated data. This training corpus includes both spontaneous and prepared unscripted speech from both adults and teenagers (teachers and students). This is not the typical starting point for such cross language experiments, but could not be avoided. As expected, the results showed poorer performance in prominence prediction for teenagers, contrasting with state-of-the-art results achieved for adults' data, regardless of the models used. This shows a clear age/status dependent effect present in our data.

Regarding the experiments conducted using the English models on Portuguese data, results showed an overall accuracy of 74.4% of prominence detection, comparable to state-of-the-art results [23]. This result further supports previous predictions using the same English models, which show that prominence detection is fairly accounted for one language to another [23]. When adapting AuToBI capabilities to train Portuguese models, as expected, the prediction of prominent events further improved in about 6% absolute. We may interpret those results as pointing out to two main directions: first, a considerable amount of prominent prosodic events may be cross-language predicted, even when tackling a sample of different typologic languages; second, albeit the reasonable cross-language prediction rate, there are language specific traits captured only with language-dependent trained models. This raises several research questions, *e.g,* on cross-language universal acoustic correlates of prominence vs. language dependent acoustic correlates; on genre free or genre dependent prosodic properties; on tonal density across languages; or even on informational structure. With this work we hope to contribute to prominence detection research, still scarcely studied.

Our preliminary results for European Portuguese are quite encouraging and a starting point to further enrich large amounts of (highly spontaneous) transcribed data available for our language. Therefore, future work will tackle the prediction of other prosodic events (phrasing and tonal boundaries), and the evaluation of the trained Portuguese models for different genres (university lectures, map-task dialogues, broadcast news, broadcast interviews, meetings, etc.). This will also allow us to address the study of age-specific and status-specific properties.

6. Acknowledgments

This work was supported by national funds through – Fundação para a Ciência e a Tecnologia, under projects PTDC/CLE-LIN/120017/2010 (COPAS) and PEst-OE/EEI/LA0021/2013, EU-IST FP7 project SpeDial under contract 611396, and by ISCTE-IUL, Instituto Universitário de Lisboa.

7. References

[1] S. Maskey and H. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech

summarization," in *Proc. of Interspeech*, Lisbon, Portugal, 2005.

[2] K. Chen and M. Hasegawa-Johnson, "How prosody improves word recognition," in *Proceedings of Speech Prosody 2004*, Nara, Japan, 2004, pp. 583–586.

[3] E. Shriberg, "Spontaneous speech: How people really talk, and why engineers should care," in *Proc. Eurospeech*, Lisbon, Portugal, 2005, pp. 1781 – 1784.

[4] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[5] H. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, no. 43, pp. 155–175, 2004.

[6] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[7] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tür, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, 2008.

[8] F. Batista, H. Moniz, I. Trancoso, and N. Mamede, "Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts," *Transactions on Audio Speech and Language Processing*, no. 20, pp. 474–485, 2012.

[9] H. Moniz, F. Batista, I. Trancoso, and A. I. Mata, "Automatic structural metadata identification based on multilayer prosodic information," 2013, DISS 2013.

[10] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proc. of Interspeech*, Antwerp, Belgium, 2007.

[11] G. Anumanchipalli, A. Black, and L. Oliveira, "Data-driven intonational phonology," *The Journal of The Acoustical Society of America*, vol. 134, pp. 4237–4237, 2013.

[12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challange," *Computer Speech and Language*, no. 27(1), pp. 4–139, 2013.

[13] S. Jun, *Prosodic typology: the phonology of intonation and phrasing*. Oxford University Press, 2005.

[14] P. Wagner, "Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates," in *Proc. of Interspeech*, Lisbon, Portugal, 2005.

[15] F. Tamburini and C. Caini, "An automatic system for detecting prosodic prominence in american english continuous speech," *International Journal of Speech Technology*, vol. 8, pp. 33–44, 2005.

[16] M. Swerts and E. Krahmer, "Facial expression and prosodic prominence: effects of modality and facial area," *Journal of Phonetics*, vol. 36, pp. 219–238, 2008.

[17] A. Rosenberg, "Automatic detection and classification of prosodic events," Ph.D. dissertation, University of Columbia, 2009.

[18] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, pp. 425–452, 2010.

[19] A. Windmann, I. Jauk, F. Tamburini, and P. Wagner, "Prominence-based prosody prediction for unit selection synthesis," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

[20] P. Prieto, M. Vanrell, L. Astruc, E. Payned, and B. Post, "Phonotactic and phrasal properties of speech rhythm. evidence from catalan, english, and spanish," *Speech Communication*, vol. 54(6), pp. 681–702, 2012.

[21] M. Mehrabani, T. Mishra, and A. Conkie, "Unsupervised prominence prediction for speech synthesis," in *Proceedings of Interspeech 2013*, Lyon, France, 2013.

[22] A. Rosenberg, "Autobi – a tool for automatic tobi annotation," in *Interspeech 2010*, 2010.

[23] A. Rosenberg, E. Cooper, R. Levitan, and J. Hirschberg, "Cross-language prominence detection," in *Proc. of Speech Prosody*, Shangai, China, 2012.

[24] V. Soto, E. Cooper, A. Rosenberg, and J. Hirschberg, "Cross-language phrase boundary detection," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[25] A. I. Mata, "Para o estudo da entoação em fala espontânea e preparada no Português Europeu," Ph.D. dissertation, University of Lisbon, 1999.

[26] A. I. Mata, H. Moniz, F. Batista, and J. Hirschberg, "Teenage and adult speech in school context: building and processing a corpus of European Portuguese," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[27] W. Labov, *Sociolinguistique*. Paris: Minuit, 1976.

[28] M. C. Viana, S. Frota, I. Falé, I. Mascarenhas, A. I. Mata, H. Moniz, and M. Vigário, "Towards a P_ToBI," in *Unpublished Workshop of the Transcription of Intonation in the Ibero-Romance Languages, PaPI 2007*, 2007.

[29] A. I. Mata, H. Moniz, T. Móia, A. Gonçalves, F. Silva, F. Batista, I. Duarte, F. Oliveira, and I. Falé, "Prosodic, syntactic, semantic guidelines for topic structures across domains and corpora," in *Proceedings of LREC 2014*, Reykjavik, Iceland, 2014.

[30] S. Frota, *Prosody and Focus in European Portuguese. Phonological Phrasing and Intonation.* New York: Garland Publishing, 2000.

[31] D. R. Ladd, *Intonational Phonology*. Cambridge University Press, 1986.

[32] S. Frota, "The intonational phonology of European Portuguese," in *Prosodic Typology II*, Sun-uh, Ed. Oxford: Oxford University Press, 2009.

[33] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA Broadcast News Workshop*, Herndon, VA, Feb. 1999.

Prosodic prominence detection in Italian continuous speech using probabilistic graphical models

Fabio Tamburini[1], *Chiara Bertini*[2], *Pier Marco Bertinetto*[2]

[1] FICLIT, University of Bologna, Italy
[2] Scuola Normale Superiore, Pisa, Italy
fabio.tamburini@unibo.it, c.bertini@sns.it, p.bertinetto@sns.it

Abstract

Prosodic prominence, a speech phenomenon by which some linguistic units are perceived as standing out from their environment, plays a very important role in human communication. In this paper we present a study on automatic prominence identification using Probabilistic Graphical Models, a family of Machine Learning Systems able to properly handle sequences of events. We tested the most promising members of such models on utterances selected from a manually annotated Italian speech corpus, obtaining very good recognition results crucially converging with the prominence detection responses provided by a pool of native speakers.

Index Terms: prosody, prominence, probabilistic graphical models, prominence annotation.

1. Introduction

A fairly uncontroversial definition of *prosodic prominence* due to Terken [29: 1768] states: "*prominence is the property by which linguistic units are perceived as standing out from their environment*". These prominent units typically contain relevant information for discourse and their correct perception is crucial for a successful communication strategy. Speakers use prominence to draw the listener's attention on specific point of the utterance, to express their emotion or attitude about the topic being discussed, to indicate the focus of an utterance, to mark the introduction of new topics, to indicate the information status of a word (new or given), to change speaking style, etc.

For all these reasons, the automatic management of prosodic prominence is crucial for both recognition and synthesis in order to build systems able to properly handle information in speech.

There is a long-standing agreement among scholars to consider the syllable as the prominence-bearing unit in connected speech. This position is not uncontroversial, however, and various studies analyse prominence at word level, especially if they mainly concern information extraction from speech utterances. In this paper we will consider the syllable, and its constituent units, as the relevant domain for prominence computation.

Several recent contributions handle prosodic prominence from a computational point of view, proposing different models (both Rule-Based and Machine Learning Systems - **MLS**) for the automatic detection of prominence in various languages, e.g. [2, 5, 11, 12, 15, 24, 25, 26]. Some of them are specifically devoted to Italian, or handle the identification of prosodic prominence in Italian among other languages [1, 8, 17, 26].

In this paper we present a procedure for the identification of prosodic prominence in Italian in the framework of MLS, based on training procedures that extract data and models from annotated corpora. These systems only take acoustic features into consideration, such as nucleus / syllable duration, energy measures in the nuclei / syllables and analysis of specific pitch profiles in the utterance.

Adopting the above reported definition [29], we consider prominence as a phenomenon establishing precise syntagmatic relations with respect to the neighbouring syllables. Its identification requires MLS able to properly model sequences of events, because the immediate context information, both in the feature sequence of the input and in the label sequence of the output, are crucial for the correct identification of syllable prominence. A syllable can be defined as prominent only by considering the relationships with other syllables, in line with the classical figure – ground contrast proposed by Gestalt psychology, rather than by considering it – and its features – in isolation.

Probabilistic graphical models (**PGM**) represent a class of MLS that, taking advantage of discriminative stochastic models, can successfully handle recognition problems that heavily depend on sequences. PGM, in some of their various configurations / models, have been applied to the task at stake with encouraging results [8, 21]. Despite these findings, PGM and their complex family of model variations have not yet been extensively applied to this problem, especially considering hidden or latent dynamics detection in speech data and the possibility of extracting and using high order relations among the acoustic features.

2. Probabilistic graphical models

PGM are powerful frameworks for representation and inference in multivariate probability distribution. They use a graph-based representation as the basis for compactly encoding a complex distribution over a high-dimensional space representing the conditional dependence structure between random variables.

In this paper we considered some of the most powerful and widespread discriminative models to identify prosodic prominence in continuous speech, as well as some recently presented new models.

PGM consists of a large family of different methods that constrain the graph structure in specific ways. Conditional Random Fields (**CRF** - see [10, 22] for general introductions) are no doubt the most used PGM in various fields. However, most CRF models use linear functions to represent the relationships between input features and the classification output and a simple graph structure for the entire model. This way of coding relations presents severe limitations for real-world applications, because: (a) in many cases the

relationships between inputs and outputs are complex and nonlinear, and (b) some problems require modelling relevant sub-structures in the label sequence.

In this work we used different PGM, each addressing in its own way the shortcomings of CRF models, considered as a baseline. Conditional Neural Fields (**CNF**) [20], inserting a small neural network between input and output, are able to capture the nonlinearities required by constraint (a) above; Latent-Dynamic Conditional Random Fields (**LDCRF**) [18], in turn, can learn latent sub-structures in output class labels. Latent-Dynamic Conditional Neural Fields (**LDCNF**) [16] can combine the advantages of both previous approaches in a single model.

PGM are able to manage sequences of input-output data predicting the output sequence considering both the input feature configuration, in a specific window centered on the generic input vector of features x_j, and the previous output sequence. Figure 1 outlines the different structures of the PGM used in this work.

Given the input sequence of local features $x_1,\ldots,x_n$, typically consisting of vectors of features, and given the output sequence $y_1,\ldots,y_n$, linear CRF assign the most likely label to output y_j conditioned by the feature vectors belonging to the local window and the previous output label y_{j-1}.

CNF extends CRF by adding one level of gate units, acting as a neuron tier (more precisely a perceptron), between the input and the output layers. These gate neurons are a sort of feature extractor able to capture nonlinear relationships between input and output.

A further, completely different way of extending CRF is implemented in LDCRF adding a layer of hidden-state units between input and output layers: these units are able to model the sub-structure of the label sequence and can learn complex dynamic behaviours between output labels.

Finally, LDCNF take advantage from both approaches, combining them in a unique structure. The LDCRF model is modified by adding the neural network introduced in CNF between input and hidden units; in this way, LDCNF models can both identify sub-structures in the output sequence and learn nonlinear relationships between input feature vectors and output class labels.

For lack of space, we refer to the cited papers for all mathematical and algorithmic details that define these approaches, especially for what concerns the learning algorithms, inference and parameter setting.

3. Corpus building and data collection

The materials used in this study were utterances extracted from the API/AVIP corpus [3] and from a selection of sentences read by a subset of the same speakers. The corpus consisted of semi-spontaneous conversations between native Italian speakers elicited with the map-task method for different Italian language varieties. The speakers used for the present purpose were from the Pisa area (central Italy).

The selected utterances presented a neutral intonation contour, without emphatic stress or pauses, and presenting at least 8 syllables. Care was taken to avoid any disturbing phenomena such as speakers' overlap, laughters, background noise, etc.

A perception experiment, divided into two different sessions (sets A-C vs D-G in table 1), was carried out using 120 selected corpus utterances (90 spontaneous and 30 read), produced by female and male speakers. The average utterance length was 18 syllables, ranging from 9 to 35. The task was performed by 35 Italian native speakers.

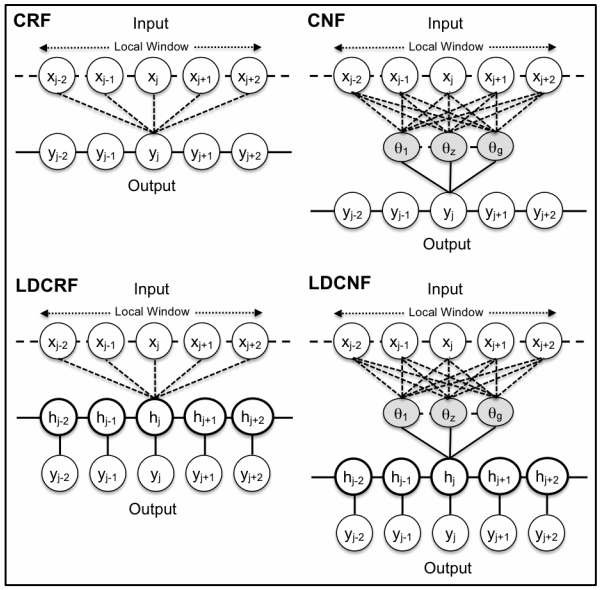


Figure 1. *The various PGM considered in this study. The gate units $\theta_1 \ldots \theta_g$ are in gray while the hidden units $h_1 \ldots h_n$ present a thick border. For clarity, only the region of the model net surrounding the generic input feature vector x_j is represented in the pictures.*

Table 1*: The latin square scheme applied to the 120 utterances composing the corpus.*

Utterance Sub-list	Annotators IDs
A (20 utt.)	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
B (20 utt.)	1, 2, 3, 4, 5, 11, 12, 13, 14, 15
C (20 utt.)	6, 7, 8, 9, 10, 11, 12, 13, 14, 15
D (15 utt.)	16, 17, 18, 19, 20, 21, 22, 23, 24, 25
E (15 utt.)	16, 17, 18, 19, 20, 26, 27, 28, 29, 30
F (15 utt.)	31, 32, 33, 34, 35, 21, 22, 23, 24, 25
G (15 utt.)	26, 27, 28, 29, 30, 31, 32, 33, 34, 35

The experimental task was to identify the sentence prominences. The participants could listen to each sentence as many time as they wanted. To reduce the task difficulty, participants were presented with a transcription of the given sentence whereby each syllable (as the possible prominence-carrying unit) was separately indicated; in addition, the lexically stressed syllables were explicitly pointed out. However, participants were warned during the training phase that not all lexically stressed syllables were actual targets of sentence prominence, while prominence could also land on lexically unstressed syllables. As soon as the participant had made her/his own choice by clicking on the square corresponding to the intended syllables, s/he was immediately presented with another sentence.

Since this task is very demanding in terms of attention, the sentences were divided into sub-lists according to a latin square scheme, so that each utterance was judged by 10 speakers. As a consequence, no participant heard/read all the sentences. Table 1 depicts the annotation scheme: the total number of utterances was divided into seven sub-lists each assigned to 10 annotators.

For each test utterance extracted from the AVIP corpus, the phonetic transcription and the phoneme level segmentation were available in the source. The selected utterances were further segmented manually in order to identify the syllable boundaries.

3.1. First-step: data annotation and overall judgment convergence

As a first step, the participants judgments were pooled together and evaluated with respect to the degree of convergence relative to the identification of any given syllable as prominent. The convergence level was assessed with respect to four criteria (60%, 70%, 80% and 90%), indicating the percentage of shared prominence identification. In practice, considering that each sentence was judged by 10 participants, the 60% criterion implied the convergence of 6 out of 10 listeners (and similarly for the other levels).

Needless to say, the 90% agreement level involves a smaller number of prominent syllables within any given sentence, since almost all participants have to agree on their judgment concerning the given syllable. By contrast, the more generous 60% level concerns a larger number of syllables. Interestingly, there was full agreement as for the last prominence of each sentence, evidently due to unequivocal durational cues, although the energy and frequency levels at the end of an utterance are usually fairly low.

As is well known, while the identification of emphatic prominences is undisputable, there usually is considerable divergence among human judges on the identification of non-emphatic prominences. As a consequence, the 80% level was selected as bench-mark for the automatic detection of prominences as a first approximation. The dataset contains 480 prominent syllables out of 2037, thus close to one prominent syllable out of four (23.56%).

3.2. Second-step: best annotators selection

Capitalizing on the well-known lack of overwhelming convergence among human judges as for the localization of sentence prominences, a second type of comparison between the automatic detector and the human judges was adopted. For each of the two sub-lists of utterances, the three most reliable judges were selected, i.e. the three participants presenting the highest level of mutual agreement according to the Fleiss K index (this schema will be referred to as the "best-3" annotation agreement). Subsequently, the syllables judged as prominent by the majority of the "best-3" were considered prominent. The correlation data are reported in Table 2. In this dataset, 33.46% of the syllables are prominent (one out of three).

4. Acoustic features

The acoustic features used in this study are the same used in some previous studies of one of the authors [25, 26]. These works proposed a rule-based system resting on four acoustic features that exhibited good performances in prominence detection. One of the major challenges in predicting syllable prominence is the correct identification of the various sources of influence, such as: fundamental frequency excursions, duration, intensity-related parameters and listeners' linguistic expectancies.

The automatic prominence detection system described in [25, 26] is based on the global prominence model proposed by Kohler [13, 14]. In his view, there are two main 'actors', at the linguistic-prosodic level, playing a relevant role in supporting sentence prominence. The first, *pitch accent*, coincides with a concept first introduced by Bolinger [4] and concerns specific movements in F0 profile. The second, *force accent*, is completely independent from the intonational profile and is connected with different acoustic phenomena, such as intensity (or spectral emphasis), segmental durations and possibly others. Both 'actors' seem to play a relevant role in supporting prominence perception at utterance level, mutually reinforcing each other.

Table 2*: The "best-3" annotators for each utterance sub-list.*

Utterance Sub-list	Best 3 annotators	Fleiss-K
A	2, 3, 6	0.875631
B	1, 2, 13	0.876340
C	6, 9, 13	0.859735
D	16, 18, 20	0.836450
E	16, 19, 20	0.857646
F	31, 33, 25	0.863555
G	31, 33, 34	0.812559

In the present study, we considered the four features used in the cited work (reported in Table 3 with brief reference to their actual computation) and added one further acoustic feature, namely syllable duration, following the good results obtained in [8]. All these features, except syllable duration, are computed within the syllable nucleus domain. Thus, using the phonetic and syllabic segmentation provided in the source corpus, all we had to do was to define the duration of the syllabic nuclei, deriving it automatically from the other two measures.

5. Results and discussion

We made a number of experiments, considering various PGM and different parameter configurations in order to maximize the agreement between the automatic procedure and the human annotators. We tested the best system on the above-described corpus, applying a random sub-sampling validation to define the training and test set (respecting a 5/1 proportion, 100/20 utterances), repeating this procedure 20 times and averaging the obtained results.

The best performances so far obtained, in comparing the automatic classifications with the gold standard, are depicted in Table 4 and 5. The first table refers to the "80%" level of annotation agreement described in section 3.1, the second refers to the "best-3" agreement described in section 3.2. There is a clear performance improvement considering the "best-3" annotation schema: the larger inter-human agreement produces a more consistent annotation and, thus, better performances of the MLS trained on these data.

The accuracy obtained by our automatic systems is very high, considering that the typical inter-human agreement accuracy reported in the literature, when annotating prominence by means of two levels (prominent vs non-prominent) is in the range 70-90%.

However, considering that the distribution between the two prominence classes is rather skewed, one should best adopt the

F-measure as a more reliable metric of the actual system performance. An F-measure of 0.770 is quite high, considering that:

a) The training corpus used to set up the model, in the various permutations composing the 20-random-sub-sampling validation, is rather small to properly train a MLS, as it only contains 100 utterances (1800 syllable, on average);
b) We only used acoustic information and did not consider any linguistic feature that might improve the system's behaviour, as showed, for example, by [21].

Table 3. *Acoustic features used to set up the PGM models for prominence identification.*

Acoustic Feature	Description
Nucleus Duration	Duration of the syllable nucleus normalised w.r.t. mean and variance duration of the syllable nuclei in the utterance (z-score), as based on the manual segmentation available in the database.
Spectral emphasis	Normalised SPLH-SPL parameter [9] (z-score).
Pitch movements	Computed as the product of A_{event} and D_{event} parameters of the TILT model representation [28] of pitch movements. The raw pitch contour is the median of three pitch tracking algorithms [27]: RAPT [23], SWIPE' [6] and YAAPT [31]. The raw pitch profile was stylised by using a quadratic spline function, interpolating the control points derived from the OpS algorithm proposed in [19].
Overall intensity	RMS energy computed in the frequency band 50-5000 Hz, normalised to mean and variance of intensity inside the utterance (z-score).
Syllable Duration	The same as the nucleus duration but referred to the entire syllable.

Table 4. *Results obtained by the PGM tested, in terms of Accuracy, Precision & Recall and F-Measure, as referred to the 80% level of annotation agreement described in section 3.1. The parameters considered with the various PGM are:* w *– local window size (symmetric),* h *– number of hidden units,* g *– number of gate neurons and* α *– regularization factor.*

Model	Parameters	Results			
		Acc.	Prec.	Rec.	F
SVM	w=2, C=0.5	0.858	0.765	0.592	0.665
CRF	w=1	0.856	0.735	0.609	0.665
LDCRF [18]	w=1, h=2	0.856	0.720	0.640	0.676
CNF [16]	w=2, g=40 α=0.5	0.871	0.784	0.642	0.705
CNF [20]	w=1, g=20	0.872	0.769	0.667	0.713
LDCNF [16]	w=1, h=4, g=40, α=0.5	**0.875**	**0.788**	**0.658**	**0.716**

Table 5. *Results obtained by the PGM tested, as referred to the "best-3" level of annotation agreement described in section 3.2. The parameters are described in the caption of Table 4.*

Model	Parameters	Results			
		Acc.	Prec.	Rec.	F
SVM	w=1, C=50	0.833	0.791	0.681	0.732
CRF	w=2	0.838	0.795	0.695	0.741
LDCRF [18]	w=1, h=2	0.842	0.792	0.713	0.750
CNF [16]	w=1, g=20 α=0.5	0.845	0.803	0.712	0.754
LDCNF [16]	w=1, h=4, g=20, α=0.5	0.851	0.823	0.706	0.759
CNF [20]	w=1, g=20	**0.855**	**0.831**	**0.718**	**0.770**

Our results cannot be directly compared with other similar studies, because there are no standard corpora for evaluation, both in general and for Italian in particular, nor specific standardised metrics. In any case, it is worth observing that the best-obtained results are equivalent or better than those of the already cited studies (e.g. [8, 15, 17]).
The best PGM for the problem at hand seems to be CNF in the implementation proposed by [20]. LDCNF and CNF from [16] obtained slightly lower performances, especially using the "best-3" annotation schema. This is probably due to the small set of utterances used to train these models, since while performing some other tests not reported here, using more utterances and a different corpus, LDCNF performed best.
In order to compare the PGM results with standard non-sequential MLS, we included in our experiments the results obtained using classical Support Vector Machines (**SVM**). All PGM exhibit significant performance improvements when compared with SVM, confirming their superiority when applied to intrinsically sequential problems.

6. Conclusion

This paper presents some experiments on the automatic detection of prosodic prominence in continuous Italian speech. Considering that in order to properly define prosodic prominence one needs to take contextual information into account, we tested a number of versions of MLS, able to correctly manage problems involving sequences of input features and sequences of output label classes, to be related in a complex way. In particular, we tested MLS belonging to the large family of PGM.
We thus performed several experiments with CRF, CNF, LDCRF and LDCNF models, obtaining very good classification results (F-measure = 0.770) despite using a small Italian corpus consisting of only 120 utterances.
Considering that, as outlined by [7, 30], prominence perception is highly influenced by the listener's linguistic expectations, there is room for large improvements in the system's performance by including linguistic features in the automatic system.
We are planning to test these models on different corpora and different languages, in order to verify the effectiveness of the proposed approach to automatic prominence detection.

7. References

[1] Abete, G., Cutugno, F., Ludusan, B., Origlia, A., "Pitch behavior detection for automatic prominence recognition", in *Proc. of Speech Prosody 2010*, 2010, Chicago.

[2] Al Moubayed, S., Beskow, J., "Prominence detection in Swedish using syllable correlates", in *Proc. of Interspeech 2010*, Makuhari, Japan, 2010.

[3] Albano Leoni F., "Tre progetti per l'italiano parlato: AVIP, API, CLIPS", in Maraschio N., Poggi Salani T., (eds.), *Italia Linguistica. Anno mille, anno duemila, Atti del XXXIV Congresso della Società di Linguistica Italiana (SLI)*, Roma, Bulzoni, 675-683, 2003.

[4] Bolinger, D., "A theory of pitch-accent in English", *Word*, 14, 1958, 109-149.

[5] Brenier, J.M., Cer, D.M., Jurafsky, D., "The detection of emphatic words using acoustic and lexical features", in *Proc. of Interspeech 2005,* Lisbon, 2005, 3297–3300.

[6] Camacho A., *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, PhD Thesis, University of Florida, 2007.

[7] Cole, J., Mo, Y., Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence". *Laboratory Phonology*, 1, 2010, 425–452.

[8] Cutugno, F., Leone, E., Ludusan, B., Origlia, A., "Investigating syllable prominence with conditional random fields and latent-dynamic conditional random fields", in *Proc. of Interspeech 2012*, Portland (OR), 2012.

[9] Fant G., Kruckenberg A., Liljencrants, J., "Acoustic-phonetic Analysis of Prominence in Swedish", in Botinis, A. (Ed.), *Intonation*, Kluwer Academic Publisher, 2000, 55–86.

[10] Fosler-Lussier, E., Yanzhang, H., Preethi, J. and Prabhavalkar, R., "Conditional Random Field an Speech, Audio and Language Processing", in *Proc. IEEE*, 101(5), 2013, 1054–1075.

[11] Goldman, J-P., Avanzi, M., Auchlin, A., Simon, A.C., "A continuous prominence score based on acoustic features", in *Proc. of Interspeech 2012*, Portland (OR), 2012.

[12] Kocharov, D., "Automatic detection of prominent words in Russian Speech", in *Proc. of the IEEE International Multiconference on Computer Science and Information Technology*, 2010, 435–438.

[13] Kohler, K.J., "Neglected categories in the modelling of prosody - Pitch timing and non-pitch accents", in *Proc. of ICPhS'03*, Barcelona, 2003, 2925-2928.

[14] Kohler, K.J., "Form and Function of Non-Pitch Accents", *Prosodic Patterns of German Spontaneous Speech, AIPUK*, 35a, 2005, 97-123.

[15] Li, K., Zhang, S., Li, M., Lo, W-K., Meng, H., "Prominence model for prosodic features in automatic lexical stress and pitch accent detection", in *Proc. of Interspeech 2011*, Florence, 2011, 2009–2013.

[16] Levesque, J.C., Morency, L.P. and Gagné, C., "Sequential emotion recognition using Latent-Dynamic Conditional Neural Fields", in *Proc. 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, Shanghai, 2013, 1–6.

[17] Ludusan, B., Origlia, A., Cutugno, F., "On the use of the rhythmogram for automatic syllabic prominence detection", in *Proc. of Interspeech 2011*, Florence, 2011, 2413–2417.

[18] Morency, L., Quattoni, A. and Darrell, T., "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, Minnesota, 2007, 1–8.

[19] Origlia, A., Abete, G. and Cutugno, F., "A dynamic tonal perception model for optimal pitch stylization", *Computer Speech & Language*, 27(1), 2013,190–208.

[20] Peng, J., Bo, L. and Xu, J., "Conditional Neural Fields", in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009.

[21] Shridar, V.K.R, Nenkova, A., Narayanan, S., Jurafsky D., "Detecting prominence in conversational speech: pitch-accent, givenness and focus", in *Proc. of Speech Prosody 2008,* Campinas, Brazil, 2008.

[22] Sutton, C. and McCallum, A., "An Introduction to Conditional Random Fields", *Foundations and Trends in Machine Learning*, 4(4), 2011, 267–373.

[23] Talkin, D., "A robust algorithm for pitch tracking (RAPT)", in W.B. Kleijn & K.K. Paliwal (Eds.), *Speech coding and synthesis*, New York: Elsevier, 1995, 495–518.

[24] Tamburini F., "Reliable Prominence Identification in English Spontaneous Speech", in *Proc. of Speech Prosody 2006*, Dresden, 2006, PS1-9-19.

[25] Tamburini F., Wagner P., "*On Automatic Prominence Detection for German"*, in *Proc. of InterSpeech 2007*, Antwerp, 2007, 1809–1812.

[26] Tamburini F., "Prominenza frasale e tipologia prosodica: un approccio acustico", in *Proc. Linguistica e modelli tecnologici di ricerca, XL congresso internazionale di studi*, Società di Linguistica Italiana, Vercelli, 2009, 437–455.

[27] Tamburini F., "Una valutazione oggettiva dei metodi più diffuse per l'estrazione automatica della frequenza fondamentale", in *Atti del IX Convegno dell'Associazione Italiana Scienze della Voce*, Venice, in press.

[28] Taylor, P.A., "Analysis and Synthesis of Intonation using the Tilt Model", *J. Acoust. Soc. Amer.*, 107(3), 2000, 1697–1714.

[29] Terken, J., "Fundamental Frequency and Perceived Prominence", *Journal of the Acoustical Society of America*, 89, 1991, 1768–1776.

[30] Wagner, P., "Great Expectations – Introspective vs Perceptual Prominence Ratings and their Acoustic Correlates", in *Proc. of Interspeech 2005*, Lisbon, 2005, 2381–2384.

[31] Zahorian, S.A., Hu, H., "A Spectral/temporal method for Robust Fundamental Frequency Tracking", *Journal of the Acoustical Society of America*, 123(6), 2008, 4559–4571.

Integrating variability in loudness and duration in a multidimensional model of speech rhythm: Evidence from Indian English and British English

Robert Fuchs[1]

[1] Englisches Seminar, Westfälische Wilhelms-Universität Münster

robert.fuchs@uni-muenster.de

Abstract

Most research on speech rhythm has focussed on duration. For example, [1] suggested the normalised pairwise variability index for vocalic intervals (nPVI-V) in order to measure the variability of vocalic durations. This paper argues that speech rhythm research should also take into account other correlates of prominence as well as their interaction. The duration-based nPVI, or nPVI-V(dur), is supplemented by an nPVI that measures variability in average loudness, nPVI-V(avgLoud). These two metrics account for variability in duration and loudness, but cannot measure if loudness and duration reinforce each other by varying simultaneously in the same direction. This simultaneous variability is accounted for by the combined nPVI-V(dur+avgLoud), which is higher than the average of the other two measures if vocalic intervals that are longer than average are also louder than average. The three metrics are subsequently applied to recordings of a reading task performed by 20 speakers of Indian English (IndE) and 10 speakers of British English (BrE). Results indicate that IndE has less variability in duration and less variability in loudness than BrE. In addition, IndE has less simultaneous variability in duration and loudness than BrE. This indicates that duration and loudness are less often used together as cues to prominence in IndE compared to BrE.

Index Terms: speech rhythm, duration, loudness, Indian English, British English

1. Introduction

Early definitions of speech rhythm held that languages fall into discrete rhythm classes: English, for example, was said to be stress-timed [2], as opposed to the syllable-timed rhythm of French [3]. However, a more nuanced view of varieties of English around the world resulted in descriptions of only some, mainly more established varieties such as British English (BrE) and American English, as stress-timed. By contrast, many of the younger national varieties of English (such as Indian, Nigerian and Singapore English), which are used in administration, education and as a national link language in their respective countries, have been classified as syllable-timed. This is often explained with transfer from the local syllable-timed languages [4–6].

The categorical view of rhythm with two separate classes of languages has in the last years given way to a gradient analysis. Accordingly, languages can be placed at any point on a continuum between a prototypically stress-timed and a prototypically syllable-timed pole. This analysis is supported by quantifications of rhythm based on the durations of vocalic intervals, consonantal intervals and syllables, and their variability, as suggested by [1,7–15]. One widely used rhythm metric is the normalised pairwise variability index for vocalic intervals, or nPVI-V [1]. It is computed by calculating the mean of the differences between successive vocalic intervals divided by their sum, multiplied by 100:

$$nPVI-V = 100 \times \frac{\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right|}{m-1}; \qquad (1)$$

where m is the number of vocalic intervals

and d_k is the duration of the kth vocalic interval

Some of these rhythm metrics have been shown to predict language discrimination in perception experiments [16]. Comparisons of vocalic and consonantal metrics by [17, 18] suggest that vocalic, speech rate normalised metrics are the most reliable measures of speech rhythm. However, even these metrics are influenced by variation between speakers, texts and transcribers. It is therefore recommended to use phonetically balanced or large samples from a greater number of speakers, and adhere to a clearly defined set of transcription rules. How crucial these suggestions are is demonstrated by [19], who was unable to replicate the results of [17, 18], perhaps because (1) she relied on samples from each language that were not representative but specifically designed to elicit sentences that differed from each other in rhythm as much as possible, and (2) because segmentation criteria and the treatment of hesitations were not specified and possibly not controlled for.

However, an approach that measures rhythm on the basis of durational variability alone may present a lop-sided account of speech rhythm. Most studies on rhythm in the last 15 years have concentrated exclusively on duration, which means that they considered only one correlate of prominence. Exceptions are [20–23], who suggested rhythm metrics based on variability in intensity, f_0, and sonority. Together, these metrics may form a multidimensional account of rhythm, as demanded by [24–26]. A comparison of two languages might show that one has less variability in all these correlates of prominence and is therefore more syllable-timed than the second language on all these dimensions. On the other hand, the first language might also turn out to have less variability in intensity, but more in duration than the second language. Any language might therefore have multiple co-existing and different rhythms [26], which contribute in different ways to a succession of elements relatively similar in prominence (syllable-timing) or relatively different in prominence (stress-timing).

2. Integrating variability in loudness and duration

While a consideration of multiple correlates of prominence contributes to a multidimensional analysis of rhythm, the metrics suggested so far rely exclusively on acoustic correlates, not perceptual correlates of rhythm. This is particularly important for the relation of intenstiy, an acoustic property, to loudness, a perceptual property. Sounds with equal intensity but different frequencies have been shown to differ in loudness [27, 28]. Therefore, this paper suggests that loudness instead of intensity should be considered as a perceptual correlate of rhythm.

A second shortcoming of a multidimensional analysis of rhythm as outlined above is that different acoustic or perceptual correlates are considered separately. In addition to an individual analysis of these correlates, their interaction should also be taken into account. For example, if stressed vowels are both longer and louder than unstressed vowels, then variability in duration and loudness may be said to reinforce each other. By contrast, if stressed vowels are only longer but not usually louder than unstressed vowels, there is no reinforcement between these two cues of prominence.

In order to derive a measure of loudness, the computer programme Praat can be used [29]. After extracting the relevant vocalic interval from the recording, first its spectrum, then excitation, and finally maximum loudness in Sone can be derived (a Praat script is available from the author upon request). As a measure of variability in loudness, a PVI can be computed by entering these loudness values (instead of duration values) into formula (1). This metric will be called nPVI-V(avgLoud), and the duration measure will be referred to as nPVI-V(dur).

The next step, accounting for simultaneous variability in loudness and duration necessitates a change to the PVI formula. For every pair of successive vocalic intervals, the difference in duration and in loudness are computed separately and divided by their sum. These relative differences in duration and loudness are then added and squared. Squaring this sum of relative differences in duration and loudness gives an advantage (higher values) to simultaneous in- or decreases in both correlates of prominence, and a disadvantage to cases where duration in- but loudness decreases (or the other way around).

$$nPVI-V(dur+avgLoud) = \quad (2)$$
$$100 \times \frac{\sum_{k=1}^{m-1} 2 \times \left(\frac{d_k - d_{k+1}}{d_k + d_{k+1}} + \frac{l_k - l_{k+1}}{l_k + l_{k+1}} \right)^2}{m-1};$$

where m is the number of vocalic intervals,

d_k is the duration of the kth vocalic interval

and l_k is the loudness of the kth vocalic interval

For example, if the second vocalic interval is twice as long and twice as loud as the first, then the square of the differences is $(\frac{1-2}{1+2} + \frac{1-2}{1+2})^2 = 0.444$. By contrast, if the second vocalic interval is twice as long as the first, but equal in loudness to the first, then the square of the differences is lower: $(\frac{1-2}{1+2} + \frac{1-1}{1+1})^2 = 0.111$. And if the second vocalic interval is twice as long as the first, but half as loud as the first, the square of the differences is zero: $(\frac{1-2}{1+2} + \frac{1-0.5}{1+0.5})^2 = 0$. The measure therefore accounts for whether loudness and duration increase (or decrease) simultaneously, thereby reinforcing each other in the generation of prominence, or whether one decreases and the other increases, which causes them to offset each other in the generation of prominence.

3. Indian English

IndE is a postcolonial variety of English used mainly in public contexts such as education, administration, business and politics, but also by Indians who travel or reside in a region whose local language they do not speak. English is the primary domestic language for only a small minority, although many others use it at home when discussing topics belonging to the public domain, as for example when a parent asks their child what happened at school that day. Around 23 % of the population of India have at least basic knowledge of English, and 4 % are fluent [30]. Based on the 2011 census [31], this means there are 50 million fluent speakers.

Although standard IndE still lacks full official recognition, it is the de facto standard taught in schools and universities is standard IndE [32, 33]. Nevertheless, this standard has not yet been fully codified, and in such a context the kind of language used by educated speakers can be used as a yardstick for what is considered acceptable by the speech community [34–36]. The phonology of IndE differs in several respects from that of BrE [37–42], likely due to transfer from Indian languages [43]. While they belong to several different language families, these languages have converged over time in several respects and form a *sprachbund* [44–46].

IndE has been suggested to be syllable-timed or more syllable-timed than BrE [32, 43, 47–52]. On the basis of these descriptions, the present paper hypothesises that IndE has lower variability of vocalic durations, lower variability in average loudness in vowels, and lower simultaneous variability in duration and loudness in IndE compared to BrE. This suggestes that nPVI-V(dur), nPVI-V(avgLoud) and nPVI-V(dur+avgLoud) are all smaller in the IndE group than in the BrE group. Additionally, it is expected that duration and loudness as correlates of prominence also reinforce each other less often in IndE than in BrE, even when lower levels of variability in loudness and duration (considered separately) are taken account of. This suggests that the difference between the group means for IndE and BrE in nPVI-V(dur+avgLoud) is greater than both the differences between the group means in nPVI-V(dur) and nPVI-V(avgLoud).

4. Data

Recordings of a text read by 10 speakers or Standard Southern BrE and 20 speakers of IndE were used. The BrE data was taken from the DyViS database [53]. The IndE speakers were recorded by the present author reading the same text. All speakers were university students at the time of recording. The IndE speakers were equallly divided between four different L1 groups, and had either Hindi or Bengali (both Indo-Aryan languages), or Telugu or Malayalam (both Dravidian languages) as first languages. L1 was determined on the basis of a sociolinguistic interview involving questions on when speakers first started using what language. They were between 20 and 28 years of age, and, with the exception of one speaker each, had exclusively attended English-medium schools and universities, and had not resided outside of India.[1]

[1] Only one speaker did not attend English-medium schools throughout, but went to a Telugu-medium primary and English-medium secondary school. Another speaker spent several years in an Arab country, but had only interaction with the local South Asian expat community in

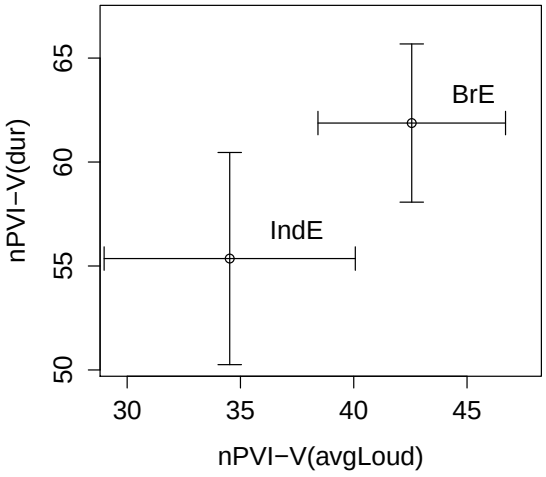


Figure 1: Group averages and standard deviations for variability in average loudness (x-axis) and in duration (y-axis).

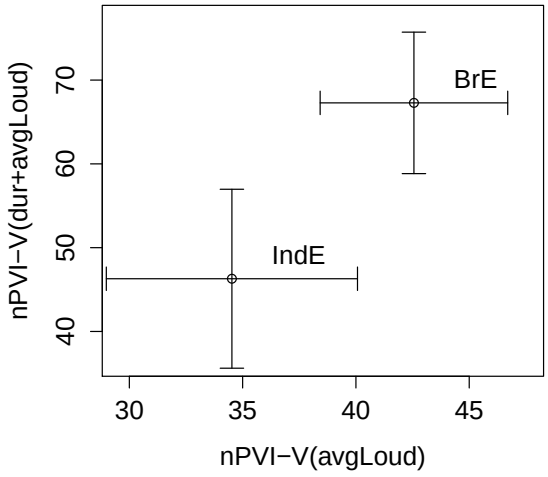


Figure 2: Group averages and standard deviations for variability in average loudness (x-axis) and simultaneous variability in duration and average loudness (y-axis).

Two thirds of the reading passage, or 392 words, were segmented according to the recommendations provided by [17, 18, 54]. nPVI-V(dur), nPVI-V(avgLoud) and nPVI-V(dur+avgLoud) were computed individually for utterances comprising at least three vocalic intervals, excluding the final interval. Next, for each individual speaker, the median of the rhythm scores of all utterances produced by this speaker was computed.

5. Results

Variability in duration was significantly lower in the IndE group (55.4) than in the BrE group (61.9, $p<0.0001$, t-test). Average variability in duration was 10.5 % lower in IndE than in BrE. Variability in average loudness was also signifcantly lower in the IndE group (34.5) than in the BrE group (42.7, $p<0.0001$). Average variability in loudness was 19.0 % lower in IndE than in BrE. The average values for variability in loudness and duration for both groups are shown in figure 1.

Scores for individual speakers are shown in figure 3, with crosses for IndE and circles for BrE speakers. Using both metrics, separation between the two groups is relatively good, although there is some overlap between the two groups. Two BrE speakers have less variability in duration and loudness (nPVI-V(dur) and nPVI-V(avgLoud)) than the other BrE speakers and are close to some of the IndE speakers. Likewise, there are two IndE speakers who have high values for variability in duration and loudness so that they are similar in this respect to the BrE group.

Combined variability in duration and average loudness was significantly lower in IndE (46.3) than in BrE (67.3, $p<0.0001$). The average combined variability in duration and loudness was 31.2 % lower in IndE than in BrE, which means that the difference in combined variability in duration and loudness (nPVI-V(dur+avgLoud)) was higher than the variability in either variability in duration (nPVI-V(dur)) or average loudness (nPVI-V(avgLoud)). The average values for variability in loudness and simultaneous variability in duration and loudness for both groups are shown in figure 2.

Scores for individual speakers are shown in figure 4. Considering simultaneous variability in loudness and duration, there are again two IndE speakers (crosses) with values for variability similar to the BrE group. There is also one BrE speakers (circles) with a variability slightly lower than the IndE speakers with the highest variability.

6. Discussion

The results have shown that IndE has

- significantly less variability in duration
- significantly less variability in loudness and
- significantly less simultaneous variability in duration and loudness

than BrE. Crucially, the difference in simultaneous variability in duration and loudness between IndE and BrE was higher than either the difference in variability in loudness or duration. This result shows that a measure of simultaneous variability in duration and loudness, the nPVI-V(dur+avgLoud) suggested in this paper, captures an important aspect of variability in prominence between successive vocalic intervals. If variability in duration on the one hand, and variability in average loudness on the other hand, were both randomly distributed, one would expect that the difference between IndE and BrE in the simultaneous variability in loudness and duration were equal to the average of the difference between the two varieties in loudness, and in duration (i.e. 14.75 %). However, this was not the case. Instead, the difference in combined variability in loudness and duration turned out to be much higher than the variability in either loudness or duration taken separately. This result also implies that in BrE, vocalic intervals that are longer than average also tend to

her daily routine.

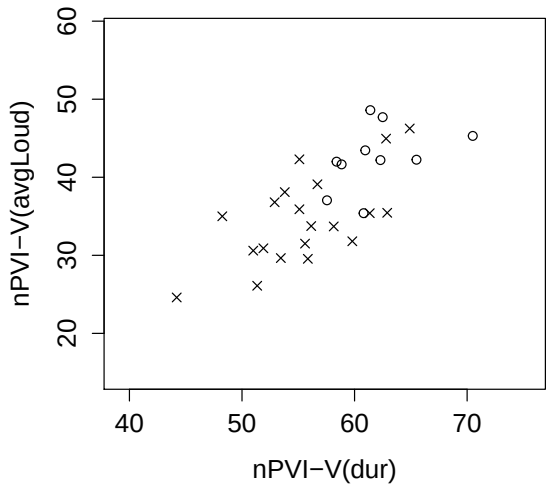


Figure 3: Variability in average loudness (x-axis) and in duration (y-axis) for individual speakers of IndE (crosses) and BrE (circles).

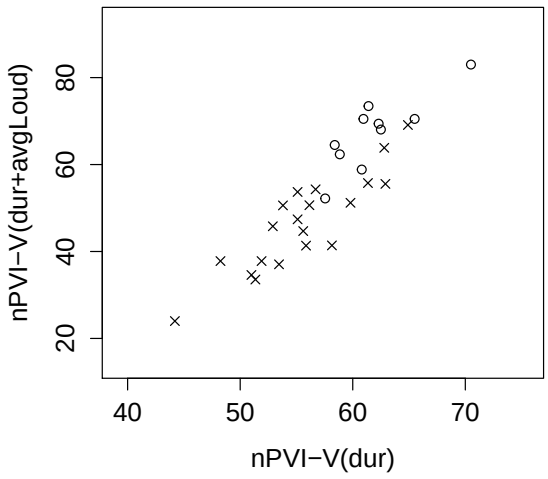


Figure 4: Variability in average loudness (x-axis) and simultaneous variability in duration and average loudness (y-axis) for individual speakers of IndE (crosses) and BrE (circles).

be louder than average, and vice versa. By contrast, in IndE, vocalic intervals that are longer than average are not as frequently and to the same extent also louder than average, and vice versa, as in BrE.

This comparison of variability in duration, loudness and simultaneous variability in loudness and duration also contributes to a better understanding of how speech rhythm is realised in different dimensions, using different acoustic and perceptual correlates of prominence. Duration is but one of these correlates of prominence. Loudness is another, and the simultaneous in- or decrease of loudness and duration is a third, and separate, aspect that needs to be considered within a multidimensional model of speech rhythm.

The results also substantiate previous descriptions of IndE as more syllable-timed than BrE [32,43,47–52]. Such a description, using a relative expression ("more syllable-timed") appears to be adequate since syllable- and stress-timing are probabaly better regarded as poles of a continuum, and perhaps as ideals that are rarely or never realised in actual speech.

Furthermore, a small number of outliers with values for variability similar to the other group occured. This shows that differences in variability in duration, loudness and simultaneous variability in loudness and duration do not constitute a categorical contrast between IndE and BrE. However, the differences are highly significant. Taking into account that a relatively long text of 392 words was used, it is very likely that the results can be generalised to other speakers of BrE and educated IndE. Considering the background of the IndE speakers, who were all engaged in university studies, it is likely that they speak a variety of IndE that is a good example of the emerging standard of IndE. This standard appears to involve a rhythm that is more syllable-timed than BrE speech rhythm.

7. References

[1] E. L. Low, E. Grabe, and F. Nolan, "Quantitative characterization of speech rhythm: Syllable-timing in singapore english," *Language and Speech*, vol. 43, no. 4, pp. 377–401, 2000.

[2] K. L. Pike, *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.

[3] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

[4] T. Platt, H. Weber, and M. L. Ho, *The New Englishes*. London/ Melbourne: Routledge, 1984.

[5] D. Crystal, "Documenting rhythmical change," in *Studies in General and English Phonetics: Essays in Honour of Professor J D O'Connor*, J. W. Lewis, Ed. London: Routledge, 1995, pp. 174–9.

[6] R. Mesthrie, "Synopsis: the phonology of english in africa and south and southeast asia," in *Varieties of English. Africa, South and Southeast Asia*, R. Mesthrie, Ed. Berlin: de Gruyter, 2008, pp. 307–319.

[7] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, pp. 265–92, 1999.

[8] V. Dellwo, "Rhythm and speech rate: A variation coefficient for deltac," in *Language and Language-Processing. Proceedings of the 38th Linguistics Colloquium*, P. l. Karnowski and I. Szigeti, Eds. Frankfurt am Main: Peter Lang, 2006, pp. 231–241.

[9] D. Gibbon and U. Gut, "Measuring speech rhythm," in *Proceedings of Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 91–94.

[10] D. Deterding, "The rhythm of singapore english," in *Proceedings of the Fifth Australian International Conference on Speech Science and Technology*, R. Togneri, Ed. Perth: Uniprint, 1994, pp. 316–321.

[11] ——, "The measurement of rhythm: a comparison of singapore and british english," *Journal of Phonetics*, vol. 29, pp. 217–230, 2001.

[12] P. Wagner and V. Dellwo, "Introducing yard (yet another rhythm determination) and re-introducing isochrony to rhythm research," in *Proceedings of Speech Prosody 2004*. ISCA, 2004, pp. 227–230.

[13] U. Gut, "Non-native speech rhythm in german," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 2003)*, M.-J. S. Daniel Recasens and J. Romero, Eds. Barcelona: Universitat Autónoma de Barcelona, 2003, pp. 2437–2440.

[14] P. M. Bertinetto and C. Bertini, "On modeling the rhythm of natural languages," in *Proceedings of Speech Prosody 2008, Campinas, Brazil*, P. A. Barbosa, S. Madureira, and C. Reis, Eds. ISCA Archive, 2008, pp. 427–30. [Online]. Available: http://www.isca-speech.org/archive/sp2008

[15] V. Dellwo, A. Fourcin, and E. Abberton, "Rhythmical classification of languages based on voice parameters," in *Proceedings of ICPhS XVI*, J. Trouvain and W. J. Barry, Eds. Dudweiler: Pirrot, 2007, pp. 1129–1132.

[16] L. White, S. L. Mattys, L. Series, and S. Gage, "Rhythm metrics predict rhythmic discrimination," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 1009–1012.

[17] L. White and S. L. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.

[18] ——, "Rhythmic typology and variation in first and second languages," *Segmental and Prosodic Issues in Romance Phonology*, vol. 282, pp. 237–257, 2007.

[19] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, pp. 351–373, 2012.

[20] E. L. Low, "Prosodic prominence in singapore english," Ph.D. dissertation, University of Cambridge, 1998.

[21] L. He, "Syllabic intensity variations as quantification of speech rhythm: Evidence from both l1 and l2," in *Speech Prosody 2012, 6th International Conference*, Shanghai, 2012. [Online]. Available: http://speechprosody2010.illinois.edu/papers/100039.pdf

[22] R. E. Cumming, "The language-specific integration of pitch and duration," Ph.D. dissertation, University of Cambridge, 2010.

[23] A. Galves, J. Garcia, D. Duarte, and C. Galves, "Sonority as a basis for rhythmic class discrimination," in *Proceedings of Speech Prosody 2002*, 2002, pp. 323–326.

[24] D. Stojanovic, "Issues in the quantitative approach to speech rhythm comparisons," *Working Papers in Linguistics (University of Hawai'i at Mānoa)*, vol. 40, no. 9, pp. 1–20, 2009.

[25] A. Loukina, G. Kochanski, B. Rosner, E. Keane, and C. Shih, "Rhythm measures and dimensions of durational variation in speech," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3258–3270, 2011.

[26] F. Nolan and E. L. Asu, "The pairwise variability index and coexisting rhythms in language," *Phonetica*, vol. 66, pp. 64–77, 2009.

[27] S. S. Stevens, "The relation of pitch to intensity," *Journal of the Acoustical Society of America*, vol. 6, p. 150, 1935.

[28] M. E. Beckman, *Stress and Non-Stress Accent*. Dordrecht/Riverton: Foris, 1986.

[29] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer (Computer Program). Version 5.3.04*, Std. [Online]. Available: www.praat.org

[30] S. B. Desai, A. Dubey, B. L. Joshi, M. Sen, A. Shariff, and R. Vanneman, *Humand Development in India*. Oxford: Oxford University Press, 2010.

[31] C. Chandramouli, *Census of India 2011. Provisional Population Totals*. New Delhi: Government of India, 2011.

[32] C. P. Masica, *The Sound System of Indian English*. Hyderabad: Central Institute of English and Foreign Languages, 1972.

[33] J. Mukherjee, "Steady states in the evolution of new englishes: Present-day indian english as an equilibrium," *Journal of English Linguistics*, vol. 35, no. 2, pp. 157–187, 2007.

[34] P. Trudgill, "Standard english: What it isn't," in *Standard English: The Widening Debate*. Routledge, 1999, pp. 177–128.

[35] A. Deumert and W. Vandenbusche, "Research directions in the study of language standardization," in *Germanic Standardizations: Past to Present*, A. Deumert and W. Vandenbusche, Eds. Amsterdam: John Benjamins, 2003, pp. 455–467.

[36] J. Edwards, *Language and Identity*. Cambridge: Cambridge University Press, 2009.

[37] O. Maxwell and J. Fletcher, "The acoustic characteristics of diphthongs in indian english," *World Englishes*, vol. 29, no. 1, pp. 27–44, 2010.

[38] ——, "Acoustic and durational properties of indian english vowels," *World Englishes*, vol. 28, no. 1, pp. 52–69, 2009.

[39] O. Maxwell, "Marking of focus in indian english of l1 bengali speakers," in *Proceedings of Speech Sciene and Technology 2010*. Australasian Speech Science and Technology Association, 2010. [Online]. Available: http://assta.org/sst/SST-10/SST2010/PDF/AUTHOR/ST100046.PDF

[40] L. Pickering and C. Wiltshire, "Pitch accent in indian-english teaching discourse," *World Englishes*, vol. 19, no. 2, pp. 173–183, 2000.

[41] C. Wiltshire and R. Moon, "Phonetic stress in indian english vs. american english," *World Englishes*, vol. 22, no. 3, pp. 291–303, 2003.

[42] C. R. Wiltshire and J. D. Harnsberger, "The influence of gujarati and tamil l1s on indian english: a preliminary study," *World Englishes*, vol. 25, no. 1, pp. 91–104, 2006.

[43] P. Sailaja, "Indian english: Features and sociolinguistic aspects," *Language and Linguistics Compass*, vol. 6, no. 6, pp. 359–370, 2012.

[44] M. B. Emeneau, "India as a linguistic area," *Language*, vol. 32, no. 1, pp. 3–16, 1956.

[45] ——, *Language and Linguistic Area*, A. S. Dil, Ed. Stanford: Standford University Press, 1980.

[46] C. P. Masica, *Defining a Linguistic Area*. Chicago, etc.: University of Chicago Press, 1976.

[47] P. Trudgill and J. Hannah, *International English*, 4th ed. London: Arnold, 2002.

[48] R. Gargesh, "Indian english: Phonology," in *A Handbook of Varieties of English*, E. W. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, Eds. Berlin: Mouton de Gruyter, 2004, vol. 1, pp. 992–1002.

[49] R. Hickey, "South asian englishes," in *Legacies of Colonial English: Studies in Transported Dialects*. Cambridge: Cambridge University Press, 2004, pp. 536–558.

[50] C. Lange, "Review of pingali sailaja *Indian English*," *Annual Review of South Asian Languages and Linguistics*, vol. 3, pp. 213–216, 2009.

[51] P. Sailaja, *Indian English*. Edinburgh: Edinburgh University Press, 2009.

[52] ——, "The standard, (non-)rhoticity and rhythm: A response to lange," *Annual Review of South Asian Languages and Linguistics*, vol. 4, pp. 183–186, 2010.

[53] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "A forensic phonetic study of dynamic sources of variability in speech: The dyvis project," in *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, P. Warren and C. Watson, Eds., 2006, pp. 13–18.

[54] P. Machač and R. Skarnitzl, *Principles of Phonetic Segmentation*. Prague: Epocha, 2009.

A Durational Study of German Speech Rhythm by Chinese Learners

Hongwei Ding[1,2] *and Rüdiger Hoffmann*[3]

[1]School of Foreign Languages, Shanghai Jiao Tong University, China
[2]School of Foreign Languages, Tongji University, China
[3]Institute of Acoustics and Speech Communication, TU Dresden, Germany
hongwei.ding@tongji.edu.cn, ruediger.hoffmann@tu-dresden.de

Abstract

This study focuses on the temporal and metrical features of the German speech produced by Chinese speakers. German is described to be a stress-timed language, while standard Chinese is regarded as a syllable-timed language. It has been suggested that the rhythm of the target language can be influenced by the learners' native language. In this study we conducted an investigation of ten sentences with 18 Chinese students in the low intermediate proficiency level in comparison with six native German speakers. We compared the duration values in terms of pairwise variability indices, and found that most of these Chinese speakers have a lower *nPVI-V* and a higher *rPVI-C* than the German speakers. We illustrate that the conventional duration measures of *nPVI-V* can be influenced by the syllable structures of the utterance and the classification approach of vocalic intervals, and a comparable *nPVI-V* can hardly be expected from different investigations. Furthermore, we argue that duration values alone cannot fully capture the rhythmic patterns of speech because other prosodic parameters such as pitch and energy also join to contribute to rhythmic characteristics of the speech.

Index Terms: durational study, learning German, Chinese learners

1. Introduction

It is well known that Pike [16] and Abercrombie [1] argue that the languages of the world can be classified into two types of rhythm patterns: *a) stress-timed rhythm* and *b) syllable-timed rhythm*. According to this hypothesis, both types of rhythm show rhythmical units of equal duration: *stress-timed* languages tend to have isochronous inter-stress intervals, while *syllable-timed* languages tend to have rather equal syllable durations. Classic examples for a stress-timed language are English and German; while Chinese belongs to syllable-timed languages [13]. The classification of rhythm-classes turned out to be based solely on intuition since a large amount of experiments carried out to provide direct correlates for the isochrony in languages remained without success [17].

In the recent decades many researchers tried to classify languages in other ways. Ramus et al. [17] proposed to divide speech into vocalic and consonantal parts, and to calculate the proportion of the vocalic intervals (*%V*) and the standard deviation of consonantal intervals (ΔC) in a sentence. They showed that stress-timed languages have a higher ΔC and a lower *%V*, whereas syllable-timed languages have a lower ΔC and a higher *%V*. Grabe and Low [9] based their pairwise comparison of successive vocalic and intervocalic intervals and calculated speech rhythm with the *Pairwise Variability Index (PVI)*, which computes the sum of the durational differences between adjacent vocalic or consonantal intervals in an utterance. They found that stress-timed languages have a higher variation in vowel durations, whereas syllable-timed languages (including Mandarin) show a lower variation in vowel length. Lin et al. [13] followed the studies of Ramus et al. [17] and Grabe and Low [9], and measured *%V*, ΔC, normalised variation of the pairs of two adjacent vowel intervals (*nPVI-V*), and raw variation of the pairs of two adjacent consonant intervals (*rPVI-C*) of Mandarin Chinese. Except for *nPVI-V*, all other measures confirmed the auditory impression of Mandarin Chinese being syllable-timed [13].

On the other hand, it has been suggested that the rhythm of the target language can be influenced by the learners' native language [10, 18]. Gut [10] described that L2 German is influenced by L1 of Chinese, English, French, Italian and Romanian in terms of ΔC, *%V* etc. In our previous study [6], we demonstrated that Chinese learners of German in the low intermediate proficiency level have a clearly higher *%V* and roughly higher ΔC than German native speakers, because they insert many vowel epentheses and speak at a slower rate. In this investigation we aim to investigate non-native rhythm of Chinese speakers in terms of *nPVI-V* and *rPVI-C* values, and to discuss the efficiencies and deficiencies of these measures with regard to our investigation material.

2. Method

This study followed the same method which was described by Grabe and Low [9] to investigate metrical features of German speech produced by Chinese speakers. The syllable-timed rhythm in the German speech of the Chinese subjects was so striking for their German teachers, it was thus interesting to find out whether these rhythm measures can reflect the perceptual impression of the rhythmic deviation for native German listeners. In order to ensure comparability, the annotation technique used by Grabe and Low [9] was adopted in the investigation. The durations of vowels and intervals between vowels (excluding pauses) in each sentence were measured. Indices of *rPVI* and *nPVI* were calculated according to equation (1) and (2) respectively. (Note: *rPVI-C* and *nPVI-V* are adopted in this paper to represent consonantal *rPVI* and vocalic *nPVI*.)

$$rPVI = \sum_{k=1}^{m-1} |d_k - d_{k+1}| \cdot \frac{1}{m-1} \quad (1)$$

$$nPVI = 100 \cdot \left(\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| \right) \cdot \frac{1}{m-1} \quad (2)$$

where *m* is the number of intervocalic or vocalic intervals in a sentence, and *d* is the duration of the *k*th interval.

With the results of this investigation we endeavour to answer the following questions:

- Do Chinese speakers exhibit different *nPVI-V* and *rPVI-C* from German native speakers? If yes, what factors might account for the differences?
- Can the pairwise variability indices produce comparable values from different investigations? Can these indices fully reflect the perceptual impression of rhythm?

2.1. Subjects

We recruited 18 native Chinese speakers, including 10 males and 8 females, who come from different parts of China, but all of them speak standard Chinese. At the time of speech data collection they had been living in Germany for one month, and they were just enrolled in the German language course for the DSH exam (the German language university entrance exam for foreign students). Their ages ranged from 22 to 28. All of them had learned German for one to one and a half years, and the length of their formal German instructions had been around 1,200 hours. These participants could be classified as having a low intermediate level, they formed a homogeneous group in terms of age, L1 background, motivation, proficiency of the German language, and the length of residence in Germany. Their German teachers found that their German speech was highly syllable-based, which deviated from the stress-timed native German speech, we thus conducted a durational investigation into their L2 speech to test whether *PVI* measures can reflect the deviated rhythm perceived. In order to provide a reference for comparison, we included six German speakers, one was male and five were female speakers. They were between 22-30 years old and were ordinary German native speakers.

2.2. Speech data collection

In order to have certain control of the speech data, reading tasks were used for analysis. The subjects were instructed to read 50 Phondat sentences in German. For the current study only ten read sentences were selected and analyzed, because they include different sentence types and the vowel and consonant percentages vary from sentence to sentence. It is better to concentrate on a small amount of data since the accuracy of annotation is essential for the measurement, which requires much carefulness and patience. Before the recording began, the subjects were given as much time as they needed to read the text to become familiarized with it. Each subject was individually recorded at 16-bit resolution with a sampling rate of 44.1 kHz by a German phonetics expert, who controlled the quality of their production.

2.3. Data analysis

The sentences were first automatically labeled by a trained aligner, and then manually corrected by the first author assisted by a German phonetics expert on Praat [3]. Sentences were first annotated at the phoneme level on the basis of both audio and visual information, and the phonemes were then grouped into vocalic and consonantal intervals. We followed the same criteria described by Grabe and Low [9] in determining the location of vowel-consonant and consonant-vowel boundaries:

- *Vocalic intervals* were characterized by vowel formants, which could contain a monophthong, a diphthong, or more vowels if the formants continue.
- *Intervocalic intervals* were defined as stretch of signals between vocalic intervals, which could include one or more consonants.

Annotation of vowel boundaries were conducted according to the generally accepted criteria [15]. Consonants such as stops, fricatives, affricates and nasals were clearly identifiable from the changes in spectrogram and formant structures. The approach to glides was also based on acoustic criteria, like that in [9]. If a clear change could be observed in the formant structure or in the amplitude as a prevocalic glide, it was excluded from vocalic portions. Otherwise, glides (especially postvocalic glides) were included in the vocalic portion. If there was a gap that was not part of the sound, it was marked as breath, and this breath gap was subtracted from the calculation of the intervals. Any two intervocalic (or vocalic) intervals split by this gap were combined into the same intervocalic (or vocalic) interval. Since we employed sentences as reading material, we computed the pairwise variability in sentences other than in a passage of speech in [9]. This procedure allows us to compare whether there are any differences among different kinds of sentences.

3. Results

The comparison results on *nPVI-V* and *rPVI-C* measures, and speaking rate are presented in the following.

3.1. Pairwise variability indices

Figure 1 demonstrates the data of the six German native speakers (de) shown as small triangles and 18 Chinese speakers (cn) indicated with small squares. *nPVI-V* values are plotted on the horizontal axis against *rPVI-C* values on the vertical axis. The index values of each speaker are the average of ten sentences.

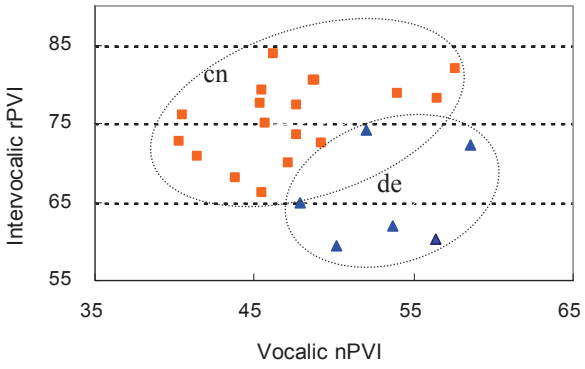


Figure 1: *Vocalic nPVI-V against intervocalic rPVI-C.*

A two-sample independent t-test shows the differences between the Chinese and the German speakers in both *rPVI-C* (t=-3.93, p=0.001) and *nPVI-V* (t=2.515, p=0.020) are significant ($p<0.05$). The Chinese speakers have a higher *rPVI-C* and a lower *nPVI-V* than the German native speakers.

3.2. Speaking rate

It is clear that the Chinese speakers spoke at a slower rate and made more pauses than the native German speakers. Without consideration of pauses or silences, the average speaking rate calculated in phonological syllables per second (syl/s) of the native German speakers and Chinese speakers are 5.95 and 3.90

respectively. A significant negative correlation (r=-0.85) between speaking rate and *rPVI-C* is found for the German speakers, but no such correlation (r=-0.17) can be found for the Chinese speakers.

Table 1: Speaking rate and correlation with *rPVI-C*

	German speakers	Chinese speakers
speaking rate (syl/s)	5.95	3.90
standard deviation	0.42	0.26
correlation with *rPVI-C*	-0.85	-0.17

Since the Chinese speakers inserted many epenthesis vowels, they produced much more additional syllables [7]. Speaking rate calculated on the basis of phonetically uttered syllables resulted in a higher rate of 5.27 syl/s (5.27>3.90), which is also negatively correlated with *rPVI-C* (r=-0.567, p=0.014).

4. Discussion

The *PVI* results obtained in this investigation can partly reflect the rhythmic deviance of L2 speech, but may not fully reflect the perceptual impression of speech rhythm.

4.1. Evaluation of *PVI*

The outcome where the Chinese speakers have a higher *rPVI-C* mirrors the result of the previous investigation [6, 10], where the Chinese speakers produced a higher standard deviation of consonantal intervals (ΔC). Since speaking rate is negatively correlated with ΔC in the German language [5], a slower speaking rate results in a higher ΔC, and also produces a higher *rPVI-C*. Moreover, the Chinese speakers can hardly reduce nonstressed vowels, and their *nPVI-V* should be much lower than the German native speakers, which is also indicated in the results. These findings support the results in previous studies [9, 8, 2]. Generally speaking, a higher *rPVI-C* and a lower *nPVI-V* can partly reflect the syllable-timed characteristic of L2 German speech by Chinese speakers.

4.2. Influencing factors

However, there are many factors which can influence the measures of *nPVI-V*, examples in this investigation will be illustrated in the following.

4.2.1. Syllable combinations

The *nPVI-V* scores are found to be influenced not only by the structure of the syllables but also by their combinations in our data. This observation supports previous findings [2, 14] that metric scores can be affected by the choice of material. If one pair of syllables exhibits quite different structures or displays quite different stress patterns, their *nPVI-V* values can be higher than other pairs. More various pairs in one sentence can result in a higher *nPVI-V*. Therefore, the mean *nPVI-V* scores in our investigation are quite different from sentence to sentence, which range from 41.98 to 74.18 for the German speakers, and from 37.06 to 55.50 for the Chinese speakers. However, *nPVI-V* values based on sentences are highly correlated between the German speakers and the Chinese speakers with r=0.67 ($p<0.05$), which explains that *nPVI-V* is partly subject to the syllable combinations in the sentence. For example, in *Wie hast du das gemacht? (How did you do that?)*, schwa in prefix /g@/ between /das/ and /maxt/ makes *nPVI-V* larger for both German and Chinese speakers.

4.2.2. Successive vowels

Successive vowels do not influence *%V*, but can influence *nPVI-V*. The conventional approach in both Ramus [17] and Grabe & Low [9] is to include adjacent heterosyllabic vowels in the same vocalic interval, which makes *nPVI-V* quite different for different sentences. Successive vowels may and may not be separated by a glottalized period.

If glottalization is found between successive vowels, these intervals were treated as silent pauses. We followed the approach in [19], the glottalized part marked between the solid lines in Figure 2 was omitted in the calculation, and the vocalic intervals before and after the glottalization were summed.

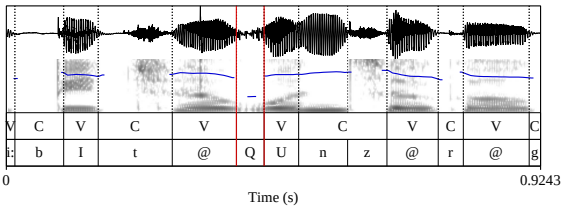


Figure 2: *Waveform, spectrogram and SAMPA annotation of "bitte unsere (please our)" with glottalization between successive vowels by a Chinese speaker.*

Another speaker produced the same phrase without glottalization, as it is shown in Figure 3.

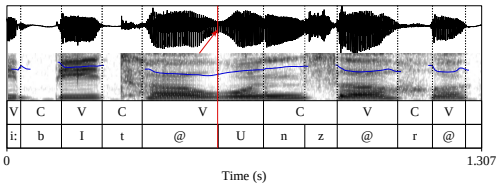


Figure 3: *Waveform, spectrogram and SAMPA annotation of "bitte unsere (please our)" without glottalization between successive vowels by a Chinese speaker.*

To put /@/ and /U/ in one vocalic interval enlarges the *nPVI-V* value greatly for the Chinese speakers. However, both a decrease in pitch and amplitude in the glottalization in Figure 2, and a decrease in the amplitude as pointed out by the arrow in Figure 3 without glottalization can be clearly perceived as a reset in prosody. It is obvious that in the above two figures the duration values of the five vowels are comparable. However, to group these two successive vowels into one vowel interval enlarges *nPVI-V*, which cannot reflect the perceptual impression of the syllable-based rhythm of Chinese speakers faithfully.

4.2.3. Syllabic consonants

Another factor which influences *nPVI-V* values is the annotation of fully reduced vowels in German. One obvious instance is that many word-final syllables in infinite verbs with *-en* are usually pronounced with syllabic consonants. For example, in

Figure 4, schwa @ in word-final with *-en* is reduced, and consonant *s* and syllabic consonant *=n* are combined without any vocalic interval between them.

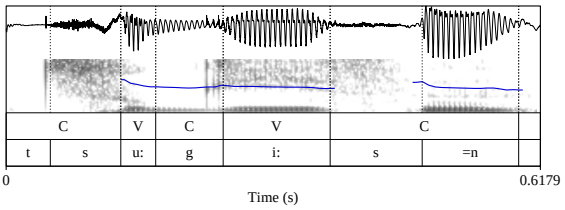


Figure 4: *Waveform, spectrogram and annotation of "zu gießen (to water)" with a syllabic consonant (=n) by a German speaker.*

If a short schwa @ is annotated, the *nPVI-V* can be greater than that with a syllabic consonant. However, a very short weak @ might not be quite different from that with a syllabic consonant perceptually, but the *nPVI-V* values may be quite different.

4.3. Measurements beyond duration

We have observed that various syllable combinations and different annotation approaches of vocalic intervals may lead to various *nPVI-V* scores, which implies that the measures of vocalic and intervocalic intervals may not fully reflect the perception of rhythm. This finding further supports the idea that rhythm metrics such as *PVI* can provide measures of speech timing and variability, but they cannot convey an overall rhythmic impression [2]. It is generally believed that rhythm is the recurring timing patterns of fundamental frequency, syllabic duration, syllabic energy, and spectral dynamics [12]. By comparing the syllable-timed L2 German speech with the stress-timed native German speech, it suggests that not only timing but also fundamental frequency and energy contribute to the rhythmic pattern over time [12]. Cumming [4] also demonstrated that f0 and duration are interdependent in the perception of rhythmic groups in speech and sentence rhythmicity.

German has full as well as spectrally reduced and shortened vowels, and the consequence is a high level of variability in vowel durations. Chinese does not have vowel reduction, and the level of vocalic variability is significantly lower. German employs sentence intonation to express intonational meanings; Chinese uses lexical tones to distinguish words. The difference between a stress-timed German speech by a native speaker and a syllable-timed German speech by a Chinese speaker can be observed in the following two figures, which are vividly displayed with ProZed [11].

In these figures, the pitch contour of the sentence is clearly demonstrated in a continuous dotted line; each circle corresponds to one vocalic or consonantal interval with the vocalic interval marked with corresponding vowels. The level on the vertical y-axis of the circle represents the pitch and the diameter represents the duration of the interval. The unit of pitch has already been normalized to the logarithmic scale $\log_2$(Hz/median), so that speakers of different fundamental frequencies can be comparably displayed on this scale. It is obvious that the vowels have more variability for the German speaker than the Chinese speaker. In Figure 5 the vowels /*u:*/ in **du** and /@/ in ***ge****macht* are very short in duration, while /*a*/ in *ge****macht*** is much longer of the German speaker. In Figure 6, except for a very short /@/ in ***ge****macht*, the variations of the other vowels (including the epenthesis /@/ after /t/ in *hast****e***) are not so much as those in Figure 5. Not only duration variations but also pitch and energy changes are different. All these prosodic parameters work together to organize speech into rhythmic chunks, which impresses us with different rhythms.

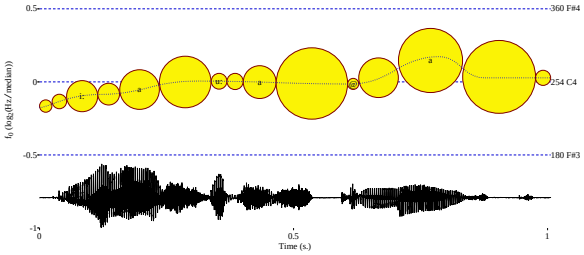


Figure 5: *Prosody display of "Wie hast du das gemacht? (How did you do that?)" produced by a German speaker.*

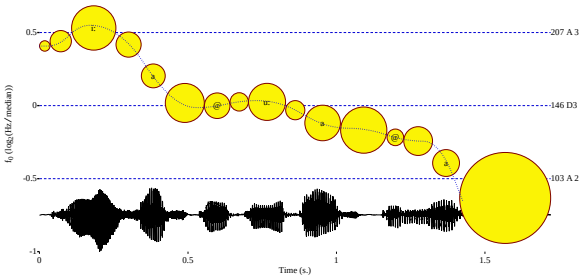


Figure 6: *Prosody display of "Wie hast du das gemacht? (How did you do that?)" produced by a Chinese speaker*

5. Conclusion

This paper analyzes L2 German speech with pairwise variability indices and demonstrates that the syllable-timed L2 German speech by Chinese speakers is characterized by a lower *nPVI-V* and a higher *rPVI-C*. However, due to many influencing factors, a comparable *nPVI-V* can hardly be expected from different investigations. We further suggest that other prosodic parameters, such as F0 and energy, can work together with duration to contribute to the acoustic analysis of rhythmicity, which can better reflect the perceptual impressions of speech rhythm. This study also suggests that in the future we should focus more on grouping patterns of prominence with more prosodic parameters than duration to investigate speech rhythm, and it would be more reasonable to link measurements of acoustic parameters in rhythm production with listeners' rhythm perception in the investigation of rhythm, as it is suggested by many researchers in [4, 12].

6. Acknowledgements

The first author is sponsored by Shanghai Social Science project (2011BYY002) and Innovation Program of Shanghai Municipal Education Commission (12ZS030) for this research work. We thank Rainer Jäckel for his help in the data collection.

7. References

[1] Abercrombie, D., “Elements of general phonetics”, Aldine: Chicago, 1967.

[2] Arvaniti, A., “Rhythm, Timing and the Timing of Rhythm”, Phonetica, 66:46-63, 2009.

[3] Boersma, P. and Weenink, D. “Praat: doing phonetics by computer [Computer program]”, http://www.praat.org/, 2013.

[4] Cumming, R. E., “Perceptually Informed Quantification of Speech Rhythm in Pairwise Variability Indices”, Phonetica, 68:256-277, 2011.

[5] Dellwo, V. and Wagner, P., “Relations between language rhythm and speech rate”, Proc. of the 15th ICPhS, 471–474, 2003.

[6] Ding, H., Jäckel, R., and Hoffmann, R., “A Preliminary Investigation of German Rhythm by Chinese Learners”, ESSV2013, 79-85, 2013.

[7] Ding, H. and Hoffmann, R., “A An Investigation of Vowel Epenthesis in Chinese Learners’ Production of German Consonants”, Interspeech, 1007–10115, 2013.

[8] Gibbon, D. and Gut, U., “Measuring speech rhythm”, Eurospeech 2001.

[9] Grabe, E. and Low, E. L., “Durational variability in speech and the rhythm class hypothesis”, in C. Gussenhoven, and N. Warner [Ed], Laboratory Phonology, 7:515–546, Berlin: Mouton, 2002.

[10] Gut, U., “Non-native Speech: A Corpus-based Analysis of Phonological and Phonetic Properties”, Peter Lang GmbH, 2009.

[11] Hirst, D., “ProZed: A speech prosody analysis-by-synthesis tool for linguists”, Proc. of Speech Prosody 2012, 15-18, 2012.

[12] Kohler, Klaus J., “Rhythm in Speech and Language: A New Research Paradigm”, Phonetica, 66:29-45, 2009.

[13] Lin, H., and Wang, Q., “Mandarin Rhythm: An Acoustic Study”, Journal of Chinese Language and Computing, 17:127-140, 2007.

[14] Loukina, A., and Kochanski, G., and Rosner, B. and Keane, E., “Rhythm measures and dimensions of durational variation in speech”, Journal of the Acoustical Society of America, 129:3258-3270, 2011.

[15] Peterson, G. E. and Lehiste, I., “Duration of Syllable Nuclei in English”, Journal of the Acoustical Society of America, 32:693-703, 1960.

[16] Pike, K. L., “The intonation of American English”, University Press: Michigan, 1945.

[17] Ramus, F., Nespor, M., and Mehler, J., “Correlates of linguistic rhythm in the speech signal”, Cognition, 72:1-28, 1999.

[18] Tortel, A., and Hirst, D. “Rhythm metrics and the production of English L1/L2”, Speech Prosody, 2010.

[19] White, L., and Mattys, S. L., “Calibrating rhythm: First language and second language studies”, Journal of Phonetics, 35:501-522, 2007.

Metrical Structure and Jaw Displacement: An Exploration

Donna Erickson [1], *Shigeto Kawahara* [2], *J.C. Williams* [3], *Jeff Moore* [4], *Atsuo Suemitsu* [5], *Yoshiho Shibuya* [1]

[1] Kanazawa Medical University, Japan; [2] Keio Institute of Language & Cultural Studies, Tokyo; [3] Independent, Kamakura, Japan; [4] Sophia University, Tokyo; [5] Japan Advanced Institute of Science & Technology, Ishikawa Pref., Japan

EricksonDonna2000@gmail.com, kawahara@icl.keio.ac.jp, catitawms@gmail.com, jeffmoore.personal@gmail.com, sue@jaist.ac.jp, yshibuya@kanazawa-med.ac.jp

Abstract

Building on Erickson et al. [9], the current electromagnetic-articulography (EMA) experiment proposes that the amount of jaw displacement—or mandible movement—may reflect the metrical organization of English sentences. The experiment also supports F1 as a reliable acoustic correlate of jaw displacement, hence metrical organization. On the other hand, the study also demonstrates that F0 does not have a similar relationship to mandible movement.

Index Terms: metrical organization, jaw movement, EMA, F1, F0

1. Introduction

Work in Metrical Phonology [1] (*et seq.*) posits that our speech exhibits patterns of organization that are akin to the metrical organization of poetry [2], [3]. Metrical structure consists of rhythmic categories, within which strong-weak patterns are assigned. Different theories propose different levels of rhythmic categories, but they generally include such units as syllable, (prosodic) word, minor phrase, major phrase (and utterance), as in Figure 2 [4], [5], [6:384]. Each syllable, based on its metrical constituency, is assigned a level of stress or prominence relative to other syllables in the utterance [4], [5], and overall prominence may be represented as a number of grid marks [7]. The acoustic prominence of each unit is often described in terms such as quality, duration, loudness, and pitch [8].

Recent X-ray Microbeam and electromagnetic-articulography (EMA) studies, inspired by Fujimura [10]-[12], with four American English speakers [9], suggest that an articulatory correlate of syllable stress in English is the amount of jaw (mandible) displacement. As a speaker opens and closes the mouth to produce each syllable, the jaw moves. It opens more for low vowels than high vowels *ceteris paribus* [13], [14]; if all the vowels in an utterance are the same, the pattern of jaw opening should mirror syllable stress levels and correspond to the metrical structure of the speaker's utterance. The greater the jaw displacement, the higher the level of syllable stress. Also, the first resonant formant of the vowel (F1) exhibits a significant correlation with jaw displacement and metrical structure, suggesting that F1 is an important acoustic correlate of English metrical structure. Figures 1 and 2, based on [9], illustrate the pattern of jaw displacement for three American English speakers uttering: "[Yes, I saw] five bright highlights [in the] sky [to]night". The greater the jaw displacement, the greater the syllable stress (syllable "magnitude" according to [10]), and this magnitude is displayed vertically in bars, making it easier to visualize how jaw displacement mirrors the metrical structure of the utterance, shown in Figure 2. F1 values are not reproduced here, but they present essentially the same pattern as jaw displacement [9]. For these three speakers, nuclear stress—the most prominent stress in a sentence [15]—is on "high" of "highlights". For the fourth speaker (not shown here), nuclear stress is on "five"; this difference of nuclear stress placement is also observed in our current experiment, as discussed in the Results section.

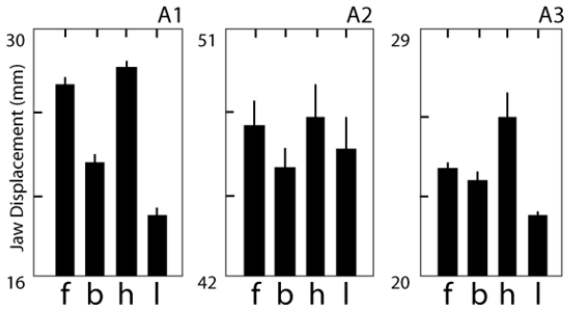


Figure 1: *Bar graphs showing the amount of jaw displacement (mm) for 3 American English speakers, for the sentence "[Yes, I saw] 5 bright highlights [in the sky tonight]." Error bars indicate standard error of the mean. Ordinate scaling is by speaker, to better display individual jaw displacement patterns.*

Major Phrase			x	
Minor Phrase	x		x	
Word	x	x	x	
Syllable	x	x	x	x
Stress level	3	2	4	1
	five	bright	high	lights

Figure 2: *Metrical structure of the utterance in Figure 1.*

One question addressed in this paper is how robust is the correlation between metrical prominence and jaw opening and between F1 and maximum jaw displacement. Additionally, do these patterns generalize beyond the single sentence tested in [9]? Is metrical prominence the only factor that affects jaw movement patterns? Does this correlation generalize to other vowels or do specific consonantal or vocalic gestures affect jaw displacement patterns? In addition, we examined F0 in

order to see what role it plays in implementing metrical stress patterns in American English.

2. Method

To address these questions, we report on jaw displacement patterns and corresponding acoustic measurements of data recorded by 3D-EMA (Carstens AG500 Electromagnetic Articulograph) at the Japan Advanced Institute for Science and Technology (JAIST, Ishikawa Prefecture, Japan) for three English sentences, as produced by two speakers of American English, one male (Speaker 1, fourth author), and one female, (Speaker 2, third author, an English-Spanish bilingual). Custom software (mview, Haskins Laboratories) was used to analyze the articulatory data. The lowest vertical position (maximum displacement) of the jaw with respect to the bite plane was located for each syllable of each utterance using a velocity-based criterion. For a more detailed description of EMA, including placement of sensors and recording procedures, see [9]. The sentences were all presumed to have the same metrical pattern; in the first and second sentences, the vowel of interest is /a/ and in the third, /e/, all followed by the palatal glide. The sentences are shown below, with the target words in the midsentence underlined.

(1) [Yes, I saw] five bright highlights [in the sky tonight].

(2) [Yes, I saw] nine nice bike fights [on the dyke tonight].

(3) [Yes, I saw] eight great playmates [in the bay today].

Sentence (1) is the same as reported in [9]; Sentence (2) contains the same diphthong /aɹ/ but different words, and Sentence (3) contains the diphthong /eɹ/. The participants practiced the sentences with pictures illustrating a scenario to match each sentence. For data recording, each sentence was presented (without parentheses or underlining) six times in random order using a PowerPoint display. The underlined syllables in each sentence were analyzed for jaw displacement, F1, and peak F0. Acoustic measurements were made using Praat [16]. Maximum F0 was measured in Hz for the steady state portion of the vowel, before the transition to the glide, and F1 was measured at that same point in time. F1 measurements in [9] were taken at the time of maximum jaw displacement. However, sometimes maximum jaw displacement occurred before the onset of voicing, so that F1 was not measurable. Measuring F1 at the time of maximum F0 has its own drawbacks, in that physiologically, lowering jaw would result in low F0 since the larynx gets lowered as well. However, this measuring method is at least consistent and obvious across all the tokens recorded.

3. Results and Discussion

Figure 3 shows a sample jaw tracing of Sentence (1) "[Yes, I saw] five bright highlights [in the sky tonight]." for the male speaker. Notice in the bottom panel of Figure 3 that the jaw opens and closes for each of these four syllables, and even though the vowel plus glide is always /aɹ/, the amount of jaw opening varies systematically.

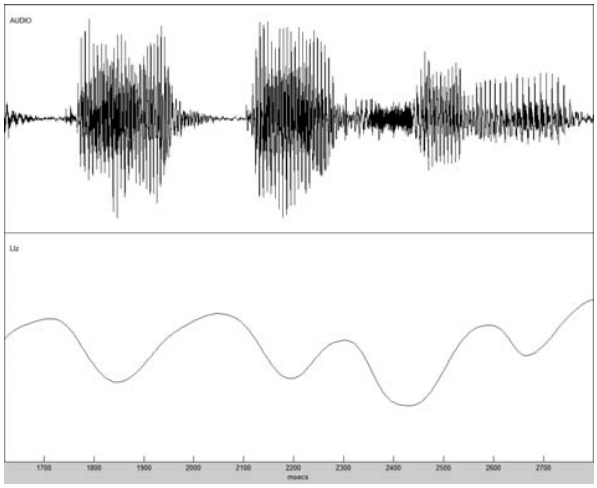


Figure 3: *Acoustic waveform (top) and jaw displacement tracing in mm (bottom) for Speaker 1 saying: "[Yes, I saw] 5 bright highlights [in the sky tonight.]"*

The patterns of jaw displacement averaged over 6 repetitions for Speakers 1 and 2 are displayed in Figure 4 such that the taller the bar, the larger the jaw displacement/syllable stress. Speaker 2 (at right in Figure 4) shows patterns similar to those reported for three of the four speakers in [9] (see Figure 1), supporting our hypothesis that metrical structure is reflected in jaw displacement patterns.

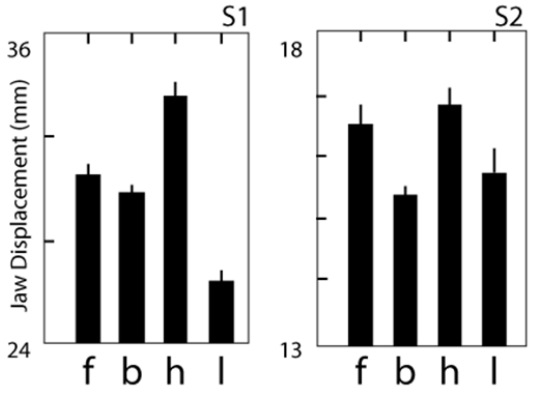


Figure 4: *Speaker 1 (L) & Speaker 2 (R) "[Yes, I saw] 5 bright highlights [in the sky tonight]" (Sentence 1).*

Figure 5 shows jaw displacement measures for Sentence (2) "[Yes, I saw] nine nice bike fights [on the dyke tonight]." Both speakers produce patterns similar to those for sentence (1), except that "nine" (the first syllable of this phrase) receives the nuclear stress. The pattern of jaw displacement for these speakers, especially for Speaker 1, is similar to that reported for the fourth speaker in [9] for "[Yes, I saw] five bright highlights [in the sky tonight]", putting nuclear stress on "five".

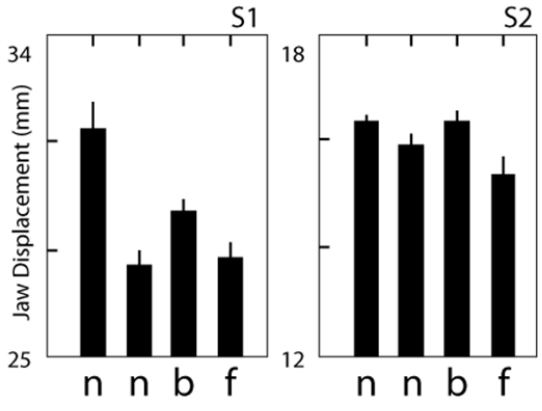


Figure 5: *Speaker 1 (L) & Speaker 2 (R) "[Yes, I saw] 9 nice bike fights [on the dyke tonight.]" (Sentence 2).*

Major Phrase	x			
Minor Phrase	x		x	
Word	x	x	x	x
Syllable	x	x	x	x
Stress level	4	2	3	2
(Yes, I saw)	nine	nice	bike	fights

Figure 6: Metrical structure of Sentence 2.

It is interesting that each speaker seems to show the jaw displacement pattern previously reported for Sentence 1, even though the words in the sentence have changed; i.e., the way they implement the metrical patterns of utterances is consistent across different sentences.

Figure 7 shows jaw displacement for a sentence with seemingly the same stress patterns, but the diphthong is /eɹ/: "[Yes, I saw] eight great playmates [in the bay today.]"

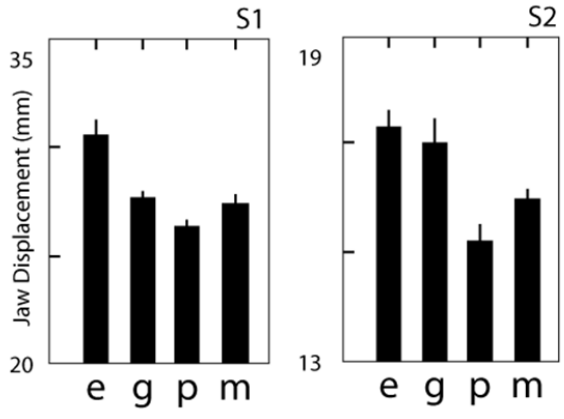


Figure 7: *Speaker 1 (L) & Speaker 2 (R). "[Yes, I saw] 8 great playmates [in the bay today]." (Sentence 3).*

Speakers 1 and 2 both show stronger jaw displacement for "eight" compared with "great", but "mates" is stronger than "play". Thus, the first phrase, "eight great", is similar to the first phrase of "nine nice bike fights" but the second phrase, "play mates", is different. It appears that the speakers treated "play mates" as a two-word phrase rather than a compound; i.e., as in the classic "bla'ckboa,rd" vs. "bla,ck boa'rd" minimal pair, where the second element in a phrasal compound receives more prominence. Thus, this sentence may have a different metrical structure than the previous two sentences.

Major Phrase	x			
Minor Phrase	x			x
Word	x	x	x	x
Syllable	x	x	x	x
Stress level	4	2	2	3
(Yes, I saw)	eight	great	play	mates

Figure 8: *Metrical structure for Sentence 3.*

What are the acoustic consequences of a speaker using the jaw to articulate the metrical structure of an utterance? Figure 9 below is a scatter plot of jaw displacement and F1 for each target word in the three sentences, produced by Speaker 1, grouped by the hypothesized stress levels.

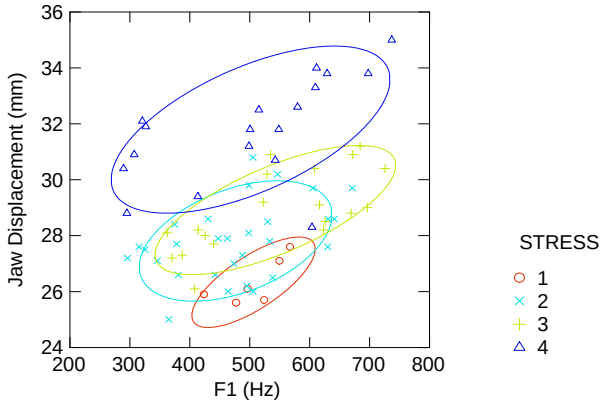


Figure 9: *Jaw Displacement and F1 (Speaker 1) as a function of stress level.*

Note that for Speaker 1, the greater the syllable stress, the larger the jaw displacement, and within each stress group, there is a significant correlation between jaw displacement and F1 for the three highest stress levels; the correlation is also high for the lowest stress group, but did not reach significance due to small N (See Table 1 for statistical results).

Table 1. *Pearson Correlation Analysis of Jaw Displacement with F1, within stress levels.*

stress level	*r*	Bartlett Chi-square statistic	df	*p*	N
1	0.737	2.743	1	0.098	6
2	0.433	5.512	1	0.019	29
3	0.694	10.823	1	0.001	19
4	0.600	6.923	1	0.009	18

Figure 10 shows a scatter plot of jaw displacement and peak F0 for each target word in the three sentences, produced by Speaker 1, grouped by the hypothesized stress levels.

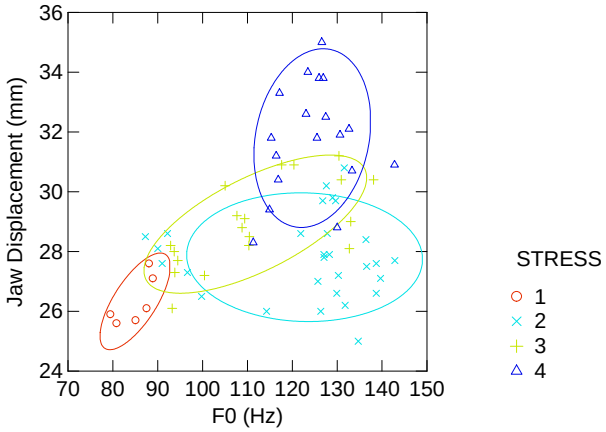


Figure 10: *Jaw Displacement and F0 (Speaker 1) as a function of stress level.*

The scatter plot does not show a consistent distribution of F0 according to the stress groups, thus suggesting that F0 is not a good indicator of stress. Stress level 2 shows a bimodal distribution, which indicates that metrical prominence cannot be the only factor that determines F0. For instance, sentence-final lowering may be responsible for the left cluster of stress level 2 [19]. In addition, F0 overlaps with a cluster of stress levels 2 and 4; and those of stress level 3 are a subset of stress level 2. From our data, it appears that F0 is not a reliable acoustic measure of metrical prominence.

For Speaker 2, we see that patterns of jaw displacement by F1 and jaw displacement by F0, when grouped by stress level, are somewhat similar to those of Speaker 1. Speaker 2's grouped stress data for jaw displacement by F1 show that the greater the stress, the larger the amount of jaw displacement (except for stress level 1); however, there is no correlation between jaw displacement and F1 within each level, as we saw for Speaker 1. Additionally, Speaker 2 does not exhibit any consistent distribution of jaw displacement by F0 according to stress groups, failing to confirm F0 as a reliable measure of metrical prominence for English sentences. The data scatter for Speaker 2 appears somewhat messy and not as clear as that seen for Speaker 1, which may reflect Speaker 2's bilingualism. The effect of bilingualism on the articulation of metrical stress is an interesting area for future exploration.

4. Conclusions

This is a preliminary report on an initial research study to investigate articulatory correlates of metrical structure in American English. Clearly, more speakers are needed, as well as more sentences. Further research, including constructing syllable triangles and boundaries according to the C/D model [10], [11], [20] will provide additional information on syllable magnitudes and differences in boundary strengths. Given that rhythm is often perceived as a pattern of "beats", the "beats" of jaw opening/closing may well be related to the rhythm/metrical structure of spoken language. Obviously, the listener does not hear jaw movement, but the resulting changes in formant frequencies, particularly in F1, may cue the listener to these "jaw beats". In our analysis of the data, peak F0 patterns fail to distinguish metrical stress levels. The results of this exploratory articulatory study reinforce the observation that metrical structure may best be reflected in jaw displacement patterns [9].

In closing, we address an issue that was raised by a reviewer -- that both of our subjects are authors. Trained speakers, including phoneticians and professional speakers, are probably able to consciously control the "rhythm" (metrical organization) of their utterances; however, it is unlikely that even these speakers can manipulate their jaw displacement patterns, since these are probably "hard-wired" from infancy [21]. Research is needed to explore this issue.

This process of deriving speech organization from mandible movement raises many questions for future investigation and provides a possible method for exploring metrical structure of language in general as well as specific languages, as demonstrated here for American English. It is hoped that this approach may offer a means to mesh the phonological abstractness of language with the phonetic instantiation of speech, resulting in new views and insights.

5. Acknowledgements

The impetus for this approach to the phonology-phonetics interface of language and its metrical structure comes from the insights of Osamu Fujimura and his life-long research in articulation, especially his theoretical framework of the C/D Model.

This work was supported by the Japan Society for the Promotion of Science, Grants-in-Aid for Scientific Research (C)#22520412 and (C)#25370444. Special acknowledgement is made to Mark Tiede for help with mview, and thanks to Jianwu Dang for the use of the EMA Lab at JAIST.

6. References

[1] Liberman, M. and Prince, A., “On stress and linguistic rhythm”, Linguistic Inquiry, 8(2):249-336, 1977.

[2] Kiparsky, P., “The rhythmic structure of English verse”, Linguistic Inquiry, 8(2):189-247, 1977.

[3] Hayes, B., “A grid-based theory of English meter”, Linguistic Inquiry, 14:357-393, 1983.

[4] Hayes, B., Metrical Stress Theory: Principles and Case Studies, University of Chicago Press, 1995.

[5] Selkirk, E., Phonology and Syntax: The Relation between Sound and Structure, MIT Press, 1984.

[6] Selkirk, E., “On derived domains in sentence phonology”, in Phonology Yearbook 3, 371-405, Cambridge U. Press, 1986.

[7] Prince, A., “Relating to the grid”, Linguistic Inquiry, 14(1):19-100, 1983.

[8] Laver, J., Principles of Phonetics, Cambridge U. Press, 1994.

[9] Erickson, D., Suemitsu, A., Shibuya, Y., and Tiede, M., “Metrical structure and production of English rhythm”, Phonetica, 69:180-190, 2012.

[10] Fujimura, O., “The C/D model and prosodic control of articulatory behavior”, Phonetica 57:128-38, 2000.

[11] Fujimura, O., “Stress and tone revisited: Skeletal vs. melodic and lexical vs. phrasal”, in S. Kaji [Ed], Cross-Linguistic Studies of Tonal Phenomena, 221-234, Tokyo University of Foreign Studies, 2003.

[12] Bonaventura, P. and Fujimura, O., “Articulatory movements and phrase boundaries”, in M.-J. Solé, P. S. Beddor, and M. Ohala [Eds], Experimental Approaches to Phonology, 209-227, Oxford University Press, 2007.

[13] Williams, J. C., Erickson, D., Ozaki, Y., Suemitsu, A. Minematsu, N., and Fujimura, O., “Neutralizing differences in jaw displacement for English vowels”, Proc. of International Congress of Acoustics, POMA 19:060268, 2013.

[14] Menezes, C. and Erickson, D., “Intrinsic variations in jaw deviation in English vowels”, Proc. of International Congress of Acoustics, POMA 19:060253, 2013.

[15] Chomsky, N. and Halle, M., Sound Pattern of English, MIT Press, 1968.

[16] Boersma, P. and Weenink, D., “Praat: Doing phonetics by computer. [Computer Software]”, Department of Language and Literature, University of Amsterdam. Online: http://www.praat.org/, accessed on 15 Dec 2013.

[17] Erickson, D. and Honda, K., “Jaw displacement and F0 in contrastive emphasis”, J. Acoust. Soc. Am., 99, 2494 (A), 1996.

[18] Edwards, J., Beckman, M, and Fletcher, J., “The articulatory kinematics of final lengthening”, J. Acoust. Soc. Am., 89:369-382, 1991.

[19] Lieberman, P., Intonation, Perception, and Language, MIT Research Monograph 38, MIT Press, 1967.

[20] Erickson, D., Kawahara, S., Moore, J., Menezes, C., Suemitsu, A., Kim, J., Shibuya, Y., “Using the C/D Model to calculate articulatory syllable duration and prosodic boundaries”, International Symposium on Speech Processing, Cologne, Germany, 2014.

[21] Riordan, J., “Anatomy and physiology of lactation”, in J. Riordan and K. Wambach [Eds], Breastfeeding and Human Lactation, 79-116, Jones and Bartlett Publishers, 2010.

Male and female speech: a study of mean f_0, f_0 range, phonation type and speech rate in Parisian French and American English speakers

Erwan Pépiot [1]

[1] Department of Anglophone Studies, University Paris 8, France

erwan.pepiot@free.fr

Abstract

Many studies have been conducted on acoustic differences between female and male speech. However, they have generally been led on speakers of only one language, and have focused on a single acoustic parameter. The present study is an acoustic analysis of dissyllabic words or pseudo-words produced by 10 Northeastern American English speakers (5 females, 5 males) and 10 Parisian French speakers (5 females, 5 males). Several prosodic parameters were measured: mean f_0, f_0 range, phonation type (through H1-H2 intensity differences) and words' duration. Significant cross-gender differences were obtained for each tested parameter. Moreover, cross-language variations were observed for f_0 range, and H1-H2 differences. These results suggest that cross-gender acoustic differences are partly language-dependent and could be socially constructed.

Index Terms: speech and gender, fundamental frequency, phonation type, speech rate, cross-gender acoustic differences, cross-language variations, Parisian French, American English.

1. Introduction

Numerous studies on acoustic differences between female and male speech have been conducted. Among the different acoustic parameters, mean fundamental frequency is commonly considered the major cross-gender difference. It would be around 120 Hz for men and 200 Hz for women [1] [2], hence a higher pitch in female speech. These values slightly vary through age [3] and are broadly lower for smokers [4]. Mean f_0 is also known to be a decisive clue in speaker's gender identification from speech [5] [6] [7].

Several authors have brought to light that vowel formants tend to be located at higher frequencies in female speakers [8] [9] [10] [11]. The scope of this cross-gender difference strongly varies from one study to another, from one formant to another, and seems to depend on vowel type. The spectral characteristics of consonants also differ as a function of speaker's gender [12] [13] [14]: once again, resonant frequencies tend to be higher in female speech.

Aside from mean f_0, other suprasegmental parameters could be gender-dependent. Some studies suggest that f_0 range would be larger for female speakers [1] [15]. Nonetheless, there is no strict consensus on this point [16]: the acoustic unit used to measure f_0 range appears to be determining. When calculated in hertz, f_0 range is almost unequivocally larger in female speech, but it is unclear whether this difference exists when it is calculated in semitones [17] [18]. This can be accounted for by human perception of pitch [16]: female speakers, who typically have a higher mean f_0 than males, have to use a larger raw range (i.e. in hertz) to reach the same perceived pitch variation (i.e. in semitones).

Phonation type also seems to depend on speaker's gender. Female voices are often considered more breathy (i.e. having a greater *glottal open quotient –GOQ*) than male voices [19] [20] [21]. Male voices, at least in American English speakers, are typically more creaky (i.e. having a very low GOQ) than female ones [22]. However, these results slightly vary from one study to another, and depend on the acoustic parameter used to estimate phonation type. Intensity difference between H1 and H2 could be the most reliable measurement [23], if used properly [24].

Potential male-female differences in speech rate have also been investigated. In a broad study led on 600 American English speakers, Byrd [25] found that mean utterance duration was 6.2 % lower in male speakers, thus indicating a faster speech rate than female speakers. Similar tendencies were found in more recent studies [26] [27]. However, several authors found no significant cross-gender differences on this parameter [28] [29].

Some of these cross-gender acoustic variations can mainly be accounted for by anatomical and physiological differences that arise during puberty. First of all, vocal folds become longer and thicker in male speakers [30], which would account for their lower mean f_0. A second relevant anatomical parameter is vocal tract length, which corresponds to the distance from the vocal folds to the lips: all things being equal, the longer the vocal tract, the lower resonant frequencies [31]. The average length of the adult male vocal tract is about 17 to 18 cm, while the average female vocal tract is 14.5 cm long [16]. It would explain, at least partially, why consonant noise and vowel formants frequencies are generally higher in female speakers.

Most of the previously mentioned studies were conducted on English speakers. Interesting facts arise when one considers other languages' data. For instance, a study reported that in Chinese Wu dialect, mean F0 was almost equivalent for male and female speakers [32]. Furthermore, if one compares various acoustic studies about vowel formant frequencies conducted on different languages [33, 34], one can notice that cross-gender differences vary from one language to another: for example, female-male differences are relatively small in Danish but appear to be much greater in Russian.

How to account for such cross-language differences? Physiological and anatomical cross-gender differences are very unlikely to explain then, and one must consider the possibility of socially constructed behaviors. Nonetheless, we have to take into account that the comparisons made by Johnson [33, 34] were based on several studies led by different authors, at different times and using different methods. Therefore, we must be very careful when interpreting such results, which need to be confirmed.

Given such facts, it seems relevant to conduct a cross-language study on acoustic differences between female and male speech. Moreover, we can notice that most studies in this field focus on a single acoustic parameter, although a multiparametric analysis would probably be much more productive. The present study is an acoustic analysis conducted jointly on Parisian French and Northeastern American English female and male speakers. It focuses on the

following prosodic parameters: mean f_0, f_0 range, phonation type and speech rate. The general hypothesis is that cross-gender acoustic differences are partly language-dependent.

2. Material and method

2.1. Linguistic material

French and English linguistic material was required for this study. Dissyllabic words and pseudo-words were used, so that many phoneme combinations could be tested. Their selection was based on two main criteria: make the two corpora as similar as possible, and limit the number of combinations by choosing only the most relevant phonemes while holding the last CV sequence constant: /pi/ was chosen as it can appear in word final position in both languages. Twenty-seven (C)VCV words or pseudo-words were finally chosen for each language:

- /C (plosive) – V – p – i / combinations: /tipi/, /tapi/, /tupi/, /dipi/, /dapi/, /dupi/, /kipi/, /kapi/, /kupi/, /gipi/, /gapi/, /gupi/ for the French corpus, /ˈti:pi/ , /ˈtæpi/, /ˈtu:pi/, /ˈdi:pi/, /ˈdæpi/, /ˈdu:pi/, /ˈki:pi/, /ˈkæpi/, /ˈku:pi/, /ˈgi:pi/, /ˈgæpi/, /ˈgu:pi/ for the English corpus.
- /C (fricative) – V – p – i / combinations: /sipi/, /sapi/, /supi/, /zipi/, /zapi/, /zupi/, /ʃipi/, /ʃapi/, /ʃupi/, /ʒipi/, /ʒapi/, /ʒupi/ for the French corpus, /ˈsi:pi/, /ˈsæpi/, /ˈsu:pi/, /ˈzi:pi/, /ˈzæpi/, /ˈzu:pi/, /ˈʃi:pi/, /ˈʃæpi/, /ˈʃu:pi/, /ˈʒi:pi/, /ˈʒæpi/, /ˈʒu:pi/ for the English corpus.
- /V – p – i / combinations: /ipi/, /api/, /upi/ for the French corpus, /ˈi:pi/, /ˈæpi/, /ˈu:pi/ for the English corpus.

English words were read by American speakers while French words were read by French speakers. There is no phonological lexical stress in French [35], but within the frame sentence used for the recordings (see 2.3) French speakers naturally produced an emphatic stress on the first syllable of each experimental word.

2.2. Speakers

Twenty monolingual speakers were recorded. Ten of them were French native speakers (5 women, 5 men) and ten others were American English native speakers (5 women and 5 men). The 10 American speakers all came from the same northeastern area of the United States (Pennsylvania, Massachusetts, New York State, or southern Vermont). The 10 French speakers all came from Paris area (Ile-de-France). Speakers were aged from 20 to 40 (SD=6.5 years). Mean age was 28.2 for US speakers (29.4 for females, 27 for males) and 26.6 for French speakers (27.2 for females, 26 for males). All speakers were non-smokers and had reported no speech disorder. Each of them received a USB memory stick for their participation in the study and was informed that the data from the recordings would be treated with confidentiality.

2.3. Recording procedure

Recordings took place in a quiet room, using a digital recorder *Edirol R09-HR* by *Roland*. English speakers read the English corpus aloud and French speakers the French one. Words were presented to the participants in an orthographical transcription. In order to make prosodic parameters consistent, words were placed into a frame sentence: "He said '*WORD'* twice" for the English corpus and "Il a dit '*MOT'* deux fois" for the French one. Speakers were asked to say each sentence twice, at a normal speech rate.

2.4. Acoustic analysis

Data analysis was conducted with *Praat* software. After having extracted the words from the frame sentence, their duration (in milliseconds) and mean f_0 (in Hertz) were obtained by creating a Pitch file for each word, and performing *Get total duration* and *Get mean* commands. This operation was automated by a Praat script.

F_0 range is the difference between the highest and the lowest f_0 frequency reached within a given linguistic unit (here, a dissyllabic word). It was collected manually through the *Pitch info* window: these data were taken in hertz as well as in semitones, which is a much more adequate scale [16].

In order to estimate phonation type, intensity differences between H1 and H2 were measured. The relative strength of H1 is correlated with glottal open quotient (GOQ): the stronger H1 is, the higher the GOQ [19, 23]. Nevertheless, H1-H2 can only be measured on open vowels: F1 would otherwise distort the results [19]. Thus, vowel [a] for French speakers and vowel [æ] for English speakers were the only ones taken into account.

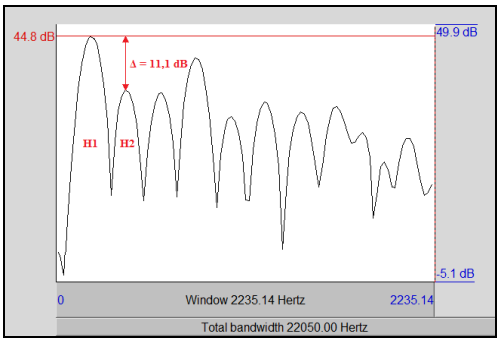


Figure 1. *Measurement of H1-H2 intensity differences on vowel [æ] extracted from word [æpi] produced by an American Female speaker.*

A 5 period selection was made on a central portion of each vowel. As shown in Figure 1, the corresponding spectrum was displayed and the difference between H1 and H2 intensity (in dB) was then calculated manually.

3. Results

3.1. Mean f_0

Mean f_0 (Hz) for French and American English speakers as a function of speaker's gender is presented in Table 1, below.

Table 1. *Mean f_0 (Hz) measured on the 27 (C)VCV words for female (n=5) and male (n=5) French speakers and female (n=5) and male (n=5) American English speakers. Standard deviation among the 135 measurements (27 words * 5 speakers) is also mentioned for the four groups.*

	French Speakers		**American speakers**	
	Females	**Males**	**Females**	**Males**
Mean f_0 (Hz) - all words	**234**	**133**	**210**	**119**
SD	*18*	*12*	*27*	*19*

Unsurprisingly, mean f_0 appeared to be much higher for female speakers in both languages. The scope of this cross-gender difference is perfectly similar from one language to another: in both cases, females' mean f_0 is 76 % higher than males'. Moreover, we can notice that mean f_0 for both genders is slightly lower in American English speakers.

In order to test if these tendencies were significant, several statistical tests were conducted. First of all, a one factor ANOVA ("speaker's gender") was led on French speakers' data. The test revealed a very strong and significant effect of this factor: $F_{(1,268)}$=3064.26 ; p<.0001. A similar statistical test was performed on American English speakers' data. Once again, the speaker's gender appeared to have a very significant effect on mean f_0: $F_{(1,268)}$=1045.21 ; p<.0001. These results confirm that mean f0 was significantly higher for female speakers in both languages.

3.2. F_0 range

F_0 range for French and American English speakers as a function of speaker's gender is presented in Table 2. It was calculated both in Hertz and in semitones.

Table 2. *F_0 range (Hz and st) measured on the 27 (C)VCV words for female (n=5) and male (n=5) French speakers and female (n=5) and male (n=5) American English speakers. Standard deviation among the 135 measurements (27 words * 5 speakers) is also mentioned for the four groups.*

	French Speakers		**American speakers**	
	Females	**Males**	**Females**	**Males**
Mean f_0 range in Hertz - all words	**90**	**41**	**74**	**40**
SD	*22*	*11*	*19*	*15*
Mean f_0 range in semitones - all words	**6.76**	**5.35**	**5.87**	**5.95**
SD	*1.81*	*1.37*	*1.45*	*1.78*

When calculated in Hertz, f_0 range was much larger for female speakers in both languages. It was 120 % wider for females in French speakers, and 85 % wider for females in American English speakers. When the data are converted into semitones, a scale that shows the perceived pitch variation, the cross-gender difference completely disappears in American English speakers. On the other hand, f_0 range remains substantially higher for females in French speakers (+ 26 %).

One factor ANOVAs ("speaker's gender") were conducted on French speakers' data. When f_0 range was calculated in Hertz, a very strong and significant effect of this factor was found: $F_{(1,268)}$=549.19 ; p<.0001. When the data were given in semitones, the analysis also reveals a strong and significant effect of the speaker's gender: $F_{(1,268)}$=51.71 ; p<.0001. Therefore, in French speakers, mean f_0 range appeared to be significantly wider for females, even when it was expressed in semitones.

Similar statistical analyses were made on American English speakers' data. When f_0 range was expressed in Hertz, the test indicated a significant effect of speaker's gender: $F_{(1,268)}$=266.23 ; p<.0001. On the other hand, when the data were expressed in semitones, no significant effect of this factor was found: $F_{(1,268)}$=.14 ; p=.71. These results confirm that f_0 range was wider in female than in male American English speakers when expressed in hertz. However, contrary to French speakers, f_0 range was no longer wider in female speakers when it was converted into semitones.

3.3. Phonation type

Mean intensity difference (dB) between the first (H1) and the second harmonic (H2) for French and American English speakers as a function of speaker's gender is presented in Table 3. It was measured on open vowels, giving a total of 9 measurements for each speaker (9 words contained vowel [a] in the French corpus while 9 words contained vowel [æ] in the English corpus).

Table 3. *H1-H2 mean difference (dB) measured on the 9 open vowels of each corpus, for female (n=5) and male (n=5) French speakers and female (n=5) and male (n=5) American English speakers. Standard deviation among the 45 measurements (9 words * 5 speakers) is also mentioned for the four groups.*

	French Speakers		**American speakers**	
	Females	**Males**	**Females**	**Males**
Mean H1-H2 difference (dB) in open vowels	**3.8**	**-1.4**	**4.0**	**-2.9**
SD	*2.5*	*1.3*	*2.6*	*2.2*

H1-H2 intensity difference appeared to be greater in female speakers. It was true in both languages. This cross-gender difference reaches 5.2 dB in French speakers and 6.9 dB in American English speakers. This cross-language variation is mainly due to American English male speakers, who had a particularly weak H1.

A one factor ANOVA ("speaker's gender") was performed on French speakers' data. The analysis revealed a significant effect of this factor over H1-H2 intensity difference: $F_{(1,88)}$=157.22 ; p<.0001. The same statistical test was conducted on American English speakers' data. Similarly to French speakers, this test indicated that there was a significant effect of speaker's gender: $F_{(1,88)}$=180.04 ; p<.0001. These results confirmed that in both languages H1-H2 difference was higher in female speakers, which suggests that females' phonation type tends to be more breathy than males'.

In order to test if cross-language variations were significant, other statistical tests were conducted. Females' data from both languages were gathered and a one factor ANOVA ("speaker's language) was performed. The analysis showed no significant effect of this factor: $F_{(1,88)}$=.11 ; p=.74. A similar test was conducted on males' data. This time, a significant effect of speaker's language was found: $F_{(1,88)}$=13.55 ; p=.0004. This analysis confirmed that American male speakers had a significantly lower H1-H2 intensity difference than French male speakers, which suggests they used a creakier phonation type.

3.4. Speech rate

Mean word duration (ms) for French and American English speakers as a function of speaker's gender is presented in Table 4, below.

Table 4. *Mean word duration (ms) measured on the 27 (C)VCV words for female (n=5) and male (n=5) French speakers and female (n=5) and male (n=5) American English speakers. Standard deviation among the 135 measurements (27 words * 5 speakers) is also mentioned for the four groups.*

	French Speakers		**American speakers**	
	Females	**Males**	**Females**	**Males**
Mean word duration (ms)	**510**	**445**	**555**	**441**
SD	*90*	*58*	*77*	*54*

Results show that mean word duration was higher for female speakers in both languages. Cross-gender difference is wider in American English speakers (+26 %) than in French speakers (+15 %). This variation can be accounted for by the difference between French and American Female speakers: the words produced by the latter appeared to be longer than those produced by the former (+9 %).

A one factor ANOVA ("speaker's gender") was conducted on French speakers' data. This test indicated that there is a significant effect of this factor on words' mean duration: $F_{(1,268)}$=48.94 ; p<.0001. The same analysis conducted on American English speakers' data reveals a strong and significant effect of speaker's gender: $F_{(1,268)}$=200.28 ; p<.0001. These results confirm that speech rate was significantly higher for male speakers in both languages.

4. Discussion - conclusions

This cross-language acoustic analysis has given several remarkable results. Mean fundamental frequency, measured on dissyllabic words, was significantly higher for women in both languages, which broadly confirms results obtained in previous studies [1] [2]. Moreover, even though mean f_0 was slightly lower for both genders in American English speakers, the scope of the female-male difference was very similar in the two languages. This suggests that cross-gender differences in mean f_0 are relatively consistent across languages, apart from a few known exceptions, such as Chinese Wu dialect, in which male speakers tend to use an exceptionally high f_0 [32].

Regarding f_0 range, measurements have highlighted an interesting cross-language variation. When data were taken in hertz, f_0 range appeared to be significantly wider for female speakers in both languages, which supports results from previous studies [1] [15]. This cross-gender difference can be accounted for by the fact that female speakers, who generally have a higher mean f_0, have to use a larger f_0 range than males (in hertz) to achieve the same perceived result in terms of pitch variation.

Indeed, when f_0 range was calculated in semitones, a scale expressing perceived pitch variation, there was no more cross-gender difference in American English speakers. However, f_0 range was still significantly larger for French female speakers compared to their male counterparts. These results clearly support a former perceptual study from Pépiot [5] that showed a tendency for French listeners to associate flat f_0 with male voices, whereas no such effect was found in American English listeners.

H1-H2 intensity measurements in open vowels gave precious indications about speakers' phonation type. In both languages, significant cross-gender differences were found. H1's relative intensity appeared to be significantly greater in female than in male speakers, which suggests they tend to speak with a greater GOQ, hence a more breathy phonation type. Moreover, American English male speakers had a significantly lower H1-H2 than French male speakers, with strongly negative values. This indicates a very low GOQ, hence a creakier voice. Such results support the claim that female speakers' breathy voice quality would have a physiological origin [16], whereas male speakers' use of creaky voice would rather be socio-phonetic and language-dependant [22].

Concerning temporal measurements, dissyllabic words' global duration was found to be significantly greater for female speakers in both languages. These data are quite similar to former results obtained by Byrd [25] and Whiteside [27]. Cross-gender difference was slightly wider in American English speakers. Nonetheless, these results may suggest that the lower speech rate in female speakers, at least in a reading task, could be fairly consistent across languages.

Overall, the present study has brought to light several cross-gender differences, but also some cross-language variation between Parisian French speakers and Northeastern American English speakers. This tends to support the general hypothesis which claimed that cross-gender acoustic differences could be partly language-dependent. Furthermore, some of the female-male differences found in this acoustic analysis, such as differences in speech rate, are unlikely to be explained by anatomical and physiological factors. A large part of cross-gender variation is likely to be accounted for by gender social construction. Therefore, these data may be of interest for improving vocal rehabilitation of transgender people [36]. They could also be useful in speech recognition and automatic gender identification from speech.

Nevertheless, such results have to be interpreted with caution. Only 5 men and 5 women were recorded for each language. Despite the restrictive selection criteria and the small intra-gender variation, it seems quite difficult to generalize the results to the whole Parisian French and American English speaker populations. Furthermore, it is known that speech task influences several acoustic parameters, such as speech rate and phonation type [37]. These corpora were made of read dissyllabic words: it is unclear whether similar results would be obtained with spontaneous speech.

5. Acknowledgements

I would like to thank the 20 speakers who agreed to participate in the recordings. I am also grateful to the *EA1569* laboratory of University Paris 8, which provided financial support for this study. Finally, many thanks to Isabelle Coydon, from NYU Paris, who was of great help for finding the American English speakers.

6. References

[1] Takefuta, Y., Jancosek, E. G., & Brunt, M., "A statistical analysis of melody curves in the intonation of American English", Proceedings of the 7th International Congress of Phonetic Sciences - Montreal (1971), 1035-1039, 1972.

[2] Boë, L-J., Contini, M., & Rakotofiringa, H., "Étude statistique de la fréquence laryngienne", Phonetica, 32: 1-23, 1975.

[3] Pegoraro-Crook, M. I., "Speaking fundamental frequency characteristics of normal Swedish subjects obtained by glottal frequency analysis", Folia Phoniatrica, 40: 82-90, 1988.

[4] Gilbert, H. R. & Weismer, G. G., "The effects of smoking on the speaking fundamental frequency of adult women", Journal of Psycholinguistic Research, 3: 225-231, 1974.

[5] Pépiot, E., "Voix de femmes, voix d'hommes : à propos de l'identification du genre par la voix chez des auditeurs anglophones et francophones", Plovdiv University "Paissii Hilendarski" – Bulgaria, Scientific Works – Philology, 49: 418-430, 2011.

[6] Pausewang Gelfer, M., & Mikos, V., "The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels", Journal of Voice, 19: 544-554, 2005.

[7] Coleman, R. O., "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice", Journal of Speech and Hearing Research, 19: 168-180, 1976.

[8] Peterson, G. E., & Barney, H. L., "Control methods used in a study of the identification of vowels", Journal of the Acoustic Society of America, 24: 175-184, 1952.

[9] Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K., "Acoustic characteristics of American English vowels", Journal of the Acoustic Society of America, 97: 3099-3111, 1995.

[10] Whiteside, S. P., "Sex-specific fundamental and formant frequency patterns in a cross-sectional study", Journal of the Acoustic Society of America, 110: 464-478, 2001.

[11] Ericsdotter, C., Articulatory-acoustic relationships in Swedish vowel sounds, PhD Thesis, Stockholm University, 2005.

[12] Schwartz, M. F., "Identification of speaker sex from isolated voiceless fricatives", Journal of the Acoustic Society of America, 43: 1178-1179, 1968.

[13] Nittrouer, S., "Children learn separate aspects of speech production at different rates: Evidence from spectral moments", Journal of the Acoustic Society of America, 97: 520-530, 1995.

[14] Jongman, A., Wayland, R., & Wong, S., "Acoustic characteristics of English fricatives", Journal of the Acoustic Society of America, 108: 1252-1263, 2000.

[15] Hwa Chen, S., "Sex differences in frequency and intensity in reading and voice range profiles for Taiwanese adult speakers", Folia Phoniatrica et Logopaedica, 59: 1-9, 2007.

[16] Simpson, A. P., "Phonetic differences between male and female speech", Language and Linguistics Compass, 3: 621-640, 2009.

[17] Henton, C. G., "Fact and fiction in the description of female and male pitch", Language and Communication, 9: 299-311, 1989.

[18] Herbst, L., "Die Umfänge der physiologischen Hauptsprechtonbereiche von Frauen und Männern", Zeitschrift für Phonetik und Sprachliche Kommunikation, 22: 426-438, 1969.

[19] Klatt, D. H. & Klatt, L. C., "Analysis, synthesis and perception of voice quality variations among female and male talkers", Journal of the Acoustical Society of America, 87: 820-857, 1990.

[20] Klatt, D. H., "Detailed spectral analysis of a female voice", Journal of the Acoustical Society of America, 80: S97, 1986.

[21] Henton, C. G. & Bladon, R. A., "Breathiness in normal female speech: Inefficiency versus desirability", Language and Communication, 5: 221-227, 1985.

[22] Henton, C. G., "Sociophonetic aspects of creaky voice", Journal of the Acoustical Society of America, 86: S26, 1989.

[23] Gordon, M. & Ladefoged, P., "Phonation types: A crosslinguistic overview", Journal of Phonetics, 29: 383-406, 2001.

[24] Simpson, A. P., "The first and second harmonics should not be used to measure breathiness in male and female voices", Journal of Phonetics, 40: 477-490, 2012.

[25] Byrd, D., "Relations of sex and dialect to reduction", Speech Communication, 15: 39-54, 1994.

[26] Fitzsimons, M., Sheahan, N. & Staunton, H., "Gender and the integration of acoustic dimensions of prosody: implications for clinical studies", Brain and Language, 78: 94-108, 2001.

[27] Whiteside, S. P., "Temporal-based acoustic-phonetic patterns in read speech: Some evidence for speaker sex differences", Journal of the International Phonetic Association, 26: 23-40, 1996.

[28] Ray, G. B. & Zahn, C. J., "Regional speech rates in the United States: A preliminary analysis", Communication Research Reports, 7: 34-37, 1990.

[29] Simpson, A. P. & Ericsdotter, C., "Sex-specific durational differences in English and Swedish", Proceedings of the 15th International Congress of Phonetic Sciences - Barcelona, 1113-1116, 2003.

[30] Kahane, J., "A morphological study of the human prepubertal and pubertal larynx", American Journal of Anatomy, 151: 11-20, 1978.

[31] Fant, G., Acoustic Theory of Speech Production (2nd ed.), Mouton, 1970.

[32] Rose, P., "How effective are long term mean and standard deviation as normalization parameters for tonal fundamental frequency?", Speech Communication, 10: 229-247, 1991.

[33] Johnson, K., "Speaker normalization in speech perception", in D. Pisoni, & R. Remez, R. [Eds], The Handbook of Speech Perception, 363-389, Blackwell Publishers, 2005.

[34] Johnson, K. "Resonance in an exemplar-based lexicon: the emergence of social identity and phonology". Journal of Phonetics, 34: 485-499, 2006.

[35] Di Cristo, A., "Vers une modélisation de l'accentuation du français : première partie", Journal of French Language Studies, 9: 143-179, 1999.

[36] Wiltshire, A., "Not by pitch alone: a view of transsexual vocal rehabilitation", National Student Speech Language Hearing Association Journal, 22: 53-57, 1995.

[37] Benoist-Lucy, A., Pillot-Loiseau, C., "The Influence of language and speech task upon creaky voice use among six young American women learning French", Proceedings of Interspeech 2013 - Lyon, 2395-2399, 2013.

Perceived Prominence Reflected by Imitations of Words with and without F0 Continuity

Hansjörg Mixdorff[1], *Angelika Hönemann*[1], *Oliver Niebuhr*[2] *and Christoph Draxler*[3]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Department of General Linguistics, ISFAS, Christian-Albrecht-University of Kiel, Germany
[3] Institute of Phonetics und Speech Processing, Ludwig-Maximilian-University Munich, Germany

{mixdorff|ahoenemann}@beuth-hochschule.de, niebuhr@linguistik.uni-kiel.de,
draxler@phonetik.uni-muenchen.de

Abstract

This paper continues our work on the perception of prominence as a function of *F0* continuity. In an earlier study the first author had shown that *F0* intervals occurring at lexically stressed syllables – and measured using the amplitude of Fujisaki model accent commands – strongly contribute to the perceived prominence of that syllable. More recent work explored how *F0* continuity influenced prominence ratings of single word utterances. The outcome indicated that listeners made use of the physically available *F0* information and therefore words containing gaps in the contour were perceived as less prominent. It was also shown that subjects were able to interpolate missing parts as long as the *F0* peak was still present. The current study explores whether subjects compensate the lack of prominence in words containing *F0* gaps by asking them to produce a word with the same accent strength as that of a spoken word stimulus, the spoken word being either the same or different from the one they are asked to utter. We evaluated word durations, *F0* intervals and intensities of the responses as correlates of prominence and found that listeners indeed seem to adjust depending on the kind of stimulus they have heard.

Index Terms: prominence, perception, Fujisaki model, *F0*

1. Introduction

The information structure of an utterance is reflected in the relative saliency of its lexical constituents. At the acoustic level we observe that accented syllables serve as anchoring points of this structure. They are emphasized or toned down by phonetic means. The perceptual correlate of this process is the so-called prominence [1]. Various segmental and supra-segmental factors have been shown to affect prominence, cf. [1][2][3], such as F0 excursions, segment durations, intensity as well as vowel type and syllable coda structure. In an earlier study [4], the first author and his co-worker investigated the relationship between perceived syllable prominence and the *F0* contour in terms of the parameters of the Fujisaki model [5]. The model was used to parameterize a subcorpus of the Bonn Prosodic Database [6]. Analysis showed that prominences labeled on a scale from 0-31 strongly correlated with the excursion of *F0* movements, but only when it was anchored to accented syllables. This indicates that the prominence judgment is partly guided by linguistic considerations. Evidence in support of this assumption has been presented for many languages, including German ([7],[8], [9]), which is the language of the present study.

While we are well aware that other factors such as syllable duration [4] influence prominence perception, we concentrate in the current work on the contribution of *F0*. As mentioned above the Fujisaki model decomposes natural *F0* contours with a defined value for each speech frame, irrespective of segmental structure. This entails that it smoothly interpolates or extrapolates *F0* gaps owing to unvoiced sounds. However, from a communicative point of view, the implicit claim of using the same underlying prosodic gesture for voiced and unvoiced sound sections is that listeners are *also* able to interpolate or extrapolate *F0* gaps. Recent evidence from a tonal scaling study [10] seemed inconsistent with this assumption.

Subjects were presented with short resynthesized utterances and asked to rate the tonal height of accent-related *F0* rises. The rises led to a peak that was either present or absent due to an unvoiced stop consonant. Tonal height ratings were made and analyzed relative to reference utterances in which the *F0* rise was replaced by a flat *F0* stretch, yielding a constant tonal height. The findings of [10] suggested that the subjective continuity of pitch contours in speech is due to the fact that the auditory system simply ignores rather than fills *F0* gaps.

In [11] we therefore examined the implications of these findings for the perception of word prominence. We investigated how gaps in the *F0* contour due to unvoiced consonants affect prominence perception, given that such gaps can either be filled or blinded out by listeners. For this purpose we created a stimulus set of real disyllabic words which differed in the quantity of the vowel of the accented syllable nucleus and the types of subsequent intervocalic consonant(s). Participants rated pairs of these stimuli in a 2AFC discrimination task for accent strength, that is, they decided which word in a pair sounded more strongly accented. Results included, inter alia, that stimuli with unvoiced gaps in the *F0* contour are indeed perceived as less prominent. The prominence reduction was smaller for monotonous stimuli than for stimuli with *F0* excursions across the accented syllable. Moreover, in combination with *F0* excursions, it also mattered whether *F0* had to be interpolated or extrapolated, and whether or not the gap included a fricative sound. The results supported both the filling-in and blinding-out of *F0* gaps, which fits in well with earlier experiments on the production and perception of pitch.

The current experiment examines whether speakers compensate for the inherent difference in prominence when they produce words with or without continuous *F0* contours. To this end we asked participants of a production experiment to reply to an acoustic stimulus of an isolated word and speak either the same or a different word with the same accent strength. Our hypothesis are: (1) if speakers indeed compensate for missing F0 parts and their prominence contribution when they reproduce a given accent strength, then they will realize higher *F0* targets in low-sonority than in high-sonority words in which *F0* is more/ all F0 is present; (2) replicating our previous findings, acoustic stimuli in which *F0* is either continuous *or* can be interpolated or filled-in will make speakers produce higher *F0* targets.

2. Stimuli and Experiment Design

The acoustic stimuli were taken from the set employed in [11]. They are shown in Table 1 with their critical segments set in bold in the SAMPA column. As can be seen, the critical segment extends from the beginning of the open vowel to the end of the intervocalic consonant(s).

Table 1. *Five target words, SAMPA transcription, English translation, type and mean energy of the critical segment.*

Word	SAMPA	English	critical segment	mean energy (dB)
Rahmen	[R**a:m**@n]	frame	long vowel (LV), voiced (vcd) nasal	74.23
Rasen	[R**a:z**@n]	lawn	LV, vcd fricative	72.10
Raten	[R**a:t**@n]	guess	LV,voiceless (vcl) plosive	68.10
Ras(s)ten	[R**ast**@n]	rest	short vowel (SV), vcl fricative+plosive	66.19
Ratten	[R**at:**@n]	rats	SV, long vcl plosive	50.68

For acoustic uniformity, the stimuli had been created using the *MBROLA* concatenative speech synthesizer driving the German male voice *de8* [12]. The base stimulus had a monotonous *F0* at 100Hz. The long vowel [a:] was adjusted to 244ms and the central consonant portion to 126ms. In the short-vowel words the V/C portions were 142ms and 228ms. These critical segment durations and the word durations, respectively, were kept constant for all stimuli, in order to avoid that durational differences influenced the prominence judgments.

Using the *FujiParaEditor* [13] and Praat PSOLA resynthesis [14] we created further stimuli by adding *F0* peak contours to the monotonous stimuli. The contour basis was laid by a phrase component, constant for all stimuli. One accent component with duration of 200ms was superimposed on the base contour. Unlike in [11] we decided to employ only medial-peak stimuli in the current study. Their *F0* maxima were aligned close to the accented-vowel offset in the long-vowel words and in the following coda in the short-vowel words, in line with previous findings [15] and observations in citations forms.

Figure 1 displays the stimuli Rahmen, Raten and Ras(s)ten with *Aa*=0.6, *Aa* being the amplitude of the underlying Fujisaki model accent command which corresponds to the interval of the *F0* excursion associated with each word (see box-shaped accent command at the bottom of Figure 1. The F0-peak range was varied in three steps, represented by three different accent command amplitudes (*Aa*): 0.4 (interval of approximately 3 semitones), 0.6 (about 3 semitones higher) and 0.8 (about 6 semitones higher). Hence, including the monotonous condition, we generated four different acoustic versions of every word and hence a total of 100 acoustic stimulus/text pairs.

The experiment was conducted using WikiSpeech [16], a framework developed at Ludwig-Maximilian-University Munich for web-based perceptual testing and speech data collection. Participants were asked to enroll on the WikiSpeech-Website and accept the download of the Speech Recorder audio recording tool. When participants executed the program on their computers, the task was explained to them on the start-up screen. Every trial consisted of the automatic playback of the acoustic stimulus, that is, one of the synthetic stimuli, followed by the display of the word to be produced written as text. A traffic light indicated when to speak. The duration of the recording slot was fixed at five seconds. After having recorded the subject's response to the given synthetic stimulus, the experiment immediately continued with the next stimulus. Each of the acoustic words was either paired with its text equivalent or that of one of the other target words. Subjects were asked to pronounce the text word with the same accent strength as the acoustic stimulus.

Prior to the experiment the subjects listened to two examples in which of an acoustic stimulus was followed by the reply of a dummy subject. Subsequently a training session started which four pairs of the same or different words had to reproduced by the subjects with the same accent strength.

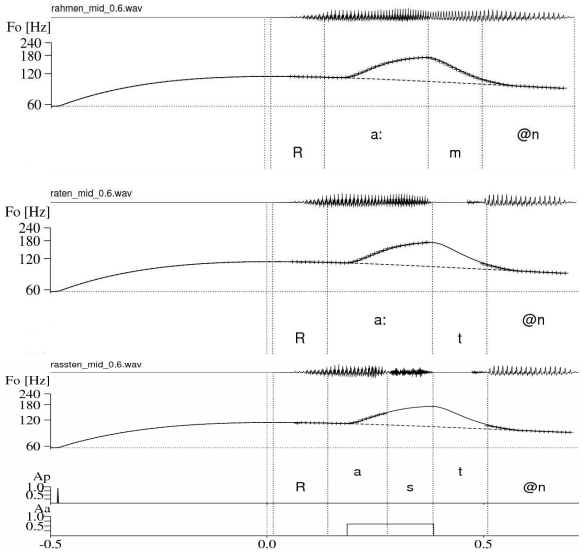


Figure 1. Examples of stimulus words Rahmen, Raten and Ras(s)ten with Aa=0.6. Panels display waveform (top), F0 contour (+++extracted, —Fujisaki model-based, middle), and underlying phrase/accent commands (bottom).

3. Acoustic Analysis of Stimuli

We recorded 19 native speakers of German (10 females, 9 males, 22-46 years old), most of them students at Beuth University Berlin or Kiel University. Since the experiment was web-based, we had no control over the recording equipment and environments. Although we had requested the use of headsets, many participants apparently used built-in microphones picking up environmental noise. Thus, the quality of recordings varied considerably with a wide range of gain settings, background noise and audible audio compression effects. But as long as the recording quality was sufficient to reliably extract prosodic parameters, and F0 in particular, we included data which would not have qualified for other acoustic analyses.

We first checked the audio files for the correct intended word, then admitted data sets with more than 70% correct word replies to further analysis, excluding, of course, all erroneous tokens within these sets. As a consequence, two participants were entirely excluded (1 female, 1 male), leaving a total of 17 data sets or 1,700 recordings.

Of the remaining sets 1,561 contained the correct word (92%). Only three of the subjects were a 100% successful. Errors chiefly concerned the words [Rast@n] and [Rat@n]. This was because, [Rast@n] (*"to rest"*) and [Ra:st@n] (past tense of *"to speed"*) have the same spelling <Rasten> in German. Since we had anticipated this interference we had spelled [Rast@n] in the non-standard way 'Rassten' to indicate the short vowel. However, the presence of 'Rasen' (infinitive

of "to speed", in addition to *"lawn"*) in the data set possibly triggered errors, especially when a long-vowel word had to be reacted to. Eleven of the participants produced [Rast@n] wrongly, in a total of 79 trials. [Rat@n] exhibited 38 two-way confusions with [Ra:t@n]. In most cases the vowel quantity of the acoustic stimulus matched the one produced erroneously. In contrast, participants reacted to the words [Ra:m@n] and [Ra:z@n] without difficulties. A few cases of errors also concerned empty audio files (9), stuttered words (8) or repetition of the acoustic stimulus word (5) when this was not requested.

After having selected the analyzable sets, their recordings were down-sampled to 16 kHz, and the first 0.85 seconds removed as they often contained audible traces of the acoustic stimulus. These traces disturbed the subsequently conducted forced alignment with the WEVOSYS LINGWAVES aligner [17]. Automatic phone segmentations were checked and corrected manually in the *PRAAT* TextGrid Editor [14].

We calculated word, syllable and phone durations in *PRAAT* based on the segmentations. *F0* values were extracted at intervals of 10 ms with *F0* floors and ceilings for male (50-300Hz) and female participants (120-400Hz). All *F0* contours were then subjected to Fujisaki model parameter extraction [18]. Results were checked and if necessary corrected in the *FujiParaEditor* [19]. In this way, we obtained a smooth, interpolated model *F0* contour even when the natural *F0* contour was interrupted by unvoiced segments. Hence the accent command amplitude *Aa* is a measure of the underlying *F0* gesture magnitude.

Intensity contours were extracted in *PRAAT* with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each phone.

In order to analyze our data with respect to the research question, it was sufficient use only the following four Fujisaki model parameters as dependent variables: *Aa* (accent command amplitude, as a measure of magnitude of *F0* excursion at the accented syllable), *Fb* (base frequency of the *F0* pattern), *Ap* (phrase command magnitude, a measure of the magnitude of *F0* reset before phrase onset), and *T1rel* (the timing of the accent command onset relative to the onset of [a:] or [a:] in the first, accented syllable). Each word exhibited one phrase and one accent command, like the acoustic stimuli in Figure 1. *F0* con-tours of several reactions to monotonous stimuli were absolutely flat, so that neither a phrase nor an accent command was extracted.

The four dependent variables provided by the Fujisaki model were complemented by five dependent derived from the segmentation and measurement procedures conducted with PRAAT. These additional five dependent variables were *duration syllable 1*, *duration syllable 2*, as well as *vowel duration*, *mean vowel intensity*, and *maximum vowel intensity*. The latter three concerned the accented vowel [a:] or [a].

It should be stressed prior to the results presentation that we did not aim to perceptually evaluate whether participants had actually succeeded in producing words of equal prominence, but explore the patterns in their reactions to the acoustic stimuli depending on the magnitude of *F0* excursion and type of word. Our only assumption is that subjects reacted to the stimuli as requested and aimed at producing equal prominences.

4. Results of Analysis

Our 4+5=9 dependent variables were analyzed statistically with a three-way multivariate ANOVA based on the fixed factors *Word Heard*, *Word Realized*, and *F0 Range Heard*. The latter factor included the four *Aa* levels of the acoustic stimuli, i.e. 0.0 (monotonous), as well as 0.4, 0.6, and 0.8 (henceforth "*F0* peak conditions"). The other two factors had five levels each, corresponding to the disyllabic target words Rahmen, Rasen, Raten, Ras(s)ten (henceforth Rasten), and Ratten.

All three fixed factors had significant effects on many dependent variables. Our results section can only summarize a subset of these findings; and since the monotonous conditions differ considerably from the *F0* peak conditions, we will deal with the two conditions separately, starting with the *F0* peak conditions.

The most important finding in the *F0* peak conditions was that the *Aa* levels in the speakers' productions were highly significantly affected by *F0 Range Heard* ($F[3,1296]= 7.582$, $p<0.001$). More specifically, the *Aa* levels produced by the participants clearly mirrored those of our acoustic stimuli, independently of the target word (cf. Figure 2, center). The *Aa* level 0.4 was on average even exactly reproduced, whereas the higher *Aa* levels 0.6 and 0.8 of the acoustic stimuli were slightly underestimated and on average produced as 0.5 and 0.6.

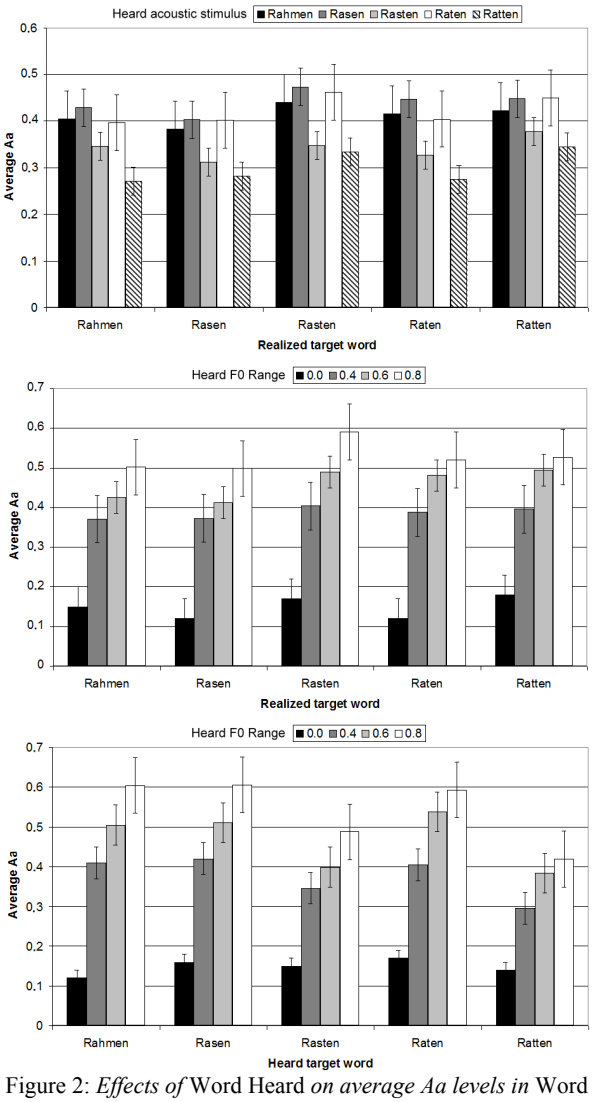


Figure 2: *Effects of* Word Heard *on average Aa levels in* Word Realized *(top), as well as combined effects of* F0 Range Heard *and* Word Realized *(center) or* Word Heard *(bottom) on the produced average Aa levels.*

Next to this general parallelism between heard and actually produced *Aa* levels, the *Aa* levels in the speakers' productions showed target word-specific differences. This is reflected in the MANOVA in a significant interaction between *Word Heard*

and *F0 Range Heard* (F[12,1296]=3.813, p<0.001). However, there was also a separate significant main effect of *Word Heard* (F[4,1296]= 35.522, p<0.001). Breaking down this main effect in terms of multiple post-hoc comparisons with Sidak correction revealed that our speakers produced higher *Aa* levels across all target words when the *F0* peaks they heard came from the more sonorous stimuli Rahmen, Rasen, and Raten rather than from the less sonorous stimuli Rasten and Ratten. The *Aa* level difference was on average 0.1 (p<0.001 in all cases, 278 ≤ n ≤ 286, cf. Figure 2, bottom). Moreover, the latter two words Rasten and Ratten yielded an additional weak *Aa* difference (p<0.05) with the Rasten stimuli triggering significantly higher *A*a level productions (of about 0.05) than the Ratten stimuli.

Exactly the opposite pattern led to a separate significant main effect of *Word Realized* (F[4,1296]= 6.829, p<0.001). We analyzed this effect by means of multiple post-hoc comparisons with Sidak correction (p<0.05 in all cases, 224 ≤ n ≤ 305, cf. Figure 2, top) and found that the *F0* peaks in the less sonorous target words Rasten and Ratten were produced with an *Aa* on average about 0.7 points higher than the more sonorous target words Rahmen, Rasen, and Raten (p<0.05 in all cases, cf. Figure 2, top and center). So, while successive desonorization and/or devoicing caused a reduction of the *F0* range at the level of perception, it triggered an extension of the *F0* range at the level of production. It is not a usual finding in the area of segment-related microprosodic perturbations that production and perception findings go in opposite directions. The present outcome suggests the existence of a compensatory strategy in *F0* production.

The speakers' productions after monotonous stimuli differed considerably from those after the *F0* peak stimuli. The nature of these differences suggests that the speakers used a stylized, singing speech mode when producing target words after monotonous stimuli. For example, *Fb* (i.e. the base *F0*) was significantly higher and *Ap* significantly smaller after monotonous than after all other stimuli, which is among others reflected in a main effect of *F0 Range Heard* (F[3,1296]= 20.752, p<0.001; F[12,1296]= 103.065, p<0.001). Furthermore, we found with respect to *F0 Range Heard* that target word productions after monotonous stimuli were softer in terms of a lower *maximum vowel intensity* (F[3,1296]= 2.253, p<0.05) and longer due to increased durations in both the first and especially the second syllable (*duration syllable 1*: F[3,1296]= 2.169, p<0.05; *duration syllable 2*: F[3,1296]= 5.095, p<0.001).

Besides these major findings, *Word Heard* resulted in an interesting further effect, which could be characterized as "transfer" or "echo effect" of the acoustic stimuli on the produced target words. For example, after hearing the short-vowel stimuli Rasten and Ratten, speakers produced their target words with shorter first syllables and vowels, independently of the target word or the quantity of its accented vowel (F[4,1296]= 10.965, p<0.001; F[4,1296]= 17.097, p<0.001). So, even Rahmen was produced with shorter first syllables and vowels when preceded by a stimulus like Rasten or Ratten. Finally, a significant main effect of *Word Realized* on *T1rel* (F[3,1296]= 5.309, p<0.001) replicated known effects of syllable structure and/or consonant type on *F0* peak timing [20], for example, in the form of an earlier timing relative to the vowel onset with increasing desonorization of the produced target words.

Last but not least, we have no indications from randomly composed sub-samples for speaker- or gender-specific effects on the crucial *Aa* variable. However, other duration and intensity variables show differences, pointing to individual F0-peak timings (cf. [24]), speaking rates or loudness levels, the latter of which could also be due to the different recording equipments.

5. Discussion and Conclusions

This paper presented results from an imitation study comparing reactions to word stimuli with either the same or a different word. Effects of *F0* on prominence perception are usually investigated with continuously voiced speech material. But what happens in more natural speech conditions, i.e. when parts of prominence-related *F0* movements are missing due to interruptions by voiceless sound segments? Do speakers and/or listeners compensate for the missing *F0* sections? The answers to these questions provided by the present findings are consistent with our hypotheses (1) and (2).

That is, in accord with previous findings on English [10], our speakers' responses to the stimuli suggest that *F0* gaps are ignored rather than filled so that word intonations in which the peak maximum and/or adjacent high *F0* section are missing sounded lower and were hence imitated with lower *Aa* on the same or other words. As in our previous study, voiceless fricatives, here the Rasten/Ratten distinction, seem to be an exception. Speakers produced higher *Aa* as a reaction to Rasten stimuli than to Ratten stimuli, which suggests that they also heard more of the prominence-related *F0* peak in the Rasten stimuli, although the *F0* gap was physically equally long as in Ratten. That is, it seems that voiceless fricatives lend themselves better to a perceptual fill-in of *F0* gaps. This matches well with the notion of "segmental intonation", developed with reference to *F0* adjusted fricative productions by the third author [21].

However, even though *F0* sections masked by voiceless segments can be restored by listeners under certain circumstances, our present findings also suggest the existence of a compensatory mechanism. Speakers did not use the same *Aa* level across all target words when they imitated the prominence level of a given stimulus. Rather, they adjusted the *Aa* level of the realized target word such that they used higher *Aa* levels for words with *F0* interruptions, hence compensating for their inherently lower prominence. As far as we know, such a compensatory mechanism is observed for the first time here, although it is known for a long time that microprosodic variation is compensated for in speech production and/or perception. For example, listeners compensate for intrinsic *F0* variation [22] or intrinsic duration variation [23]. Such compensatory mechanisms are typically only partial; and a comparison between the *Aa* levels in perception and production/imitation suggests that the same also applies to our findings.

Finally, given the general parallelism between heard and realized *Aa* levels, the clear distinction between short and long vowels, the observed "transfer/echo effect" (short/long vowels in the stimuli resulted in generally shorter/longer syllable productions), and the replication of known interactions between syllable structure and F0 peak timing, we can be confident that our speakers generated valid data and did a good job in reproducing the prosodic prominence patterns of acoustic stimuli on the same of different target words. Nevertheless, follow-up studies should aim at replicating the present findings with a larger number of speakers. and probably without the monotonous *F0* stimuli, as it is unclear how their very special character biased the prominence imitations in the *F0* peak stimulus conditions. Future work should also take up the observed "echo effects", which could point to phonetic accommodation or the way in which acoustic properties are mapped onto perceptual measures, which is another promising field for imitation tasks.

6. Acknowledgements

Special thanks go to all students at Beuth University Berlin and Kiel University participating in this experiment.

7. References

[1] Fry, D.B., “Experiments in the perception of stress”, Language and Speech 1, 126-152, 1958.

[2] Gay, T., “Physiological and acoustic correlates of perceived stress. Language and Speech 21, 347-353, 1978.

[3] Koreman, J., Van Dommelen, W., Sikveland, R., Andreeva, B., Barry, W.J., ”Cross-language differences in the production of phrasal prominence in Norwegian and German”, in M. Vainio, R. Aulanko, O. Aaltonen (eds.): Nordic Prosody X. Frankfurt: Peter Lang, 139-150, 2009.

[4] Mixdorff, H., Widera, C., “Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model”, Proc. Eurospeech 2001, Aalborg, Denmark, 403-406, 2001.

[5] Fujisaki, H., Hirose, K., “Analysis of voice fundamental frequency contours for declarative sentences of Japanese”, Journal of the Acoustical Society of Japan 5, 233-241, 1984.

[6] Heuft, B., “Eine prominenzbasierte Methode zur Prosodieanalyse und –synthese”, in W. Hess and W. Lenders (eds.): Computer Studies in Language and Speech, Vol. 2, Peter Lang, Frankfurt am Main, 1999.

[7] Niebuhr, O., “Interpretation of pitch patterns and its effects on accentual prominence in German”, Proc. 3rd Tone and Intonation in Europe Conference (TIE3), Lisbon, Portugal, 2008.

[8] Niebuhr, O., “F0-based rhythm effects on the perception of local syllable prominence”, Phonetica 66, 95-112, 2009.

[9] Kleber, F., Niebuhr, O., “Semantic-context effects on lexical stress and syllable prominence”, Proc. 5th Speech Prosody, Chicago, USA, 1-4, 2010.

[10] Barnes, J., Brugos, A., Veilleux, N., Shattuck-Hufnagel, S., “Voiceless Intervals and Perceptual Completion in F0 contours: Evidence from scaling perception in American English”, Proc. 16th ICPhS, Hong Kong, China, 108-111, 2011.

[11] Mixdorff, H., Niebuhr, O., “The Influence of F0 Contour Continuity on Prominence Perception”, Proc. Interspeech 2013, Lyon, France, 230-234, 2013

[12] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vreken, O., “The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes”, Proc. ICSLP, Philadelphia, USA, 1393-1396, 1996.

[13] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009.

[14] Boersma, P., “Praat, a system for doing phonetics by computer”, Glot International 5, 341-345, 2001.

[15] Niebuhr, O., Ambrazaitis, G.I., “Alignment of medial and late peaks in German spontaneous speech”, Proc. 3rd Speech Prosody, Dresden, Germany, 161-164, 2006.

[16] Draxler, Chr., Jänsch, K., “WikiSpeech - A Content Management System for Speech Databases”, Proc. Interspeech 2008, 1646-1649, Brisbane, 2008.

[17] http://www.wevosys.com/products/lingwaves/lingwaves.html

[18] Mixdorff H., “A novel approach to the fully automatic extraction of Fujisaki model parameters”, Proc. ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.

[19] Mixdorff, H., “FujiParaEditor”, http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html, 2009.

[20] Wichmann, A., House, J. Rietveld, T., “Discourse constraints on F0 peak timing in English”, in A. Botinis (ed.): Intonation. Dordrecht/Norwell: Kluwer Academic Publishers. 163-182, 2000.

[21] Niebuhr, O., “At the edge of intonation – The interplay of utterance-final F0 movements and voiceless fricative sounds”, Phonetica 69, 7-27, 2012.

[22] Niebuhr, O., “Intrinsic pitch in opening and closing diphthongs of German”, Proceedings of the 2nd international conference of speech prosody, Nara, Japan, 733-736, 2004.

[23] Gussenhoven, C., “Explaining two correlations between vowel quality and tone: the duration connection2”, Proceedings of the 2nd international conference of speech prosody, Nara, Japan, 179-182, 2004.

[24] Niebuhr, O., D’Imperio, M., Gili Fivela, B., Cangemi, F., “Are there "shapers" and "aligners"? Individual differences in signaling pitch accent category”, Proc. 17th ICPhS, Hong Kong, China, 120-123, 2011.

The Structure of Japanese Phrase in Accordance with Speaking Modes

Toshiyuki Sadanobu

Graduate School of Intercultural Studies, Kobe University

sadanobu@kobe-u.ac.jp

Abstract

While English is often spoken in an increment of clause (i.e. subject and predicate), Japanese of a smaller phrase called "bunsetsu" (e.g. noun phrase and case particle). Previous studies on Japanese language, however, have traditionally been focusing on clause structure, and little attention has been paid on the structure of "bunsetsu" (non-predicate one, especially). This paper describes the basic structure of non-predicate "bunsetsu" from grammatical point of view, and elucidates that the structure of non-predicate "bunsetsu" varies in accordance with four speaking modes ((i) Sentence mode A; (ii) Sentence mode B; (iii) "Bunsetsu" mode; and (iv) Character mode), which are identified on the criteria of compatibility among seven phenomena attested in Japanese speech. To be more concrete, this paper shows that it is only the mode (iii) that enables copula, "bunsetsu"-final particle ("Kantoujoshi" in Japanese), final leaping, and combination of breaking and prolongation in non-predicate "bunsetsu").

Index Terms: intonation unit, speaking mode, Japanese

1. Introduction

While English speech often employs the clause (subject + predicate) as a unit, ([1][2]), Japanese speech often employs the phrase ("bunsetsu") as a unit ([3][4][5]). However, Japanese language studies through now have concentrated on structural analysis of the sentence, with few exceptions ([5][6]) almost never looking at the structure of the phrase (particularly the non-predicate phrase). By focusing on the relation between phenomena, this paper groups speech modes (speech consciousness) into four main types ((i) sentence mode a, (ii) sentence mode b, (iii) phrase mode, (iv) character mode), and shows that the structure of a non-predicate phrase varies by speaking modes.

Section 2 will introduce the phenomena discussed in this paper. Section 3 will show how they can be grouped into four categories connected with four speaking modes ((i) Sentence mode A; (ii) Sentence mode B; (iii) Bunsetsu mode; and (iv) Character mode) based on the relation between these phenomena, and describe that the structure of non-predicate phrase can be complex only on Bunsetsu mode. Lastly, Section 4 will summarize this paper's findings.

2. Seven phenomena

This paper observes seven types of phenomena in total. A brief introduction to each is provided below

2.1. Disappearance of lexical accent (DA)

While traditionally it has been thought that in Japanese intonation does not break down the form of lexical accent ([7]), in actual everyday communication it is not uncommon for intonation alone to be reflected in tone, with lexical accent drowned out ([8]).

For example, in the database of conversations between a mother and her child commonly known as "Emi-chan Data," ([9]), the entire sentence of the joking threat "Emi no otete tabechau zoo" ("I'm going to eat your hand!") is spoken in a gently rising tone, ignoring lexical accent on all words and phrases (no. 14, side B, approx. 49: 32).

In this paper, we will call this phenomenon in which tone reflects only one intonation and lexical accent is not reflected "disappearance of lexical accent" (DA).

2.2. Preposed predicate phrase (PPP)

As seen in the example "Chan to kare ni itta yo" ("I already told him"), in a Japanese sentence the predicate phrase (in this example, "itta yo") is positioned at the very end. However, this order of occurrence of phrases is not strictly unchanging, and in some cases the predicate phrase may appear in a position more to the front of the sentence, as in "Itta yo, chan to kare ni." Furthermore, in some cases the predicate phrase may appear in positions other than the end of the sentence with modifiers, as in "Chan to itta yo, kare ni" or "Kare ni itta yo, chan to." In this paper, we will call this phenomenon, including all of these cases, "preposed predicate phrase" (PPP).

This paper will not pursue the issue of whether the predicate phrase actively is placed in the front in all cases referred to as PPP or whether some cases should properly be described as "rear placement" (late appearance) of other phrases. Instead, it will refer to every such case uniformly as PPP.

2.3. Occurrence of copula (OC)

In the example "Nihon no shuto wa Tokyo da" ("The capital of Japan is Tokyo"), the copula (in this example, "da") occurs at the end of the predicate phrase. Traditionally, the copula has been known to appear at the end of other phrases as well ([7]). As one specific example, in "Shikamo da" ("And that too") the copula (in this example, "da") occurs at the end of the adverbial phrase "shikamo," which is not a predicate phrase.

As used in this paper, "occurrence of copula" (hereafter OC) will refer only to cases in which a copula occurs at the end of a phrase other than a predicate phrase. It will not include cases when a copula occurs at the end of a predicate phrase.

Besides "da", there is another copula "desu", which often appears at the end of non-predicate phrase as well as "da". Moreover, it is possible to combine these two copulas through the past particle "ta" to construct a complex form "da-tta-desu."

Aside from "da" and "desu," other copula widely seen to occur at the end of non-predicate phrases include "ja," as well as "ssu," "zamasu," "degozaru" and "deojaru" etc. As

well as in predicate phrases, form of copula in non-predicate phrases varies in accordance with the speaker's types ([8][9]). For example, a speaker of "old man" type utters "ja" instead of "da", and "desu-ja" instead of "desu". And instead of "desu", athletic speaker utters "ssu", upper-class matron speaker utters "zamasu", samurai/ninja utters "degozaru", and Heian noble man utters "deojaru."

2.4. Occurrence of interjectory particle (OI)

Some final particles appearing at the end of a predicate phrase (i.e., those other than particles such as "zo," "ze," or "wa") traditionally are known also to appear at the end of phrases other than predicate phrases, as interjectory particles ([7]). While some advocate eliminating the distinction between final particles and interjectory particles ([10]), in this paper we will maintain the distinction between the two, referring to the phenomenon of an interjectory particle ("ne" in the following examples) appearing at the end of a phrase other than a predicate phrase ("shikamo") as in the example "Shikamo ne" as "occurrence of interjectory particle" (hereafter OI).

In addition to "ne," widely seen interjectory particles likely to occur in this way include "na," "yo," and "sa," as well as "no(o)" as spoken by elderly people and "nya(a)" as spoken by "cat-like" people.

In some cases both a copula and an interjectory particle may appear together at the end of a non-predicate phrase. In such a case, as seen in the example "shikamo da ne" the copula ("da" in this example) is positioned first, with the interjectory particle ("ne" in this example) coming later.

2.5. Final leaping of intonation (FLI)

The phenomenon in which, for example, the phrase "kare ga," generally pronounced with a High-Low-Low tone (hereafter HLL), is instead pronounced with a HLH or the phrase "sore ga," generally pronounced with a LHH tone is instead pronounced with a LH-Ultra High, pronouncing the final component (in these examples, "ga") with a higher intonation as if it is leaping from the phrase, including cases in which this is followed by a return to a lower intonation ("kare ga(a) [HLHL], "sore ga(a)" [LH-Ultra High-L]), has been known for some time and is referred to by various names. In this paper we will refer to it simply as "final leaping of intonation" (FLI), separating this into "returned final leaping of intonation" (returned FLI) and "unreturned final leaping of intonation" (unreturned FLI) as necessary.

While returned FLI has a negative image as childish, unintelligent way of speaking among young people, in fact this intonation is observed broadly in the speech of young and old, male and female, with the exception of some high-ranking characters. Unless repeated frequently and consecutively, this intonation does not actually give such a bad impression.

Although it is difficult for a predicate phrase or final phrase to appear in the case of a returned FLI, this is merely a general tendency. In strong expressions such as "Hara ga tatsuu!" ("It makes me angry!"), "Koitsuu!" ("You!"), "Mo, otosan-ttaraa!" ("Oh, you, father!"), "Itsumo so nan da karaa!" ("It's always that way!"), "Wakatta yoo!" ("I get it!"), or "Kaeshite yoo!" ("Give it back!"), returned final leaping intonation may occur with a predicate phrase or final phrase ([12]).

2.6. Combination of breaking and prolongation (CBP)

Japanese permits various types of halting of speech. For example when saying the title "Dokurokamen" hesitatingly in a form such as "Saikin terebi de hayatteru anime arimasu yo ne. Eeto, _____ deshita-kke" ("There's a popular animated series on TV lately. What's it called, _____?"), it would be natural to say "Do, Dokurokamen" (breaking off and then repeating from the start), "Do, kurokamen" (breaking off and then continuing), "DooDokurokamen" (extending the first sound and then repeating from the start), or "Dookuro kamen" (extending and then continuing). At the same time, a certain connection can be seen between halting and feeling (i.e., hesitation or surprise), as in the way the only halting used to express surprise is "Do-, Dokuro Kamen" (breaking off and then repeating from the start) ([6]).

As used here, "combination of breaking and prolongation" (hereafter CBP) refers to a complex kind of halting in which, as in the examples "koozookaikaku, uuno" ("structural reforms" in genitive case) or "koozookaikakuo, oo," ("structural reform" in accusative case) after once halting pronunciation at the end of a word ("koozookaikaku") or phrase ("koozookaikakuo") ("koozookaikaku,""kozokaikakuo,") the final vowel sound ("u" or "o") is extended. This type of halting is used in official situations by adult characters who have authority (while taking care to be noncommittal). It is not used by children speakers.

2.7. High pitch accent on first mora (HF)

When, for example, reading the alphabet one letter at a time as it is pronounced in Japanese, every letter is pronounced with a high pitch accent on first mora, from "Ee" ("A") (HL) through "daburuyuu" ("W") (HLLL), "ekkusu" ("X") (HLLL), "wai" ("Y") (HL), and "zetto" ("Z") (HLL). Similarly, when reading the Kanji characters in the name "Koobe Daigaku" ("Kobe University") one at a time as "koo," "be," "dai," "gaku," again each character is pronounced with a high pitch accent on first mora: "koo" (HL), "be" (H), "dai" (HL), and "gaku" (HL). This phenomenon will be referred to as "high pitch accent on first mora" (hereafter HF). In general, when adding honorifics such as "-san," "-chan," and "-kun" to people's names the pitch of the immediately preceding mora is continued unchanged, as in "Tanaka-san," "Tanaka-chan," or "Tanaka-kun" (LHH-HH) or "Shimizu-san," "Shimizu-chan," or "Shimizu-kun" (HLL-LL). However, in all cases when there is a high pitch accent on the first mora, as in "Ka-san," " Gi-chan," "Go-kun," "Shu-san," "Jo-chan," "So-kun," "Pe-san," or "Ra-chan," "-san," "-chan," and "-kun" are pronounced with a low pitch (HLL). Furthermore, when pronouncing a foreign name such as Xiaoming Chang in a way patterned on the tone of its original language (Chinese), such as "Chen Shaomin"(LHHLLH), "-san" appended at the end has a low pitch as well. These cases too are due to the phenomenon of reading with a high pitch accent on the first mora. When saying, for example, "'Sada' to iu ji wa ko kakimasu" ("The character 'sada' is written this way") as well, the accent is high on the first mora, and this too is due to the phenomenon of reading with a high pitch accent on the first mora.

3. Relation between these phenomena

Observation of whether the seven types of phenomena introduced above are likely to co-occur shows that while three types of phenomena are unlikely to co-occur with others, four types may co-occur with and affect on each other. That is, three types of phenomena may be considered each to result from peculiar speech modes of the speaker, while four may be considered to result from a common speech mode.

3.1. Peculiarity of DA

Among these seven types of phenomena, DA does not co-occur with the other phenomena. No copula or interjectory particle is added, for example by saying in a gently rising tone without lexical accent something such as "??Emi no o-tete da ne tabechau zoo" ("I'm going to eat your hand!"). (Hereinafter, double question marks "??" at the start of the sentence indicate that the sentence is an unnatural expression.)

3.2. Peculiarity of PPP

The PPP also does not co-occur with the other phenomena. For example, the sentence "Itta yo, chan to kare ni" is not spoken in a gently rising tone without lexical accent or spoken with a copula or interjectory particle added ("??Itta yo, chan to da ne, kare ni saa.")

3.3. Peculiarity of HF

HF does not co-occur with other phoneme phenomena (i.e., DA or FLI). The possibility of its co-occurrence with non-phoneme phenomena (i.e., PPP, OC, OI, or CBP) relates to the details of what is described as "speakership" [13] in this paper.

If the speaker is an "animator" merely reading something given to him or her, he or she may pronounce each character individually with a high pitch accent on the first mora—that is, HF may co-occur—in a sentence with a PPP ("Itta yo, chan to kare ni"), a phrase in which a copula or interjectory particle occurs ("Shikamo da ne"), or a phrase in which CBP occurs ("kozo kaikaku, u-wo"). However, if the speaker is the "author" choosing his or her own words, then co-occurrence with these phenomena is not possible.

3.4. Interrelation between other phenomena

Unlike the above three types of phenomena, the remaining four types may co-occur and affect each other in complex ways. I shall begin to show this by describing the relationship between OC and OI.

OC tends to make OI easier. To be more concrete, "zo" and "ze" tend to be natural at the end of non-predicate phrase only with copula. For example, "shikamo da zo" and "shikamo da ze" are natural unlike "??shikamo zo" and "??shikamo ze". (We might be able to say that "shikamo da zo" and "shikamo da ze" are sentences to some degree. However this does not mean that they are not phrases at all. One reason for this is that "ze" and "zo" in "shikamo da zo" and "shikamo da ze" bear a prosodic restriction which are not found for them in the end of predicate phrase. That is to say, "ze" and "zo" in "shikamo da zo" and "shikamo da ze" are necessarily pronounced in rising intonation but never in "FLI".) The variety of interjectory particles at the end of non-predicate phrase still increases when we add the past particle "ta" to the copula, as in "tookyoo kara da-tta kashira" and "(tashika) tookyoo kara da-tta wa" (cf. "?? tookyoo kara kashira" "?? tookyoo kara wa"). And "wa" at the end of non-predicate phrase is uttered in rising intonation. Another example of increase of naturalness by additing "ta" is some unnatural phrases with the particle "ka" at the end of non-predicate phrase: "?tookyoo kara ka" (cf. "doko kara ka") vs. "tookyoo kara da-tta ka"

OI can make non-predicate phrases more natural, as in "soreo desune" (cf. "??soreo desu"). More striking is the case with the complex copula "da-ttadesu", as in "kare to da-tta desu kane" (cf. "??kare to da-tta desu"). However this effect of interjectory particles can be seen only when OC realizes.

We can see still more relationship between the four phenomena. This will be discussed in practical terms below, looking chiefly at the phenomenon of FLI. First we will look at the relationship between FLI, OC, and OI.

When no copula or interjectory particle appears at the end of a phrase, both returned FLI ("kare ga" ("he . . ."), "dakara" ("because . . ."); HLH) and unreturned FLI ("kare ga(a)," "dakara(a); HLHL) are possible.

When a copula but no interjectory particle appears at the end of a phrase, while returned FLI is not possible unreturned FLI may occur for a particle ("kare ga da"; HLHL). After all, unlike the way the copula of a predicate phrase may have a high tone (the "da" in "Tokyo da" is High, while the "desu" in "Tokyo desu" is High-Low), the copula of a non-predicate phrase always is low and is not subject to FLI. This is similar to the way a quotation marker "to" spoken to oneself as a separator when composing a sentence, as in "'Bunsetsu matsubi ni' to, 'kopyura ga' to, 'arawareru ga' to, . . ." ("'at the end of a phrase,' and . . . 'a copula,' and . . . 'appears, but, . . . ").

When no copula but only an interjectory particle appears at the end of a phrase, both unreturned FLI ("da kedo ne" ["that's true, but"; HLLH) and returned FLI ("da kedo ne(e)"; HLLHL) are possible.

When both a copula and an interjectory particle appear at the end of a phrase, FLI may occur for the part immediately preceding the copula or for the interjectory particle immediately following it. That is, it may occur up to two times in a single phrase ("kare ga da ne(e)"; HLHLHL). While only unreturned FLI may occur for the part immediately preceding the copula ("??kare ga(a) da ne(e)"; HLH(L)LHL), both returned FLI and unreturned FLI are possible for an interjectory particle.

Next, let's look at the relationship between combination of brealing and prolongation and continuation and the other phenomena. Since as described above in Section 2.6 CBP and continuation is used in official situations by adult characters who have authority, when considering the possibility of co-occurrence with this halting we must look only at copula such as "desu" and "da" and interjectory particles such as "ne" and "na." However, it is hard to judge whether "ee"and "aa"in "sore o da ne, ee"and "sore o desu na, aa" are part of this halting or are merely the fillers "ee"and "aa," or if there may be instances of both cases. That is, the possibility of co-occurrence of CBP and OC or OI is unclear.

However, it is clear that co-occurrence of CBP and FLI is possible, and what's more these two phenomena affect each other as shown below.

When CBP occurs at the end of a noun (“koozookaikaku, uuno”), both returned and unreturned FLI are possible for the end of the phrase (“koozookaikaku, uuno”; LHHHHLLL, L-H; “koozookaikaku, uuno(o)”; LHHHHLLL, L-HL). The extended part in CBP is simply a continuation of the immediately preceding sound, so that if the immediately preceding sound is high (“kaikaku”; LHHH), the extended part also will be high (“kaikaku, uuno(o)”; LHHH, H-Ultra High(-L)).

When CBP occurs at the end of a phrase (“koozookaikaku no, oo”), while again both returned and unreturned FLI are possible (“koozookaikaku no(o), oo; LHHHHLLLH(L), L-), the extended part is not subject to the FLI, so that no matter how high the tone of the immediately preceding sound the extended part always will be low.

3.5. Four types of speech consciousness

We have grouped the seven types of phenomena into four categories based on the possibility of co-occurrence between phenomena. These four categories can be understood as differences in speaking mode.

First of all, HF can be considered a phenomenon that occurs when the speaker is pronouncing characters as a list of characters without being conscious of their meaning.

Also, DA can be considered a phenomenon that occurs at the sentence level when the speaker is conscious of the entire sentence, since a single intonation is expressed for the entire sentence. The speaker is conscious of a strong feeling connected to this intonation (i.e., a joking threat if the intonation is gently rising).

PPP also can be considered a phenomenon that occurs at the sentence level when the speaker is conscious of the entire sentence, since it manipulates the order of phrases within the sentence. However, in the case of preposed predicate the speaker is not conscious of a particularly strong feeling.

The four phenomena of OC, OI, FLI, and CBP occur when the speaker is conscious of the smaller units of individual phrases rather than the entire sentence. For this reason all the more, they may co-occur and affect each other.

4. Conclusion

By looking at seven phenomena in spoken Japanese, focusing on the relations between phenomena, this paper grouped the speaker’s speaking mode into four categories. These four categories are (i) when the speaker is conscious of the sentence as a whole and is conscious of a strong feeling connected to a specific intonation, (ii) when the speaker is conscious of the sentence as a whole but is not conscious of such a strong feeling, (iii) when the speaker’s consciousness is at the phrase level, and (iv) when the speaker is conscious only of the characters, without being conscious of their meanings.

While past studies of grammar and prosody have concentrated their conditions on category (ii), in fact there are various other types of speech consciousness, and the form of speech varies with these types. In particular, if we admit that Japanese speech often takes place in units of phrases, then perhaps we should turn our observation to category (iii) as well.

This paper has made it clear that it is only in category (iii) that a copula, interjectory particle, FLI, or CBP may occur within a phrase, and in doing so it has showed that in this sense the structure of a phrase varies with the speech consciousness of the speaker. We also have observed the complexity of the relationship between the four phenomena in the case of category (iii). In order to unravel this complexity, there probably is a need to further deepen our understanding of the structure of the phrase in the future.

5. Acknowledgements

I thank many members of Onsei Bumpo Kenkyukai, especially Nick Campbell, and late Miyoko Sugito for extensive advice and for helping in many other ways. This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 23320087, and (C), 24500256.

6. References

[1] Chafe, Wallace, The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production. Ablex, 1980.

[2] Chafe, Wallace, “Cognitive constraints on information flow”, in Russell S. Tomlin [Ed], Coherence and Grounding in Discourse: Outcome of a Symposium, Eugine, Oregon, June 1984, John Benjamins, 21-51, 1987.

[3] Clancy, Patricia, “Written and spoken style of Japanese style”, in Deborah Tannen [Ed], Spoken and Written Language: Exploring Orality and Literacy, Ablex, 55-76, 1982.

[4] Maynard, Senko K., Japanese Conversation, Ablex, 1989.

[5] Iwasaki, Shoichi, “The structure of the intonation unit in Japanese”, in Soonja Choi [Ed], Japanese/Korean Linguistics, Vol. 3, 39-51, 1993.

[6] Sadanobu, Toshiyuki, Sasayaku Koibito, Rikimu Reporter:Kuchi no naka no Bunka, Iwanami, 2005.

[7] Amanuma, Yasushi, Kazuo Ootsubo, and Osamu Mizutani, Nihongo Onseigaku. Kurosio, 1978.

[8] Sadanobu, Toshiyuki, “The competitive relationship between Japanese accent and intonation”, Japanese Speech Communication, No. 1, 1-27, http://www.hituzi.co.jp/epublish/academic_journal/nhng_onsei/index.html, 2013.

[9] Sugito, Miyoko, “Nyuuyoujito hahaoyatono taiwa onsei deetabeesu-Emichan: sono shoukaito riyou nitsuite”, Onsei Kenkyu, Vol. 9, No. 3, 52-57, 2005.

[10] Tanaka, Akio, “Shuujoshi to kantoujoshi”, in Kazuhiko Suzuki and Ooki Hayashi [Eds], Joshi, Meijishoin, 209-247, 1973.

[11] Sadanobu, Toshiyuki, Nihongo Shakai Nozoki Kyarakuri: Kaotsuki, Karadatsuki, Kotobatsuki. Sanseido, 2011.

[12] Sadanobu, Toshiyui, and Miliang Luo, “Bumpou, paragengojouhou, character ni motozuku nihongo meishisei bunsetsu no tougoutekina kijutsu”, Journal CAJLE, Vol. 12, 77-95, 2011.

[13] Fujiwara, Yoichi, Bumpougaku. Musashinoshoin, 1994.

[14] Sadanobu, Toshiyuki, “Bunsetsuto bunno aida”, Onsei Bumpou Kenkyuukai [Ed], Bumpou to Onsei, Vol. 5, 107-133, Kurosio, 2006.

[15] Goffman, Erving, Forms of Talk. University of Pennsylvania Press, 1981.

8 Wednesday 1

Stability in perceiving non-native segmental length contrasts

Yuki Asano

Department of Linguistics, University of Konstanz, Konstanz, Germany

yuki.asano@uni-konstanz.de

Abstract

Previous studies have demonstrated that listeners show high sensitivity in discriminating non-native segmental length contrasts thanks to auditory memory [1, 2]. We tested the limits of discriminating Japanese consonantal length contrasts with three groups of listeners (German learners of Japanese, German non-learners and Japanese natives) under increasing task demands. We increased memory load through a longer inter-stimulus interval (= ISI) (2500ms vs. 300ms) and added psycho-phonetic complexity (trials with task-irrelevant pitch falls that occurred simultaneously with the consonant vs. with monotonous pitch).

Results showed very good discrimination in all groups when the task demands were the lowest. With increasing task demands, only the non-natives' discrimination abilities decreased: non-learners were strongly affected by both ISI and pitch, while learners only by pitch. The psycho-phonetic complexity of the stimuli had a stronger impact on performance than the increased memory load.

Our findings suggest that L2 learners can establish novel phonological representations, but the ability to use them can be applied only under favourable listening conditions with no distracting acoustic information. The non-native listeners' reduced sensitivity under increasing task demands appears to be the reason why even advanced learners still face difficulties in natural learning situations.

Index Terms: task demands, discrimination, non-native length contrast, Japanese, German, geminate

1. Introduction

The discrimination of non-native length contrasts is expected to be difficult for L2 listeners when such contrasts do not exist in an L2 lexicon (theoretical supports for segmental contrasts see e.g. [3, 4]). Indeed, better performances have been oft found in L1 listeners rather than in L2 listeners [5, 1, 6]. At the same time however, despite this prediction, even non-native listeners have occasionally been reported to discriminate non-native prosodic contrasts fairly better than expected, given the lacking L2 phonological category in a listeners' L1. Hayes and Masuda [1] and Wilson *et. al* [2] report that even non-learners who have no consonantal length contrasts in their L1 could discriminate them without any exposure to the L2, by simply relying on auditory memory [7] and absolute durations [8, 2]. We postulate that these findings do not contradict each other, but do relate to task demands.

Under challenging situations, speech perception becomes more demanding. This is because the listener's cognitive load, operationalized as attention control (efficient attention shift among foregrounding and backgrounding of task-relevant and -irrelevant information [9]) and memory load becomes higher in such situations (e.g. [10, 11]). That appears to be all the more true when it comes to L2 perception. Previous studies show that the difficulty of listening to speech in noise or with greater talker variability is more exacerbated in L2 perception when compared to L1 perception [12, 13, 14, 15]. Given that the cognitive resources in L2 speech processing are fewer due to reduced L2 proficiency in comparison to native speakers [12], L2 perception is expected to be more "vulnerable" under such demanding listening conditions.

Werker and Logan [16] for instance show a performance decrease in the L2 listeners' discrimination of non-native segmental contrasts once an ISI becomes longer. ISI is known to increase task demands affecting the memory load (the longer the higher) and involving different levels of speech processing (the longer the more higher phonological perception) [17, 18, 19, 7, 20, 16]. Their findings suggest that L2 listeners had difficulties in discriminating non-native segmental contrasts once memory load increased and the processing tapped into the phonological one. We will investigate non-native listeners' limitations in discriminating non-native segmental *length* contrasts (and not segmental contrasts) under increasing task demands by increasing ISI [16] and by adding another dimension of task demands, psycho-phonetic complexity relating to attention control.

A discrimination task is a suitable method to investigate said issue, because it appears to require low task demands under a certain condition (see the aforementioned fairly good non-native performance) and it is therefore easy to enhance task demands in other conditions. In the current study, we conducted a speeded same-different-task to discriminate consonantal length contrasts testing Japanese natives, German learners and non-learners. Consonantal length is lexically contrastive in Japanese, but not in German. We expect therefore that the discrimination of consonantal length contrasts will be exacerbated to a greater extent only for German learners and even more for German non-learners under increasing task demands. Increasing task demands in our study are defined along two dimensions: attention control through psycho-phonetic complexity of the stimuli (e.g. [9, 21]) and memory load (e.g. [22]), the capacity to hold memory for a limited period of time.

psycho-phonetic complexity was increased by adding a task-irrelevant pitch movement. We built two stimuli conditions, one with pitch falls that occurred simultaneously with the consonant and the other with flat (namely monotonous) pitch. Attention control in the former condition is expected to be higher, because ignoring the task-irrelevant pitch movement is required in this condition. In case listeners cannot ignore it, they need to process both pitch and length simultaneously in the consonantal length contrast. In the flat pitch condition, listeners have one prosodic cue less to process (only duration). In case of an unsuccessful attention control, the trials with a falling pitch are expected to become psycho-phonetically more complex and more difficult for the discrimination of consonantal length contrasts than those with a flat pitch.

Previous studies consistently report that the complexity of speech perception in a bad speech quality or in unfavourable circumstances (e.g. in noise) impairs successful speech perception [10, 11, 23] and that is more the case for L2 perception [12, 13, 14]. It may be assumed that L2 listeners' are less successful in ignoring or shutting down the task-irrelevant information, e.g. noise in background. In our study we therefore hypothesize that the psycho-phonetic complexity of the stimuli constitutes is an impediment to a greater extent for L2 listeners than for L1 listeners, because the former are less likely to ignore the task-irrelevant pitch movement. Moreover, this additional falling pitch movement mirrors the phonetic form of a Japanese lexical falling pitch accent. The study will therefore contribute to understanding the L2 perception of a language with lexical tone or pitch accent.

The other dimension, memory load, was manipulated by increasing the duration of ISI (in one condition 300ms and in another condition 2500ms). In order to eliminate the risk of backward masking, ISI should be longer than 250ms [17, 24, 25]. Around 250ms after the offset of a sound, information is recognized at the sensory level, but is not yet identified or categorized [26]. The discrimination ability increases rapidly between 100 and 500ms and falls gradually as the ISI increases further [17, 7, 20]. The decrease after 500ms may be interpreted as the effect of gradually decaying auditory information in short-term memory. In order to test the acoustic comparison of successive stimuli without risking a backwards masking effect, we decided to use 300ms in one condition (slightly longer than 250ms) to ensure the acoustic trace was available. Then, it is claimed that this uncategorized acoustic information is maintained for a while (approximately 2000ms) in the precategorical acoustic storage [26] or in the short-term phonological storage as an auditory image [27, 28]. To keep our experiment to a duration that would not cause the participants to diminish their concentration or motivation, but at the same time to make sure that the the processing taps into the categorical level after 2000ms [27, 28, 26], we decided to use 2500ms as the long ISI.

At the segmental level, numerous studies have investigated the effect of the durations of ISIs on the discrimination of non-native *segmental* contrasts. With a short ISI with which researchers hypothesize that the lexical knowledge or phonological category is not consulted, no difference between L1 and L2 listeners was found. When the ISI became long, however, the L1 listeners' performance surpassed the L2 listeners' one. No study has been carried out so far that explores the discrimination of non-native *segmental length* contrasts by varying ISIs. We assume that segmental length contrasts are perceived relatively in a larger unit in relation to the adjacent words than segmental contrasts. Therefore, it might be more difficult to discriminate two stimuli with a segmental length contrast by directly relying only on auditory memory.

2. Experiment

2.1. Methods

2.1.1. Participants

Twenty-four native Japanese participants (henceforth= JNs, 10 male, 20-31 years), 24 native German who were not learning Japanese (= non-learners, henceforth= GNs, 8 male, 19-30 years) and 48 German learners of Japanese (henceforth= GLs, 30 male, 20-34 years) took part for a small fee. They were unaware of the purpose of the experiment. None of the learners had prior training in Japanese phonology.

2.1.2. Materials

We used disyllabic triplets that differed segmentally only in the length of the first vowel or in the length of the second consonant (e.g., [pɯnɯ],[pɯːnɯ], [pɯnːɯ]). Twenty-one triplets were evaluated in a pretest with Japanese and German native listeners (different from those of the main experiment) to select only stimuli that did not activate a word via phonological analogy. Participants were presented with one stimulus at a time and were required to write down the first word coming to their mind. We analysed the responses of 24 Japanese natives and 24 German natives separately. From there, six non-word triplets with the lowest association strength in both groups were selected (word association rate between 29.8 % and 45.3%, mean = 34.5%, while the one of all 21 triplets ranged between 29.8% and 100.0%, mean = 52.3%). The selected stimuli differed in manner of articulation and voicing of the medial consonant (= phon), *punu, gunu, gupu, gubu, zusu, sufu*. The materials were recorded by a female speaker of Japanese in two pitch conditions; high flat pitch and falling pitch (with a pitch fall during the medial consonant pitch tracks, see Figure 1). Each stimulus was recorded six times in order to have different tokens of the same type (N = 216). To ensure the same pitch between the two stimuli presented together (= A and X), the pitch was manipulated by using a method which is based on the representation of F_0 contours with B-splines [29] and on a smooth warping of the time axis allowing us to move selected time boundaries to desired positions (see [30]). More specifically 6 tokens of each triplets of a given phon (thus N = 18 for each phon) realized in the same pitch pattern were aligned on the average pitch across tokens (in the flat pitch: average = 1.3 semitones, range = 1.0 – 1.6 semitones; in the falling pitch: average = 13.0 semitones, range = 10.5 –16.4 semitones). Finally, a female native speaker of Japanese and a male native speaker of German selected the most naturally sounding tokens as experimental items (for each phon, N = 3 for the stimuli with a long vowel or a long consonant, N = 4 for the stimuli with short vowels and consonants). There was no disagreement on the decisions.

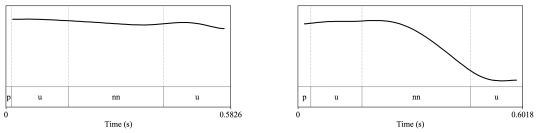


Figure 1: *Pitch track of geminate stimuli in the flat and falling pitch condition. F_0 range is shown between 100 and 350 Hz.*

To verify the durational differences of the selected stimuli, mean durations from the three different tokens of the respective long and short consonants and vowels were analyzed. A linear mixed effects regression model with the critical vowel or consonant *duration* as dependent measure and *pitch* (flat vs. falling), *segmental condition* (short vs. long vowel or consonant) as fixed factors and *phon* as a random factor including random slopes for the fixed factors [31, 32] showed a significant interaction for vowel contrasts ($p < 0.01$), but not for consonantal contrasts ($p > 0.1$). Then we analyzed the durations in the flat and falling pitch conditions separately for vocalic contrasts. Results of paired t-tests showed that, on average, long vowels in the flat pitch condition were 3.3 times longer than short vowels ($t(5) = 20.0$, $p < 0.001$) and those in the falling pitch condition were 3.0 times longer ($t(5) = 28.1$, $p < 0.001$). For consonant contrasts, we analyzed the durations in the flat

and falling pitch conditions together. Results of paired t-tests showed that geminates were on average 3.2 times longer than singleton consonants in the flat pitch condition ($t(10) = 25$, $p < 0.001$). These duration measurements ensured that the acoustic criteria for the length distinction in vowels and consonants were met (ratio for Japanese vowels was approximately 3.2:1, [33] and for consonants 3.2:1 [34, 35]).

To compare the spectral quality for long and short vowels /u:/ and /u/, the first and second formants at the midpoint of the vowel were automatically extracted. A linear mixed effects regression model with *formant* as dependent measure and *pitch* (flat vs. falling), *formant condition* (F1 vs. F2) as fixed factors and *phon* as a random factor including random slopes for the fixed factors [31, 32] showed that an interaction approached significance ($p = 0.07$). Results of paired t-tests separately in the two pitch conditions showed no difference in F1 nor in F2 (p-values were overall > 0.15). On average, F1 was 501.3 Hz for long vowels and 484.8 Hz for short vowels in the flat pitch condition and 467.9 Hz for long vowels and 429.9 Hz for short vowels in the falling pitch condition. F2 was 1499.0 Hz for long vowels and 1530.8 Hz for short vowels in the flat pitch condition and 1510.0 Hz for long vowels and 1495.9 Hz for short vowels in the falling pitch condition.

2.1.3. Procedures

A speeded AX-task was used to test the subjects' sensitivity for consonantal length contrasts. One base list was assembled by presenting all possible pairings of the stimuli (N = 84). The short ISI was 300ms lSI (session 1) and the long one 2500ms lSI (session 2). In both sessions, the intertrial-interval was 1000ms. Each trial began with a beep of 44100 Hz (500ms). No feedback was provided during the experiment. Each session began with 10 training trials using the phones that were not used as the experimental ones (*guna* and *puna*) followed by a pause before the experimental session started. The test lasted approximately 20 minutes. The order of presentation was automatically randomized using *Presentation* (Neurobehavioral Systems).

2.2. Results

2.2.1. Sensitivity to contrast: d' scores analyses

Participants' sensitivity to the contrasts was calculated using d' [36]. We calculated d' scores for each participant for each of the consonantal and vocalic length contrasts, flat and falling pitch as well as short and long ISI. We normalized them by subtracting those for vocalic length contrasts (as baseline) from those for consonantal length contrasts (note that there were no differences in the d' scores for vocalic length contrasts across *language groups*, overall $p > 0.2$). The value of 0 in the plot means that the d' for consonantal and vocalic length contrasts were almost the same. A linear mixed effects regression model with *d' scores* as dependent measure and *language groups*, *pitch* (flat vs. falling), *ISI* (short vs. long) as fixed factors and *participants* as a random factor including random slopes for the fixed factors [31, 32] showed a significant three-way interaction ($p < 0.003$). To investigate the nature of this interaction, the data were split by *pitch*. In the flat pitch condition, an interaction was found between *language group* and *ISI* (GNs' d' scores decreased in the long ISI condition, but the d' scores of the other two groups did not, $p < 0.03$). JNs' and GLs' d' scores did not differ from each other. In the falling pitch condition, a main effect of *language groups* was found (the JNs' d' scores higher than the GLs' ones, $p < 0.01$, the GLs' ones higher than the GNs' ones, $p< 0.01$); see Figure 2. In order to compare the d' scores between the two plots, we analyzed the d' scores for both pitch conditions in each *language groups*. JNs' d' scores were not affected by *pitch* or *ISI* (overall $p > 0.7$). In the GLs' data, the effect of *pitch* approached significance (flat $>$ falling, $p = 0.09$). The GNs' data showed a significant interaction between *pitch* and *ISI* (d' scores only in the flat pitch condition decreased in the long ISI condition, $p < 0.03$).

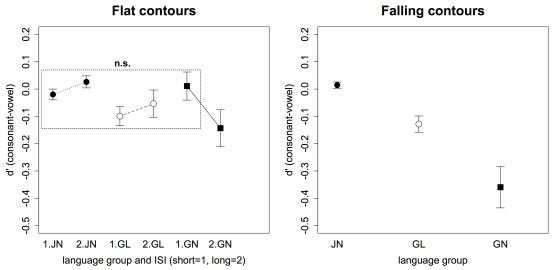


Figure 2: *Results of the normalized d' scores*

2.2.2. Processing difficulty: reaction time analyses

The reaction time (RT) analyses were performed to investigate task difficulty for the discrimination of *length conditions* (vowel vs. consonant) [37, 38]. We only analysed RTs in trials with different pairs. To account for participant-specific RT-differences, we normalized the raw RT data in the following way: We discarded RTs longer than 2000ms and aggregated the data for each *participant*, *pitch*, *length condition* and *ISI*. Then, the averaged RTs for vocalic length contrasts were subtracted from those for consonantal length contrasts.

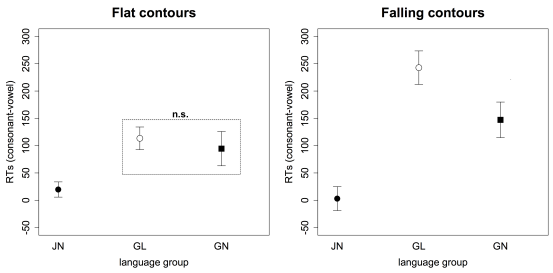


Figure 3: *Results of the normalized RTs*

In the following, we will describe the results of a linear mixed effects regression model with normalized *RTs* as dependent variable, *language groups*, *pitch* and *ISI* as fixed factors and *participants* as a random factor including random slopes for the fixed factors. Results showed a main effect of *language groups* (JNs faster than other two groups, $p < 0.001$) and an interaction between them (only GLs slower in the falling condition than JNs and GNs were, $p < 0.03$); see Figure 3. In order to compare the data between the two plots in Figure 3, we further analysed the RTs in both two pitch conditions in each *language groups* separately (*pitch* as a fixed factor and *participants* as a random factor including random slopes for the fixed factor). The analysis revealed that the JNs' and GNs' RTs in the two pitch conditions did not differ ($p > 0.9$, $p > 0.2$ respectably), while those of GLs did ($p < 0.001$).

3. Discussion

To investigate the (in)stability of the discrimination of non-native prosodic contrasts, we conducted discrimination tasks with various conditions that may impair the processing.

We found a consistent effect of pitch on the discrimination of consonantal length contrasts in the d' score analyses: the discrimination of consonantal length contrasts presented in falling pitch was more difficult for learners than in flat pitch and this especially for non-learners. The findings consistently confirm that the psycho-phonetic complexity affected only the non-natives' discrimination ability, indicating the non-natives' decreasing attention control along the increasing psycho-phonetic complexity. Additionally, the difference between the learners' and the non-learners' d' scores suggests a positive L2 learning effect in establishing higher attention control and in obtaining a more stable L2 processing of non-native prosodic contrasts.

The RT analyses also showed longer RTs in the learners' group when the pitch was falling. Contrary to the results of the d' scores, the non-learners' RTs did not change in both pitch conditions, suggesting that the psycho-phonetic complexity did not impact on a *perceivable* task difficulty for non-learners just as for native listeners. In this way, however, the learners' performance regarding the RTs became worse than the non-learners' one in the falling pitch condition. Such a similar tendency was also found in Altmann *et al.* [5] that investigated the non-native (= German) discrimination ability of Italian consonantal length contrasts. They found higher d' scores for learners than non-learners, but no significant difference between the learners' and non-learners' RTs. We postulate that learners and non-learners applied different strategies for the task. When the stimuli became psycho-phonetically more complex, the direct auditory comparison of the stimuli became more difficult. Therefore listeners started to rely on phonological knowledge. Such learners' indecisiveness demonstrates that the learners already had a degree of access to a newly forming category and thus needed to decide between two possible mapping representations. Non-learners on the contrary, required a shorter time for the decision without the "selection" processing. Another possible interpretation would be the following: this difference between the learners and non-learners was found to be independent from the two ISI conditions, suggesting a direct co-activation or mapping of phonological representations from acoustic speech signals [39]. Here again, since learners established L2 phonological representations, they needed more time to select one between the L1 and the L2 phonological representations, but non-learners did not. We assume that this difference was found only in the increasing psycho-phonetic complexity, because this temporal effect of the selection was relatively small. In future experiments further attention needs to be paid to the underlying mechanisms for non-learners and learners in future experiments.

As for the other dimension of increasing task demands, memory load, native-like good performance by learners and even non-learners was observed only in the condition with the lowest task demands (short ISI and flat pitch). In the long ISI condition, however, the non-learners' d' scores decreased. Such decrease is the evidence that they had difficulties in discriminating non-native consonantal length contrasts, when the memory load increased and once the processing tapped into the phonological one. Learners were not affected by the increasing memory load just as Japanese natives were not. This effect of memory load was found in a so to say "greenhouse condition" for the consonantal length contrasts, when it was not disturbed with another prosodic cue, pitch, but the processing only dealt with length contrast without any effects of pitch. Once pitch came into play, the effect vanished. Recall that the d' scores in the falling pitch condition were generally lower than those in the flat pitch condition in both the learners' and the non-learners' group, while those of native listeners did not change between the two pitch conditions. Therefore, the reason for not finding the temporal decrease in the falling pitch condition may be explained with a floor effect.

Taken together, the effect of psycho-phonetic complexity was stronger than the one of memory load. Both learners and non-learners were distracted by the stimuli with greater psycho-phonetic complexity, because their attention control became lower. As learners become more familiar with the situation, attention demands were eased and automatic processes developed [21]. Therefore, our finding suggests that learners, who even established the phonological representations of non-native consonantal length contrasts still were not successful in automatising the L2 speech processing fully.

The task demands controlled in our study may be translated into various distracting factors in our natural speech perception. Therefore, the performance decreases found in our study suggest the instability of the L2 perception under various task demands in our daily life. Moreover, the performance decrease due to the psycho-phonetic complexity suggests that the Japanese lexical pitch movement (e.g. a falling lexical pitch accent or an initial low [40]) makes the perception of non-native consonantal length contrasts more difficult.

4. Conclusions

We examined the discrimination of non-native consonantal length contrasts under increasing task demands. Even non-learners without any exposure to the L2 could discriminate them in the lowest task demands, simply by comparing two stimuli at the auditory level. However, such a reliance on the auditory comparison could not last long. Once the ISI became longer and memory load became higher, so that the auditory information began to be processed phonologically, their discrimination ability decreased and differed from the one by learners or native listeners. The learners' performance did not differ from the native listeners' one. This result suggests that the exposure to the L2 helped them to establish the phonological representations of non-native consonantal length contrasts. However, even learners who were not distracted by the increasing memory load could not overcome the performance decrease due to the increasing psycho-phonetic complexity of the stimuli. It was difficult for both learners and non-learners to ignore the task-irrelevant pitch and to focus on their attention only to the task-relevant information. The finding indicates the difficulty to automatise the L2 processing even after establishing or being exposed to the L2 categories. To summarize, the non-natives' performance was native-like good only under favourable listening conditions with no distracting acoustic information and with lower memory load.

In our natural listening situations, there are numerous distracting factors that might impair L2 perception. The overall decreases indicate why L2 perception remains difficult in the daily-life situations, despite the still successful exercise in L2 class room situations.

5. Acknowledgements

The study was funded thanks to a grant from the Young Scholar Funds at the University of Konstanz.

6. References

[1] R. Hayes-Harb and K. Masuda, "Development of the ability to lexically encode novel l2 phonemic contrasts," *Second Language Research*, vol. 24, no. 1, pp. 5–33, 2008.

[2] A. Wilson, H. Kato, and K. Tajima, "Native and non-native perception of phonemic length contrasts in japanese: Effects of speaking rate and presentation context," *The Journal of the Acoustical Society of America*, vol. 1, no. 117, p. 2425, 2005.

[3] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of nonnative consonant contrasts varying in perceptual assimilation to the listeners native phonological system," *Journal of Acoustical Society of America*, vol. 109, pp. 775–794, 2001.

[4] J. E. Flege, M. J. Munro, and I. Mackay, "Effects of age of second-language learning on the production of english consonants," *Speech Communication*, vol. 16, pp. 1–26, 1995.

[5] H. Altmann, I. Berger, and B. Braun, "Asymmetries in the perception of non-native consonantal and vocalic length contrasts," *Second Language Research*, vol. 28, no. 4, pp. 387–413, 2012.

[6] B. Kabak, T. Reckziegel, and B. Braun, "Timing of second language singletons and geminates," in *ICPhS in Hong Kong, 17-21 August 2011*, 2011, pp. 994–997.

[7] D. B. Pisoni, "Auditory and phonetic memory codes in the discrimination of consonants and vowels," *Perception & Psychophysics*, vol. 13, pp. 253–260, 1973.

[8] K. Tajima, K. Hiroaki, R. Amanda, A.-Y. Reiko, and M. Kevin G., "Training english listeners to perceive phonemic length contrasts in japanese," *The Journal of the Acoustical Society of America*, vol. 123, pp. 397–413, 2008.

[9] T. Isaacs and P. Trofimovich, "Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech," *Applied Psycholinguistics*, vol. 32, pp. 113–140, 2011.

[10] K. M. Dallett, "Intelligibility and short-term memory in the repetition of digit strings." *Journal of Speech and Hearing Research*, vol. 7, pp. 362–368, 1964.

[11] P. A. Luce, T. C. Feustel, and D. B. Pisoni, "Capacity demands in short-term memory for synthetic and .natural speech," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 25, no. 1, pp. 17–32, 1983.

[12] M. Antoniou, P. C. M. Wong, E. Ingvalson, and S. Wang, "Cognitive factors contribute to speech perception: Implications for sound change actuation," in *Poster presented at Workshop on Sound Change Actuation*. Chicago, USA: University of Chicago, 2013.

[13] A. Cutler, M. Cooke, M. L. G. Lecumberri, and P. D., "L2 consonant identification in noise: Cross-language comparisons," in *Proceedings of Interspeech*, Antwerp, Belgium, 2007, pp. 1585–1588.

[14] M. L. G. Lecumberri and M. Cooke, "Effect of masker type on native and non-native consonant perception in noise," *Journal of Acoustical Society of America*, vol. 119, pp. 2445–2454, 2006.

[15] M. Sonu, H. Kato, K. Tajima, R. Akahane-Yamada, and Y. Sagisaka, "Non-native perception and learning of the phonemic length contrast in spoken japanese: training korean listeners using words with geminate and singleton phonemes," *Journal of East Asian Linguistics*, vol. 22, no. 4, pp. 373–398, 2013.

[16] J. F. Werker and J. S. Logan, "Cross-language evidence for three factors in speech perception," *Perception & Psychophysics*, vol. 37, no. 1, pp. 35–44, 1985.

[17] N. Cowan and P. Morse, "The use of auditory and phonetic memory in vowel discrimination," *Journal of Acoustical Society of America*, vol. 79, pp. 500–507, 1986.

[18] E. Gerrits, "The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of acoustic speech cues," Ph.D. dissertation, Universititeit Utrecht, 2001.

[19] K. Johnson, "Cross-linguistic perceptual differences emerge from the lexicon," in *Proceedings of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*, G. Agwuele, W. Warren, and S.-H. park, Eds. Sommerville, MA: Cascadilla Press, 2004, pp. 26–41.

[20] M. Schouten and A. J. Van Hessen, "Modelling phoneme perception: I. categorical perception," *Journal of Acoustical Society of America*, vol. 92, pp. 1841–1855, 1992.

[21] B. McLaughlin, T. Rossman, and B. McLeod, "Second language learning: An information-processing perspective1," *Language Learning*, vol. 33, no. 2, pp. 135–158, 1983.

[22] A. Baddeley and B. A. Wilson, "Prose recall and amnesia: implications for the structure of working memory," *Neuropsychologia*, vol. 40, no. 10, pp. 1737–1743, 2002.

[23] A. L. Francis and H. C. Nusbaum, "Effects of intelligibility on working memory demand for speech perception," *Attention, Perception, & Psychophysics*, vol. 71, no. 6, pp. 1360–1374, 2009.

[24] T. Imada, R. Hari, N. Loveless, L. McEvoy, and M. Sams, "Determinants of the auditory mismatch response," *Electroencephalogr. Clin. Neurophysiol.*, vol. 87, no. 3, pp. 144–53, 1993.

[25] E. Sussman, "Integration and segregation in auditory scene analysis," *Acoustical Society of America*, vol. 117, no. 3, pp. 1285–1298, 2005.

[26] R. G. Crowder and J. Morton, "Precategorical acoustic storage (pas)," *Perception & Psychophysics*, vol. 5, pp. 365–373, 1969.

[27] A. D. Baddeley, *Working memory*. Oxford: Oxford University Press, 1986.

[28] A. Baddeley, "The episodic buffer: a new component of working memory?" *Trends in Cognitive Sciences*, vol. 4, no. 11, pp. 417–423, 2000.

[29] C. de Boor, *A practical guide to splines*. New York: Springer, 2001.

[30] M. Gubian, Y. Asano, S. Asaridou, and F. Cangemi, "Rapid and smooth pitch contour manipulation," in *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013, pp. 31–35.

[31] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, no. 3, pp. 255–278, 4 2013.

[32] I. Cunnings, "An overview of mixed-effects statistical models for second language researchers," *Second Language Research*, vol. 28, no. 3, pp. 369–382, 2012.

[33] S. Akaba, "An acoustic study of the japanese short and long vowel distinction," 2008.

[34] M. S. Han, "Acoustic manifestations of mora timing in japanese," *Acoustical Society of America*, vol. 96, no. 1, pp. 73–82, 1994.

[35] Y. Homma, "Durational relationships between japanese stops and vowels," *Journal of Phonetics*, vol. 9, no. 3, pp. 273–281, 1981.

[36] N. Macmillan and C. Creelman, *Detection theory: a user's guide*. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2005.

[37] J. Borràs-Comes, V. M. M., and P. Pieto, "The role of pitch range in establishing intonational contrasts in catalan," in *Proceedings of the 5th International Conference on Speech Prosody*, Chicago, 2010, pp. 1–4.

[38] F. Tomaschek, H. Truckenbrodt, and I. Hertrich, "Processing german vowel quantity: Categorial perception or perceptual magnet effect?" in *Proceedings of the 17th International Congress of the Phonetic Sciences*, 2011, pp. 2002–05.

[39] I. Darcy, L. Dekydtspotter, R. A. Sprouse, J. Glover, C. Kaden, M. McGuire, and J. H. Scott, "Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in l1 english– l2 french acquisition," *Second Language Research*, vol. 28, no. 1, pp. 5–40, 2012.

[40] S. Haraguchi, *The tone pattern of Japanese: An autosegmental theory of tonology*. Tokyo: Kaitakusha, 1977.

Investigating the relationship between accentuation, vowel tensity and compensatory shortening

Jessica Siddins, Jonathan Harrington, Ulrich Reubold, Felicitas Kleber

Institute of Phonetics and Speech Processing, Ludwig-Maximilians-Universität Munich, Germany

jessica | jmh | reubold | kleber@phonetik.uni-muenchen.de

Abstract

The aim of this study was to investigate the relationship between compensatory shortening and coarticulation in German tense and lax vowels and to determine whether this relationship was influenced by prosodic accentuation. While previous studies focussed on temporal vowel reduction due to compensatory shortening, and often found conflicting results, our study extends previous results by including a formant analysis of spatial reduction in two types of compensatory shortening. Specifically, we tested for polysyllabic shortening (monosyllabic vs. disyllabic words) and incremental coda shortening (words with final singletons vs. final clusters). Speakers produced minimal pairs differing in vowel tensity in accented and deaccented contexts for both shortening conditions. Vowel duration was influenced primarily by vowel tensity as well as by accentual lengthening for tense but not lax vowels. While vowel duration was not affected by compensatory shortening, formant analyses revealed an effect of coda cluster for tense vowels as well as clear effects of accentuation and vowel tensity. There was no effect of polysyllabic shortening on formants.

Further to previous studies on compensatory shortening, these results reveal that compensatory shortening is not limited to temporal reduction, but can have an impact on vowel quality as well.

Index Terms: coarticulation, speech timing, compensatory shortening, accentual lengthening, target undershoot, sound change

1. Introduction

This study investigated the effects of two types of compensatory shortening on the acoustic duration and formant values of vowels in rhythmically strong syllables. Our general goal within a larger series of experiments is to better understand the relationship between prosodic weakening, coarticulation and sound change.

The first type of compensatory shortening we investigated, polysyllabic shortening, refers to the compression of a vowel spoken in a polysyllabic compared with a monosyllabic word. Thus, the vowel in English *sleep* is longer than the same vowel in English *sleepy* [1]. Polysyllabic shortening has been well-documented in Germanic languages such as English [1, 2, 3, 4, 5, 6], Dutch [7] and Swedish [8, 9, 10, 11, 12], although [13] found no evidence of polysyllabic shortening in a study of read speech in English.

The second type of compensatory shortening we investigated, incremental coda shortening, refers to the compression of a vowel spoken before a consonant cluster compared with a consonant singleton. [14] found acoustic shortening of vowels before consonant clusters in their study of three speakers of English, and [15] replicated these effects for obstruent clusters (but not for sonorant clusters in syllable-final position). A recent study on German by [16] found no results of incremental coda shortening in production, but they did find that listeners expected shorter vowel durations before complex clusters in a perception experiment.

Prior research has largely failed to establish the effect of accentuation on compensatory shortening (see [17]), in that studies have mostly examined stressed syllables or accented words only. In German, as in other Germanic languages, a word is accented when a pitch accent is associated with the rhythmically strongest syllable [18]. In addition to the f0 changes caused by a pitch accent and any adjacent boundary tones, pitch accented syllables in many Germanic languages are also hyperarticulated and produced with greater duration [19].

Thus, this study aimed to investigate vowel reduction and the extent to which compensatory shortening interacts with prosodic accentuation and vowel tensity. Based on previous studies, we had three main hypotheses for the durational analysis as well as analogous hypotheses for the formant analysis. Firstly, we expected acoustic vowel shortening and thus also vowel undershoot before coda clusters compared with coda singletons [14, 15] as well as in disyllabic compared with monosyllabic words [11, 20]. Secondly, we expected compensatory shortening and undershoot induced by this shortening to be more marked in accented than in deaccented contexts [4, 17, 21, 22]. Thirdly, we expected more shortening and hence more shortening-induced undershoot of tense than of lax vowels [4, 23, 24].

2. Method

2.1. Participants

Twenty-nine L1 speakers of Standard German (11 male, 18 female) were recorded at a sampling rate of 44 100 Hz in a sound-attenuated booth. Speakers had no known speech or language impairments and were paid for their participation.

2.2. Stimuli

We chose a tense-lax target vowel pair believed to vary mostly in quantity and only minimally in quality in order to study the effects of compensatory shortening. In German, tense vowels are phonologically long, while lax vowels are phonologically short, for example /ˈbiːtn̩/ (to offer) vs. /ˈbɪtn̩/ (to request). The German vowel that differs least in quality between its tense and lax versions is the open central /a/, which varies mainly in vowel height (F1), for example a lower F1 for lax *Kamm* /kam/ (comb) and a higher F1 for tense *kam* /kaːm/ (came) [23, 24]. [23, p341] claim that even the durational contrast is minimised in prosodically weak contexts. Our stimuli were real German words with

lax /a/ and tense /aː/. We restricted the target vowels to one consonantal context in order to avoid varying effects of CVC coarticulation. The target words were embedded in phrase-medial position in the carrier sentence *Anna hatte [target word] verstanden* (Anna understood [target word]) in order to avoid utterance or phrase-final lengthening.

	Cluster Shortening		
Lax Vowels	/zak/	/zakt/	/zaktə/
Tense Vowels	/zaːk/	/zaːkt/	/zaːktə/
		Polysyllabic Shortening	

Table 1: The 6 target words, spoken in both accented (A) and deaccented (U) contexts (= 12 target words).

A minimal pair paradigm was created using one consonantal context adjusted for Syllabicity (monosyllabic vs. disyllabic), Coda (final singleton vs. final cluster), Accentuation (accented vs. deaccented) and Vowel Tensity (tense vs. lax). The monosyllabic level of the Syllabicity condition overlapped with the cluster level of the Coda condition (see Table 1), leading to a total of 12 target items. Each speaker produced 10 repetitions of each item, leading to a total of 120 utterance tokens for each speaker (29 subjects x 120 utterances = 3 480 utterances). Filler words and sentences were also included to disguise the object of the experiment.

2.3. Procedure

The stimuli were presented and recorded in randomised order using SpeechRecorder software [25]. Participants were first presented with a question designed to elicit a narrow focus on the target word for the accented context and a broad focus for the deaccented context: either *WAS hatte Anna verstanden?* (WHAT did Anna understand?) or *WER hatte [target word] verstanden?* (WHO understood [target word]?). The target sentence was then presented with the accented word in capital letters. The experimenter asked subjects to repeat the sentence if they made a mistake (either segmentally or suprasegmentally).

Three subjects were eliminated from further analysis as they were unable to elicit the correct accentuation patterns of the stimuli. The remaining 26 speakers (9 males, 17 females; mean age 24 years) were included in the analysis.

The entire corpus was automatically segmented and labelled using the Munich Automatic Segmentation System [26] and corrected manually. During this procedure several rare cases of incorrect accentuation were discovered and removed from the database.

2.4. Analysis

For the durational analysis, the dependent variable was normalised vowel duration. We normalised the vowel durations by dividing the duration of each target vowel by the duration of each /a/ in the utterance-final word *verstanden* in order to control for changes in speech tempo throughout the experiment. *Verstanden* was chosen as it was considered least likely to be affected by the varying accentuation pattern, which was confirmed by a visual analysis of the data. The normalised vowel duration was then averaged per speaker and per condition.

For the formant analysis, the dependent variable was maximum F1 measured in Hz from the central third of the vowel. We calculated formants with a frame shift of 5 ms and a window length of 12.5 ms for female speakers and 20 ms for male speakers using Emu [27]. Formant errors were hand-corrected in Emu when necessary.

Both dependent variables were tested individually alongside within-subjects factors Vowel Tensity, Accentuation and Syllabicity (for the polysyllabic shortening analyses) or Coda (for the cluster shortening analyses) and random factor Speaker in a repeated measures ANOVA design using the `ez` package in `R` [28, 29].

3. Results

3.1. Cluster Shortening

For the cluster shortening condition, we compared the monosyllables with and without final clusters (the stimuli coloured light grey in Table 1). Disyllabic stimuli were excluded from analysis.

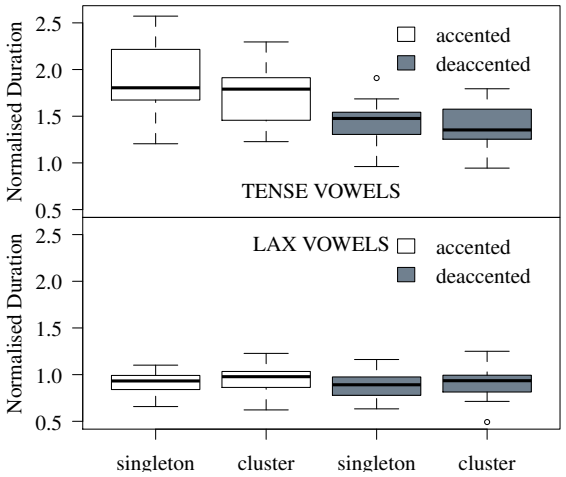


Figure 1: Effects of incremental coda shortening, accentuation and vowel tensity on normalised vowel duration

The durational analysis revealed main effects of both Vowel Tensity ($F[1, 25] = 278.7; p < .001$) (top vs. bottom of Figure 1) and Accentuation ($F[1, 25] = 54.6; p < .001$) (white vs. grey in Figure 1), but there was no effect of Coda on the normalised vowel duration. We found significant interactions between Vowel Tensity and Coda ($F[1, 25] = 6.3; p < .05$) and Vowel Tensity and Accentuation ($F[1, 25] = 82.2; p < .001$). Post-hoc Bonferroni-corrected t-tests revealed accentual lengthening of tense ($p < .001$) but not lax vowels. The post-hoc tests did not reveal the reason for the interaction between Vowel Tensity and Coda, but Figure 1 shows a slight tendency toward compensatory shortening of tense vowels which cannot be seen for lax vowels.

The formant analysis revealed main effects of Vowel Tensity ($F[1, 25] = 105.7; p < .001$) (black vs. grey in Figure 2), Accentuation ($F[1, 25] = 102.8; p < .001$) (solid vs. dashed in Figure 2) and Coda ($F[1, 25] = 5.8; p < .05$) on the maximum F1, with a significant interaction between Vowel Tensity and Coda ($F[1, 25] = 8.5; p < .01$) . Pairwise post-hoc comparisons for the Vowel Tensity vs. Coda interaction indicated that the effect of Coda on F1 is restricted to tense vowels only, although the effect is weak (see Figure 3).

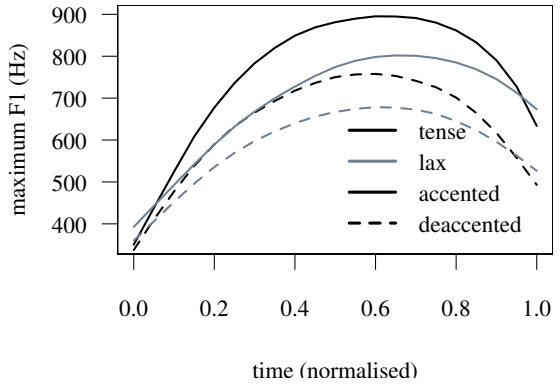


Figure 2: F1 trajectories (linear time-normalised) for the cluster shortening experiment as a function of vowel tensity and accentuation

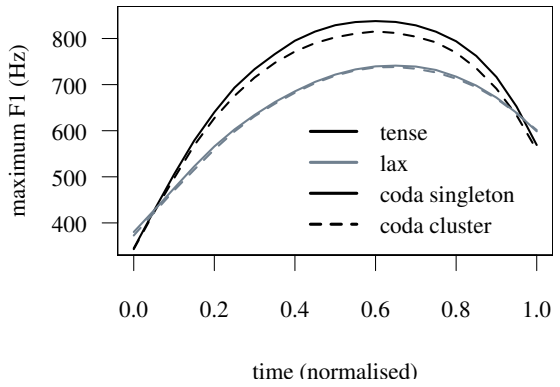


Figure 3: F1 trajectories (linear time-normalised) for the cluster shortening experiment as a function of vowel tensity and coda condition

3.2. Polysyllabic Shortening

For the polysyllabic shortening condition, we excluded all monosyllabic words with final singleton from the analysis and compared the monosyllabic and disyllabic words with clusters at the end of the first syllable (the stimuli coloured dark grey in Table 1).

While we found main effects of both Vowel Tensity ($F[1, 25] = 33.2; p < .001$) (top vs. bottom of Figure 4) and Accentuation ($F[1, 25] = 401.9; p < .001$) (white vs. grey in Figure 4), there was no effect of Syllabicity on the normalised vowel duration. We also found a significant interaction between Vowel Tensity and Accentuation ($F[1, 25] = 31.6; p < .001$). Post-hoc Bonferroni-corrected t-tests revealed a significant effect of accentual lengthening on tense vowels ($p < .001$), but not on lax vowels.

The F1 analysis found main effects of Vowel Tensity ($F[1, 25] = 100; p < .001$) (black vs. grey in Figure 5) and Accentuation ($F[1, 25] = 103.7; p < .001$) (solid vs. dashed in Figure 5), but not of Syllabicity. There was a significant interaction between Vowel Tensity and Accentuation ($F[1, 25] = 9.3; p < .01$). In view of Figure 5, this interaction is likely due to a slightly larger difference between tense and lax vowels in accented contexts than in deaccented contexts (see Figure 5).

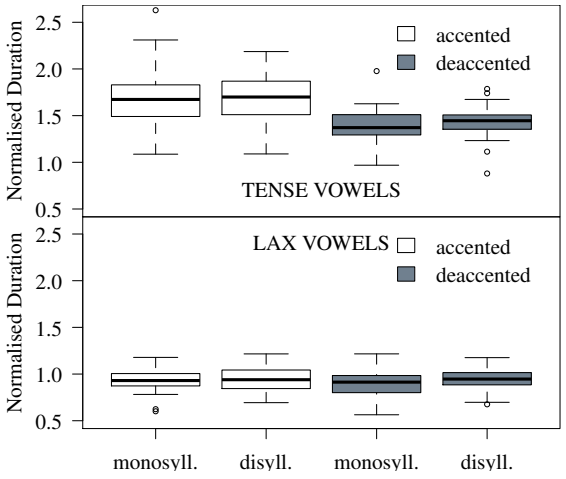


Figure 4: Effects of polysyllabic shortening, accentuation and vowel tensity on normalised vowel duration

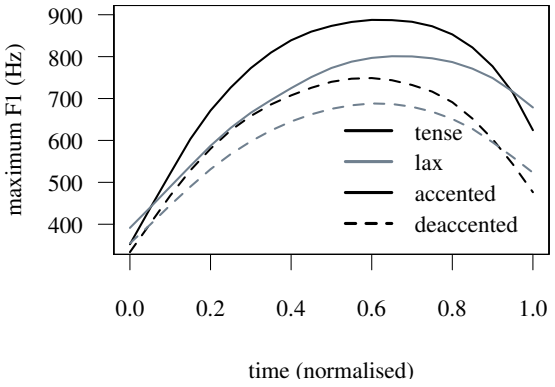


Figure 5: F1 trajectories (linear time-normalised) for the polysyllabic shortening experiment as a function of vowel tensity and accentuation

4. Discussion

We expected compensatory shortening and target undershoot of vowels before coda clusters (incremental coda shortening) and in disyllabic words (polysyllabic shortening), and for these effects to be more marked in accented than in deaccented words. In addition, we expected tense vowels to undergo more compensatory shortening and shortening-induced vowel reduction than lax vowels.

4.1. Vowel Tensity

We confirmed that the tense-lax distinction (at least of low vowels) is defined by both vowel quantity and quality in German.

The quantity distinction was maintained even in deaccented contexts, and there was accentual lengthening of tense but not lax vowels. In terms of quality, we found lower first formants in lax than in tense vowels and in deaccented tense vowels than in accented tense vowels (black vs. grey in Figures 2, 3 & 5), contrary to [24, p14]'s conclusion that "tense and lax low vowels [i.e. /a, aː/] in [Standard] German differ consistently only in duration, but not in formant structure". A clear distinction in F1 between tense and lax low vowels may be necessary because the

durational distinction between tense and lax vowels lessens (but is not neutralised) in deaccented contexts (see Figures 1 and 4; see also [23, 24]).

4.2. Accentuation

We found clear effects of accentual lengthening of tense but not lax vowels. In line with [30], who found larger jaw movements in stressed than in unstressed syllables, there was a significant effect of accentuation on the F1 of both tense and lax vowels, and this effect was slightly stronger for tense vowels (see also [23]). As a result, there is less difference in duration between tense and lax /aː/ and /a/ in deaccented contexts. In Figures 2 and 5, a large amount of overlap between tense and lax vowels is visible: /a/ has a higher first formant (indicating greater jaw opening and greater tensity) in the accented condition than /aː/ in the deaccented condition.

4.3. Cluster Shortening

Contrary to our hypothesis and some previous studies [14, 15], we found no effect of complex coda on the duration of the target vowel. Our findings thus match those of [16] and are compatible with the c-centre hypothesis [31], which predicts incremental onset shortening but not incremental coda shortening.

The conflicting results of incremental coda shortening emphasise the need to investigate whether shortening in fact induces other types of reduction. Indeed, we did find significant spatial reduction of F1 in tense vowels before complex clusters. That is, incremental coda shortening (at least in German) appears to induce F1 undershoot of primary-stressed tense vowels, even if there is no durational shortening. This provides support for our hypothesis that there is greater undershoot before clusters than before singletons, thus leading to a smaller difference between tense and lax vowels before coda clusters (see Figure 3).

There is no evidence for our hypothesis that undershoot is greater in accented than in deaccented contexts: that is, the degree of separation in F1 between tense and lax vowels before singletons or clusters is largely unaffected by accentuation. Thus, while accentuation does influence vowel quality and quantity, it does not interact with incremental coda shortening.

4.4. Polysyllabic Shortening

In this study, we found no effect of polysyllabic shortening on vowel duration. This result is contrary to the main body of research on polysyllabic shortening, but in line with [13], who found no polysyllabic shortening in connected speech. However, their study differed from others in that it was based on syllable shortening in stress groups rather than in polysyllabic words. In addition, they counted syllables with secondary stress as stressed syllables.

According to [4]'s incompressibility theory, "a vowel that is shortened by one rule becomes less compressible to additional shortening influences" [4, p1103]. As a result, one might argue that there was no polysyllabic shortening of the target vowel because both the monosyllabic and the disyllabic condition contained a final coda cluster, which can induce incremental coda shortening. However, as there was no effect of incremental coda shortening on vowel duration, it is unlikely that the final coda cluster prevented shortening of the target vowel in the disyllable compared with the monosyllable.

In addition, there was no effect of polysyllabic shortening on the first formant, which is known to be correlated with jaw opening. There is no evidence from our results that polysyllabic shortening induces greater durational compression in accented contexts than in deaccented contexts and in tense vowels than in lax vowels. As a result, there is no evidence from this study that polysyllabic shortening induces any type of vowel reduction, neither temporal nor spatial.

4.5. Implications for Sound Change

The results of this study show slightly less durational contrast between tense and lax vowels in deaccented contexts (white vs. grey in Figures 1 and 4) and a slight decrease in the quality contrast before coda clusters (see Figure 3). In general, there is a great amount of overlap in the first formant of accented lax vowels and deaccented tense vowels (see Figures 2 and 5).

If tense and lax vowels are more difficult to distinguish before clusters based on their vowel quality, this might be the reason the durational contrast between tense and lax vowels was maintained in our analysis of acoustic shortening before coda clusters. Alternatively, the tense-lax contrast may be diminishing in German [32], and this might be a context in which listeners could misperceive the vowel, leading to a mini sound change [33].

4.6. Outlook

In view of the above, there may be a greater risk of confusing /aː/ and /a/ before clusters, not because of acoustic shortening, but because of F1 undershoot. In order to determine how listeners parse this undershoot, we are now running perception experiments in which a synthetic vowel continuum is created and spliced into two contexts: /zaːk/-/zak/ and /zaːkt/-/zakt/. For the target vowel, we chose an ambiguous vowel duration (the mean of tense and lax tokens of a speaker) and varied only the height of the first formant. If listeners attribute vowel undershoot before coda clusters to compensatory shortening, we would expect a higher F1 at the category boundary for the continuum ending in coda cluster than the one ending in coda singleton. Alternatively, if listeners incorrectly attribute vowel undershoot before coda clusters to (lax) vowel quality, i.e. they are not aware of the effects of compensatory shortening on vowel quality, we would expect the continuum before coda singleton and the continuum before coda clusters to have the same category boundary.

5. Conclusions

This study examined the relationship between vowel tensity, accentuation and compensatory shortening and in particular how these factors affect vowel quantity and quality. At least for the vowel pair we tested, vowel tensity in German is clearly marked by both duration and quality. In addition, accentuation affects both duration and quality, but not equally for tense and lax vowels. This asymmetry leads to considerable overlap between tense and lax vowels. We found no effects of compensatory shortening on duration, but we did find F1 undershoot before coda clusters. Thus, the quality contrast is diminished before clusters. A perception experiment is being carried out to determine whether or not listeners correctly attribute cluster-induced vowel undershoot to its source. If not, this context could be a source of sound change.

6. Acknowledgements

This research was supported by European Research Council grant 295573 to Jonathan Harrington.

7. References

[1] I. Lehiste, "The timing of utterances and linguistic boundaries," *JASA*, vol. 51, pp. 2018–2024, 1972.

[2] D. Jones, "Chronemes and tonemes," *Acta Linguistica*, vol. 4, no. 1, pp. 11–10, 1944.

[3] T. P. Barnwell, "An algorithm for segment durations in a reading machine context," Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, Tech. Rep. 479, 1971.

[4] D. H. Klatt, "Interaction between two factors that influence vowel duration," *The Journal of the Acoustical Society of America*, vol. 54, no. 4, pp. 1102–1104, 1973.

[5] ——, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[6] R. F. Port, "Linguistic timing factors in combination," *The Journal of the Acoustical Society of America*, vol. 69, no. 1, pp. 262–274, 1981.

[7] S. G. Nooteboom, "Production and perception of vowel duration: a study of durational properties of vowels in Dutch." Ph.D. dissertation, University of Utrecht, 1972.

[8] S. Öhman, "Syllabic function of vowel length," *STL-QPSR*, vol. 1, pp. 7–9, 1961.

[9] C. C. Elert, *Phonologic Studies of Quantity in Swedish*. Stockholm: Almqvist & Wiksell, 1964.

[10] B. Lindblom, "Temporal organization of syllable production," *Speech Transmission Lab. Quarterly Progress Status Report*, vol. 2, no. 3, pp. 1–5, 1968.

[11] B. Lindblom and K. Rapp, "Reexamination of the compensatory adjustment of vowel duration in Swedish words," *Occasional Papers, University of Essex*, vol. 13, pp. 204–224, 1972.

[12] ——, "Some temporal regularities of spoken Swedish," *Papers in Linguistics from the University of Stockholm*, vol. 21, pp. 1–59, 1973.

[13] T. H. Crystal and A. S. House, "Articulation rate and the duration of syllables and stress groups in connected speech," *Journal of the Acoustical Society of America*, vol. 88, pp. 101–112, 1990.

[14] K. Munhall, C. H. Fowler, S. Hawkins, and E. Saltzman, ""Compensatory shortening" in monosyllables of spoken English." *Journal of Phonetics*, vol. 20, no. 2, pp. 225–239, 1992.

[15] S. Marin and M. Pouplier, "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model," *Motor Control*, vol. 14, no. 3, pp. 380–407, 2010.

[16] S. Peters and F. Kleber, "Compensatory vowel shortening before complex coda clusters in the production and perception of german monosyllables," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 4202–4202, 2013.

[17] L. White, "English speech timing: a domain and locus approach," Ph.D. dissertation, University of Edinburgh, 2002.

[18] J. Pierrehumbert and M. Beckman, *Japanese Tone Structure*. Cambridge: MIT Press, 1988, vol. Linguistic Inquiry Monograph 15.

[19] K. de Jong, "The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 491–504, 1995.

[20] I. Lehiste, *Suprasegmentals*. Cambridge, Massachusetts: The MIT Press, 1970.

[21] A. E. Turk and S. Shattuck-Hufnagel, "Word-boundary-related duration patterns in English," *Journal of Phonetics*, vol. 28, no. 4, pp. 397 – 440, 2000.

[22] L. White and A. E. Turk, "English words on the Procrustean bed: Polysyllabic shortening reconsidered," *Journal of Phonetics*, vol. 38, pp. 459–471, 2010.

[23] C. Mooshammer and S. Fuchs, "Stress distinction in German: simulating kinematic parameters of tongue-tip gestures," *Journal of Phonetics*, vol. 30, no. 3, pp. 337 – 355, 2002.

[24] M. Jessen, "Stress conditions on vowel quality and quantity in German," *Working Papers of the Cornell Phonetics Laboratory*, vol. 8, pp. 1–27, 1993.

[25] C. Draxler and K. Jänsch, "SpeechRecorder - a Universal Platform Independent Multi-Channel Audio Recording Software," in *Proc. of the IV. International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004, pp. 559–562.

[26] F. Schiel, C. Draxler, and J. Harrington, "Phonemic segmentation and labelling using the MAUS technique," Philadelphia, PA, USA, 2011. [Online]. Available: http://epub.ub.uni-muenchen.de/13684/

[27] J. Harrington, *Phonetic Analysis of Speech Corpora*. Wiley Publishing, 2010.

[28] M. A. Lawrence, "Package 'ez'," Sep. 2013. [Online]. Available: http://cran.r-project.org/web/packages/ez/ez.pdf

[29] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/

[30] R. Kent and R. Netsell, "Effects of stress contrasts on certain articulatory parameters," *Phonetica*, vol. 24, pp. 23–44, 1971.

[31] C. P. Browman and L. Goldstein, "Competing constraints on intergestural coordination and self-organization of phonological structures," *Bulletin de la Communication Parlee*, vol. 5, pp. 25–34, 2000.

[32] G. Seiler, "How contrastive vowel quantity can become non-contrastive," in *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, vol. 40, no. 1. Chicago Linguistic Society, 2004, pp. 349–363.

[33] J. J. Ohala, "The listener as a source of sound change," in *Papers from the Parasession on Language and Behavior*, C. S. Masek, R. A. Hendrick, and M. F. Miller, Eds. Chicago: Chicago Linguistic Society, 1981, pp. 178–203.

The form and use of uptalk in Southern Californian English

Amanda Ritchart[1], *Amalia Arvaniti*[2]

[1] University of California, San Diego
[2] University of Kent
aritchart@ucsd.edu, a.arvaniti@kent.ac.uk

Abstract

This study examines the phonetics, phonology and pragmatic function of uptalk, utterance-final rising pitch movements, as used in Southern Californian English. Twelve female and eleven male speakers were recorded in a variety of tasks. Instances of uptalk were coded for discourse function (statement, question, confirmation request, floor holding) based on context. The excursion of the pitch rise and the distance of the rise start from the onset of the utterance's last stressed vowel were also measured. Confirmation requests and floor holding showed variable realization. Questions, on the other hand, showed a rise that typically started within the stressed vowel and had a large pitch excursion, while uptalk "proper", i.e. uptalk used with statements, exhibited both a smaller pitch excursion and a later rise that often started after vowel offset. This pattern suggests that statements have a L* L-H% melody while questions have L* H-H%. Gender differences were also found: female speakers used uptalk more often than males, and showed greater pitch excursion and later alignment, all else being equal. Other social parameters, however, such as social class and linguistic background, did not affect the use of uptalk.

Index Terms: intonation, HRT, uptalk, English, sociolinguistics, gender

1. Introduction

Rising melodies used with statements, commonly referred to as *uptalk* or *high rise terminals* are common in many varieties of English. Here we use the term *uptalk* which better reflects the Southern Californian patterns that are the focus of our investigation. Research on uptalk in some varieties is quite extensive, but has often been impressionistic [1]. The varieties that have been most investigated include those spoken in Australia and New Zealand as well as UK varieties from Glasgow and Belfast [1]–[5] (and references therein).

These studies document that different tunes are used for uptalk across varieties. Thus, [1] report that Australian uptalk is realized as either L* H-H% or H* H-H%. For Glasgow, L*H H-L% is proposed for the "rise-plateau-slump" type of uptalk, with suspension of the rule that in other English varieties upsteps a L% after a H- phrase accent [2]. In [1], New Zealand uptalk is analyzed as reflecting two main patterns, LH* H-H% and L* H-H% (based on [3]), but a newer study suggests that New Zealand English may exhibit change in progress with respect to uptalk [4].

In addition to differences in form, uptalk across varieties of English is used for different purposes. Thus, [1] report that in Australian English upstep is used both with questions and declarative statements; upstepped statements are particularly frequent when the speaker wishes to hold the floor. This leads [1] to suggest that the intonational difference between statements and questions and that between statements and continuation is neutralized in Australian English. New Zealand English also uses uptalk for both statements and questions but the tunes used for each function are becoming increasingly distinct [4]. Research on Glasgow and Belfast English, e.g., [2] and [5], focused on form rather than function, but recent research suggests that uptalk, in Belfast at least, may have its origins in list intonation [6].

One of the varieties that is stereotypically known as exhibiting use of uptalk is Californian English, particularly the varieties spoken in the south (henceforth *SoCal*). The use of uptalk in SoCal is often referred to as "valley girl speak" and is often assumed to be a feature of younger females only, though no studies exist, to our knowledge, confirming or refuting this general lay perception.

Here we present data from SoCal English which show that the use of uptalk is widespread in this variety and exhibits gender-related variation. We further show that SoCal uptalk tunes are different from those reported for other varieties of English, and that speakers retain systematic differences between uptalk used in statements and other types of uptalk, such as pitch rises used with questions. Differences apply both to the tunes employed and to the scaling of the rise.

2. Methods

2.1. Speakers

Twenty-three speakers were recorded for the study, eleven male and twelve female. They were all native speakers of SoCal English, from San Diego (N = 7), Orange (N = 6), Los Angeles (N = 8), and Riverside (N = 2) counties. Fifteen were monolingual, while the other eight reported being bilingual in English and one of the following languages: Vietnamese (N = 3), Japanese (N = 1), Armenian (N = 1), Assyrian (N = 1), Spanish (N = 1), and Cantonese (N = 1). The speakers' ethnic backgrounds varied: twelve self-identified as Asian, six as Hispanic and five as White.

The MacArthur Scale of Subjective Social Status (cf. [7], [8]) was used to determine the speakers' socioeconomic status or *SEC*, a rather fluid concept in California. Participants found the use of the scale easy and intuitive. They were classed into three groups based on their responses: lower (rungs 1-4, N = 4), middle (rungs 5-7, N = 13) and upper (rungs 7-10, N = 6).

2.2. Materials, Tasks and Procedures

Recordings took place in the recording studio of the UCSD Phonetics Lab, using an AD converter at 48 KHz and 16-bit quantization. Four types of data were collected from each speaker: (a) map task; (b1) reading of the transcript of a popular sitcom scene; (b2) retelling of the sitcom scene; (c) controlled materials consisting of isolated questions and statements. For the first 17 participants, tasks were presented in the following order: (c), (b), (a). (For task (b), the retelling of the clip always followed the reading of the transcript.) To control for possible order effects, the order of tasks was

counterbalanced for the other recordings and each participant was randomly assigned to one of three possible orders (Latin square design): abc, bca, or cab. However, given that the data consist largely of spontaneous speech it is unlikely that order could have severely biased speaker productions with respect to uptalk.

For the map task, maps with local (or local sounding) landmarks were designed as illustrated in Figure 1. In the map task, the participants acted as leaders with the follower being either the first author or an undergraduate research assistant (both females and native SoCal speakers). For task (b1), a scene from either *Scrubs* or *How I met your mother* was used; the show chosen was the one the participant was less familiar with. Lack of familiarity was sought so that speakers would not imitate the actors' accents. The scene was muted and participants were given a transcript of the dialogue while they watched the clip. When they were ready, they chose which character they were most comfortable reading from, and participated in reading aloud the transcript in a dialogue with the experimenter. In task (b2), participants had to retell the same scene in their own words. For task (c), the participants read aloud a list of 49 sentences. These were statements and questions constructed for the study. In these sentences, the number of syllables and position of stress was controlled in order to examine the realization of specific tonal events; see (1) for an illustration. Here we report on the results from tasks (a) and (b2).

(1) a. *Did Anne and Mel eat the lime?*
b. *Did Annabelle and Melinda eat the lime?*

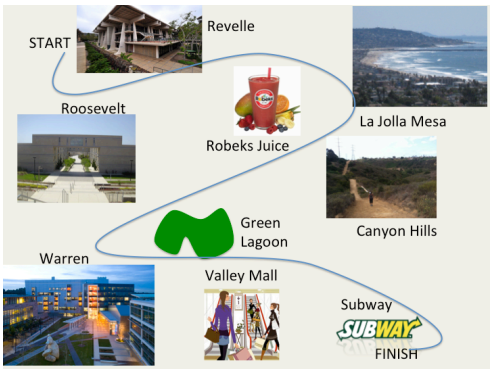


Figure 1: *The Instruction Giver's map used in the map task.*

2.3. Analysis and Measurements

The analysis involved both a categorization of each instance of uptalk in terms of its discourse function and acoustic measurements with respect to the alignment and scaling of the pitch rise associated with each uptalk token.

Specifically, instances of uptalk were classed in one of four discourse functions: question, statement, holding the floor, and confirmation request. A token was considered to be a question when it was syntactically marked as such, for example by showing inversion. Confirmation requests were indirect questions: they were not syntactically questions, but the context and interlocutor response indicated that the speaker was indirectly asking if their interlocutor was paying attention, agreed or understood. Holding the floor was defined as an utterance indicating that the speaker did not intend to cede the floor, in that s/he continued talking with either a minimal or no pause and was not interrupted by their interlocutor. All other instances of uptalk were identified as statements. These were regular declaratives for which no other discourse function was apparent from context; e.g., such utterances did not elicit information from the interlocutor. For cases in which the discourse context was ambiguous, a forced choice was made by the first author who is a native speaker of the dialect.

In addition, the scaling and alignment of the rise was annotated using the facilities of Praat [9]. The beginning of the rise was manually located as the point at which an upward trend was apparent in which successive F0 values differed by more than 5 Hz (this was done to exclude microprosodic variation); see Figure 2 for an illustration. Scaling was measured in Hz and defined as the difference between the F0 at the beginning of the rise and the highest F0 point at its end. After the F0 information was extracted, values in Hz were converted to ERB in order to better compare male and female voices. The alignment of the F0 rise was defined as the distance of the point annotated as the start of the rise from the onset of the last stressed vowel in the utterance. This measurement was based on the assumption (supported by the data) that the last content word is typically the one carrying the nuclear pitch accent.

3. Results

Results presented here are related to the function, scaling and alignment of rises and to differences in gender. We note that ethnicity, SEC status and bilingualism did not affect the use of uptalk; thus they will not be discussed further. All significance testing was determined using linear mixed-effects models with Speaker as a random intercept. P-values are given with respect to model comparisons and are reported with the χ^2 statistic, which compares the (reduced) model without the fixed effect and the (full) model with the fixed effect.

3.1. Discourse Functions and Distribution of Uptalk

The best-fit model for comparing uptalk against other utterances in the corpus included discourse function, task type, gender and an interaction between gender and discourse function as fixed effects. Uptalk was more frequent in the map task than in clip retell: 34% of the utterances in the map task ended in uptalk as opposed to only 20% of utterances in clip retell [$\chi^2(1) = 37.4$, $p < 0.001$]. Uptalk was also used more frequently, approximately twice as often, by female than male speakers: uptalk comprised 42% of the female speakers' utterances vs. 20% of the male speakers' utterances [$\chi^2(1) = 14.1$, $p < 0.001$].

Gender also interacted with discourse function [$\chi^2(3) = 16.9$, $p < 0.001$]. First, no gender or discourse differences were found for uptalk in questions and confirmation requests in the corpus: uptalk was used for both types of utterances in 100% of the tokens independently of speaker gender. Statements and floor holding, on the other hand, showed different frequencies for uptalk, with floor holding being signaled by uptalk significantly more frequently than statements: 45% of floor holding ended in uptalk vs. 16% of statements [$\chi^2(3) = 244.7$, $p < 0.001$]. However, while females and males used uptalk with statements equally frequently, females used uptalk to hold the floor significantly more frequently than males; indeed females used uptalk more than twice as much as males for floor holding. This is illustrated in Figure 3.

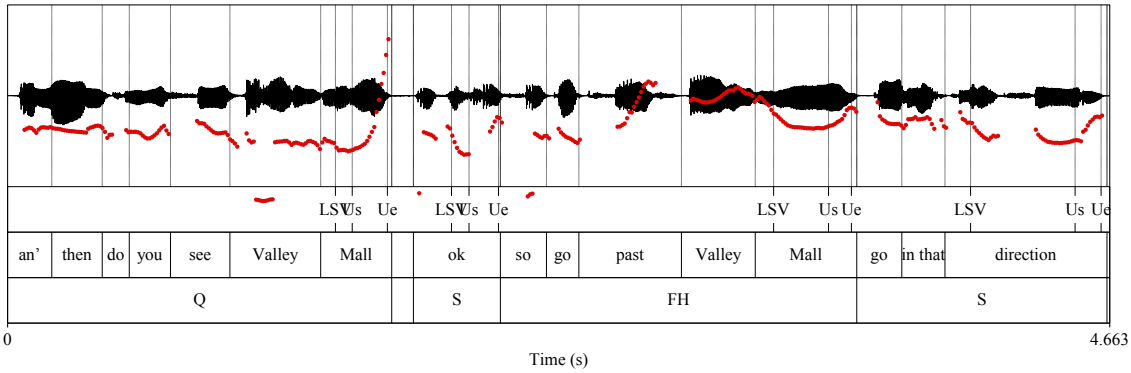


Figure 2: *Example of data annotation from the map task. LSV = last stressed vowel; Us = start of uptalk rise; Ue = end of uptalk rise; Q = question; S = statement; FH = floor holding. The follower's response ("yes, I do") which followed the question in this example has been removed for clarity.*

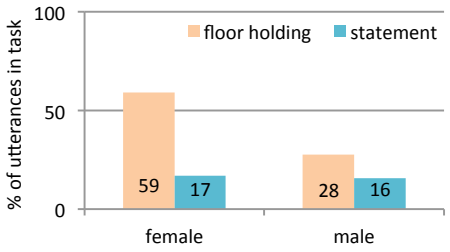


Figure 3: *Proportion of uptalk used by discourse function and gender.*

3.2. Alignment of Uptalk Rise

The best-fit model for the alignment of the uptalk rise included discourse function and gender as fixed effects. In this model, only two levels of discourse function were included, statement and question. Floor holding and confirmation requests were omitted from the model as their alignment was too variable.

The results from statements and questions showed a consistent difference between the onset of the rise in statements vs. questions, with the former having significantly later alignment than the latter [$\chi^2(1) = 19.3$, $p < 0.001$]. Specifically, the rise in the questions included the last stressed vowel (which is presumed to carry the nuclear pitch accent) while in statements the rise started after this vowel. The difference in the alignment of the rise in statements and questions is illustrated in Figure 4, which also shows the effect of gender on alignment. Specifically, uptalk produced by female speakers showed later alignment than uptalk produced by male speakers both for statements and questions [$\chi^2(1) = 5.6$, $p = 0.02$]. The differences were quite substantial, particularly for the questions: male speakers started the rise just before the last stressed vowel on average, while the rise for the female speakers started within this vowel.

3.3. Scaling of the Uptalk Rise

The best-fit model for the scaling of the rise included discourse function [$\chi^2(3) = 19.4$, $p < 0.001$], gender [$\chi^2(1) = 27.01$, $p < 0.001$] and task type [$\chi^2(1) = 20.03$, $p < 0.001$] as fixed effects. The major difference in pitch excursion with respect to discourse function was between statements and the other functions, with statements showing approximately half the pitch rise than questions, confirmation requests and floor holding (see Figure 5). Differences between these last three discourse functions were also statistically significant but minimal in actual terms [questions, confirmation requests > floor holding].

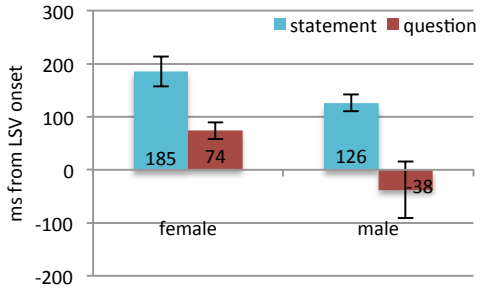


Figure 4: *Mean rise alignment (with standard error bars) per type of discourse function and gender. Negative values represent a rise beginning before the onset of the last stressed vowel (LSV).*

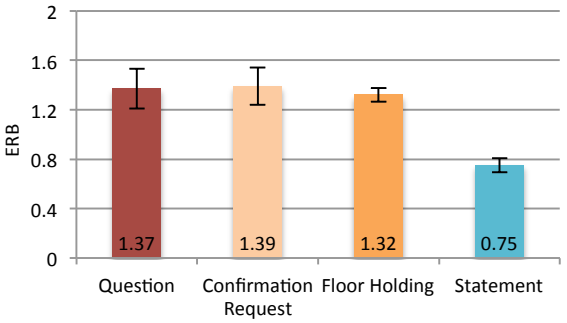


Figure 5: *Mean scaling of rises (with standard error bars) per discourse function.*

In addition, the data showed that female speakers had generally greater pitch excursions than males (see Figure 6a), presumably a reflection of gender differences in the use of the

frequency and effort codes [10]. Further, pitch excursions associated with uptalk were significantly larger in the map task than in clip retell (see Figure 6b). Neither result interacted with discourse function, however, suggesting these are independent effects and not the result of, e.g., female talkers asking more questions, or speakers in general making more confirmation requests in the map task than in clip retell.

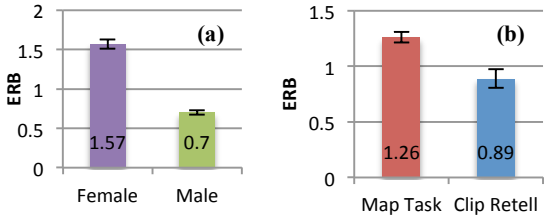


Figure 6: *On the left, mean scaling of uptalk (with standard error bars) by gender; on the right, mean scaling of uptalk (with standard error bars) by type of task.*

4. Discussion and Conclusions

Given the above results, we propose that the melody *typically* used with questions in SoCal English is L* H-H% and that used with statements is L* L-H%. The difference in phonological composition accounts both for the difference in alignment reported above but also for the difference in the scaling of the pitch rise: L-H% results in a lower rise than H-H%. Questions, as noted, can show a rise on the stressed syllable, a contour that could be interpreted as the reflex of a bitonal LH accent. However, the auditory impression is that of a low pitch accent, while the use of either L*H or LH* in questions is pragmatically doubtful (cf. [11] on the pragmatics of L*H when followed by a rise). Independently of the representation adopted for the question tune, the fact remains that questions and statements are not relying on the same melody as is often assumed; to put it differently, SoCal statements with uptalk do not sound like questions.

Our results further show that SoCal English makes a distinction between uptalked statements and questions even when the same melody is used (as happens occasionally). In particular, although the distinction is typically realized as a choice of tune, as noted above, it can also be signaled by just differences in the pitch scaling of the final rise (cf. the question and statements in Figure 2). The difference in pitch rise scaling is particularly evident when the tune used is H* H-H%, a variant that was attested but was not as frequent in our data as the L* accent variants. If such differences in the scaling of the rise turn out to be used by listeners to interpret the pragmatic intent of an utterance, this would suggest the need to incorporate scaling contrasts beyond H vs. L in phonological representations of intonation.

The two main melodies L* H-H% and L* L-H% are also used for floor holding and confirmation requests except that these two functions do not have as consistent a connection with a specific melody. In the case of confirmation requests this could be due to their dual role as questions and statements: speakers are making a statement but simultaneously requesting that their interlocutor confirm that what is said is understood or accepted. Thus, speakers use L* H-H%, L* L-H% or H* H-H% in these instances. Regarding floor holding, one of the most noticeable features was the use of high plateaux, rather than rises per se. Plateaux are particularly prevalent when speakers are listing items or instructions in the map task (cf. [12] on the intonation of lists). Plateaux are possible realizations of high tones [13] and thus they can perhaps create the impression of a rise; however, in our data they were clearly different from uptalk "proper" both acoustically and impressionistically and thus best represented phonologically as L* H-L% where the L% is upstepped.

The patterns described above document the use of tunes that are different from those described for other varieties of English that use uptalk. In particular, the prevalence of L* is not reported for other varieties of English (but see [4] on New Zealand English). As noted, for example, Australian English uses mostly a H* accent and it is precisely this use that has given rise to the term *High Rising Terminal*. Thus, the present study underlies the importance of including dialectal variation in the investigation of intonation and gives support to the claim that such variation exists even within dialectal areas often described as uniform, like the USA West [14].

Regarding the demographic factors in our study, we note that there are consistent differences between genders, with females using uptalk twice as often as males. This difference is presumably what has given rise to the stereotype that uptalk is used by females only; among women uptalk is sufficiently frequent to be identified as a distinctive characteristic of their way of speaking. Contrary to the popular stereotype, no gender differences in the use of uptalk were observed for statements: approximately 16% of statements ended in uptalk in the speech of both men and women in our sample. However, differences are evident in the use of uptalk for floor holding: in this use, uptalk is twice as frequent in the data from female speakers (a result that in itself suggests that the similar frequency of uptalk with statements cannot be attributed to the fact that our male speakers interacted with female researchers). At present we do not have a good explanation for this but offer some suggestions. One possibility is that women wish to hold the floor longer and use uptalk as a device to indicate this intent. This explanation however does not quite tally with existing research suggesting that women do not take longer turns than men ([15] and references therein). Another possibility is that women wish to indicate their intent to hold the floor because they are generally interrupted more often than men [15]. Again, this is not entirely satisfactory as our data were based on monologues (clip retell) and a cooperative task in which the interlocutor was always female. Thus, this aspect of the data clearly requires further investigation. At the same time, we do find that gendered use of uptalk did not interact with task. From this we can infer that the gender effect is not due, e.g., to women asking more questions, but rather to their general preference for certain uses of uptalk.

Unlike gender, which was a clear determiner of the frequency, function and form of uptalk, we did not find differences relating to ethnicity, SEC status or the language background of our speakers. Although it is possible that such differences could emerge with a larger sample, the ubiquitous use of uptalk in our corpus rather suggests that uptalk is sufficiently widespread in SoCal to transcend social barriers. In turn this tallies with the speakers' attitude to uptalk: for SoCal speakers it is not a feature that attracts attention.

5. Acknowledgements

We thank our participants, our research assistants Annabelle Cadang and Andy Hsiu for help with constructing and labeling the corpus, and the members of the UCSD Phonetics Lab for valuable feedback on this project.

6. References

[1] J. Fletcher, E. Grabe, and P. Warren. "Intonational variation in four dialects of English: the high rising tune," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S-A. Jun, Ed. Oxford: Oxford University Press, 2005, pp. 390-409.

[2] C. Mayo, M. Aylett, and D. R. Ladd. "GlaToBI prosodic transcription of Glasgow English: An evaluation study of GlaToBI," in *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications*, 1997, pp. 231-234.

[3] N. Daly and P. Warren, "Pitching it differently in New Zealand English: Speaker sex and intonation patterns," *Journal of Sociolinguistics*, vol. 5, no. 1, pp. 85-96, 2001.

[4] P. Warren, "Patterns of late rising in New Zealand: Intonational variation or intonational change?" *Language Variation and Change*, vol. 17, no. 2, pp. 209-230, 2005.

[5] E. Jarman and A. Cruttenden, "Belfast intonation and the myth of the fall," *Journal of the International Phonetic Association*, vol. 6, no. 1, pp. 4-12, 1976.

[6] J. Sullivan. "The why of Belfast rises," in *New Perspectives on Irish English,* B. Migge and M. Ní Chiosáin, Eds. John Benjamins Publishing Company, 2012, pp. 67–84.

[7] N. E. Adler, E. S. Epel, G. Castellazzo, and J. R. Ickovics, "Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy white women," *Health psychology: official journal of the Division of Health Psychology, American Psychological Association*, vol. 19, no. 6, pp. 586-592, 2000.

[8] A. Singh-Manoux, M. G. Marmot, and N. E. Adler, "Does subjective social status predict health and change in health status better than objective status?" *Psychosomatic Medicine*, vol. 67, no. 6, pp. 855-861, 2001.

[9] P. Boersma and D. Weenik. (2013). *Praat: doing phonetics by computer* [Computer program], Version 5.3.59. Available: http://www.praat.org

[10] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge University Press, 2004.

[11] J. Hirschberg and G. Ward, "The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English," *Journal of Phonetics*, vol. 20, pp. 241-251, 1992.

[12] M. Liberman and J. Pierrehumbert. "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, M. Aronoff and R. Oehrle, Eds. Cambridge, MA: MIT Press, 1984, pp. 157-233.

[13] R-A. Knight and F. Nolan, "The effect of pitch span on intonational plateaux," *Journal of the International Phonetic Association,* vol. 36, no. 1, pp. 21-38, 2006.

[14] W. Labov. "The three dialects of English," in *Handbook of Dialects and Language Variation*, M. D. Linn, Ed. San Diego: Academic Press, 1998, pp. 39-81.

[15] J. Coates, *Men, women and language*. Pearson Education, 2004.

9 Wednesday 2

Rhythmic structure of utterances in native and non-native Polish

Agnieszka Wagner

Institute of Linguistics, Adam Mickiewicz University in Poznan, Poland

wagner@amu.edu.pl

Abstract

This paper presents results of an ongoing study concerning speech rhythm in native and non-native Polish. The goal of the analyses described in the paper was to characterize rhythmically Polish utterances realized by native and non-native speakers with German and Korean accent. The analyses are limited to the domain of duration, but in the future other prosodic parameters will also be investigated. In the current study, different rhythm metrics (%V, ΔV, ΔC, PVIs and Varcos) were applied to provide quantitative description of temporal patterning in native and non-native Polish. Following the assumption that perceived speech rhythm is the effect of meter and grouping which are closely related to prominence and phrasing, durational marking of various levels of prominence and prosodic edges was also analyzed between the three accents (native Polish and German- and Korean-accented Polish). The analyses aimed also at rhythmic classification of Polish – for that purpose the results of quantitative description with rhythm metrics and phonotactic properties of the speech material used in the current study were compared with the data for other languages presented in the literature.

Index Terms: speech rhythm, rhythm metrics, prominence, phrasing, native and non-native Polish

1. Introduction

Speech rhythm can be defined as a systematic temporal organization of prominent and less prominent speech units. In linguistics and phonetics, the notion of speech rhythm is traditionally related to the notion of *isochrony* [1, 2]. Isochrony underlies *rhythm class hypothesis* according to which every language represents a specific rhythm class, i.e. syllable-timed (Spanish, Italian), stress-timed (English, German) or mora-timed (Japanese), on the basis of its temporal organization of syllables. The lack of experimental evidence for stress-, syllable- and mora-based isochrony caused that in rhythm research the focus moved from duration measurements of syllables and feet to investigation of phonetic and phonological factors which affect the timing of syllables and feet, most importantly degree of vowel reduction in unstressed syllables, and phonotactic complexity of syllables. As a results rhythm has been redefined as the perceptual effect of interaction of a number of components: phonetic and phonological on the one hand, and segmental and prosodic on the other [3].

1.1. Rhythm metrics

The differences in vowel reduction, stress-based lengthening and syllable complexity between stress- and syllable-timed languages became the basis of *rhythm metrics* – formulas that measure durational variability in consonantal (C) and vocalic (V) intervals (or syllables). The most widely used metrics include %V-ΔC [4], PVIs [5] and Varcos [6]. Languages with the stress-timed rhythm are expected to exhibit higher nPVI and lower %V and higher consonantal rPVI and ΔC that result from vowel reduction and quite complex syllable structure respectively. Syllable-timed languages generally do not have spectrally reduced and shortened vowels and most syllables have a simple CV structure. Consequently, these languages are expected to exhibit higher %V and lower values of nPVI and the consonantal metrics. There is a growing body of evidence from experimental studies showing that rhythm metrics are not perfect. First of all, they are sensitive to differences in text materials, speech elicitation methods, measurements and changes in speech rate [7, 8]. This sensitivity explains why rhythm metrics take different values in different studies. Second major problem is that metrics reduce rhythm to *timing*, whereas apart from duration, perceived speech rhythm is also the product of properties such as melodic (F0) change and, to a lesser extent, intensity and vowel quality. As a result, different metric scores do not necessarily reflect perceptually different rhythms [9]. Perceived rhythm is the effect of *meter* and *grouping* and therefore research on rhythm should incorporate among others investigation of prominence (which functions as a meter) and phrasing (which is related to grouping, see [10]). Despite criticism, rhythm metrics still constitute the methodological basis of studies concerning topics as diverse as discrimination between rhythm classes [11], acquisition of timing patterns in L1 [12] and L2 [13, 14, 15, 16, 17, 18], detection of speech impairments [19] or dialect discrimination [20].

1.2. Phrasal properties of speech rhythm

It is well known that higher levels of prosodic structure such as prominence marking and prosodic phrasing strongly influence the organization of timing across languages. Analyses based on various languages (including Polish) showed that stressed and pitch accented syllables are produced with additional lengthening compared with unstressed syllables [21, 22, 23, 24]. Final lengthening at the edges of phrasal prosodic constituents is also very widespread [23, 25, 26]. Therefore, it has been acknowledged by many authors (e.g., [10], [27], [28], [29]) that perception of rhythm classes should be examined in relation to prosodic timing phenomena – the durational marking of prominences and phrase boundaries, because they are correlated with the potential rhythm class distinctions and constitute an important ingredient in the rhythm percept across languages.

1.3. Features of Polish, German and Korean rhythm

The speech material used in the current study comes from speakers whose native language was Polish, German or Korean. Polish exhibits phonological properties that are associated with both types of rhythm class [21, 23, 30, 31, 32, 33, 34]. Increased duration of prominent syllables and some compensatory shortening effects together with phonotactic complexity of syllables indicate stress timing. On the contrary, the lack of vowel reduction and a fixed lexical stress on the penultimate syllables indicate syllable timing. Therefore, having some features distinctive of stress-timed languages and others distinctive of syllable-timed languages, Polish would be

predicted to behave rhythmically as an intermediate or mixed language. Yet the acoustic evidence in this respect is contradictory: According to PVIs [5], Polish is close to syllable-timed languages, but, according to %V – ΔC, it is grouped with stress-timed English and Dutch [4].

German is a typical example of a stress-timed language with a phonemic vowel length contrast and vowel reduction in non-prominent unaccented syllables that is manifested acoustically mainly by differences in duration and quality, and allows for longer sequences of consonants in both onset and coda position [35]. This in turn leads to lower %V and higher ΔC, nPVI and rPVI [5, 7].

A perceptual study reported in [36] brought evidence that Korean, like Japanese, has mora-timed rhythm. However, this is at odds with the results of other studies (e.g. [7], [37], [38]) which suggest that Korean has "mixed" rhythm that is closer to syllable-timing than to stress-timing. Thus, at the acoustic level Korean rhythm should be manifested by higher %V and lower ΔC and PVIs than in stress-timed languages.

In the analysis of non-native speech rhythm we expect to find the influence of speakers' L1 on the realization of the durational variability of vocalic and consonantal intervals in Polish. It is assumed that non-native accent should be manifested by values of the metric scores intermediate between those reported for speaker's L1 (German or Korean) on the one hand and those obtained for Polish on the other. The perceptibly weaker non-native accent of German speakers (see sec. 3.1) should be manifested by smaller differences in the metric scores between German-accented and native Polish than between Korean-accented and native Polish.

1.4. Objectives of the current study

The objective of the analyses presented in this paper is to characterize Polish rhythm in utterances realized by native and non-native speakers with German and Korean accent and to provide rhythmic classification of Polish whose rhythmic status is unclear. For this purpose different rhythm metrics will be applied – they will be analyzed in terms of their stability (with respect to speech rate) and robustness (with respect to discrimination between the three accents). The quantitative description of speech rhythm using selected rhythm metrics will be complemented by results of analyses of the phonotactic structure of the utterances and study of durational marking of prosodic heads and phrase edges.

2. Methodology

2.1. Speakers and text

The speech material includes recordings of a literary fairy tale "The teapot" (by H. Ch. Andersen), read by 15 speakers: 5 Polish native speakers, 5 speakers with L1 German and 5 with L1 Korean. The text consists of 19 phonetically and prosodically rich sentences. Recordings of speakers with L1 German come from the *Euronounce* corpus [39], whereas the part representing Korean-accented and native Polish was recorded for the purpose of this study. Recordings were carried out in a sound-treated booth, directly to a disk and with a sampling frequency of 16 kHz. German and Korean speakers represented an intermediate level of proficiency in Polish, but the Koreans were less proficient in pronunciation and took Polish Phonetics course to improve their skills. The non-native speakers were competent, but with a perceivable non-native accent which was more salient in Korean-accented Polish (sec. 3.1). The subjects were asked to read the text once (sentence after sentence), at their own pace. Sentences containing disfluencies or mispronunciations were re-recorded.

2.2. Annotation and measurements

The whole speech material was segmented into vocalic and consonantal intervals on the basis of automatic transcription and segmentation [40] which was verified and manually corrected following standard segmentation criteria. All vowels were marked as vocalic intervals and all consonants (except for post-vocalic glides) – as consonantal intervals. A vocalic interval could contain a single vowel or 2-3 subsequent vowels, or a vowel followed by a glide. Intervals could span across syllable and word boundaries. As in [7] prepausal intervals were not excluded from measurements and segments separated by a pause were treated as two distinct intervals. Despite the applied recording procedure the utterances by non-native speakers are not entirely free of mispronunciations and all insertions and substitutions were included in the C and V intervals. On the one hand, differences in the phonetic realization will affect temporal structure of utterances, but on the other hand, if they are small and few they should not significantly affect metric scores. This issue was investigated using analysis of distribution of various types of syllables between the accent groups (sec. 3.2).

For each sentence the following rhythm metrics have been calculated:

- **%V** – the proportion of vocalic intervals, **ΔV** and **ΔC** – the standard deviation of the duration of vocalic and consonantal intervals respectively [4]
- **rPVI-V** (raw Pairwise Variability Index) and **nPVI-V** (vocalic normalized Pairwise Variability Index): the mean of the duration differences between successive C intervals and the mean of the duration differences between successive V intervals divided by the sum of the same intervals respectively [5]
- **VarcoV/VarcoC**: standard deviation of vocalic/consonantal interval duration divided by mean vocalic/consonantal duration [6]

Prosodic annotation included marking of four levels of prominences and two levels of phrasing. Each syllable and vowel was labeled as unstressed, stressed but unaccented, accented and nuclear accented, and as non-final, final in an intermediate phrase or final in an intonational phrase (see [29], [41], [42]). Syllable boundaries were determined according to criteria presented in [43]. Annotation and duration measurements were done in *Praat*. The data was exported to a spreadsheet where classification into C and V intervals was carried out and values of the rhythm metrics were calculated. For statistical analyses *Statistica 10* was used.

3. Results

3.1. Perception test

The aim of the test was to determine the strength of foreign accent in speech produced by non-native speakers. Seven subjects (students at the Institute of Linguistics) listened to utterances realized by Polish, German and Korean speakers, and marked the strength of foreign accent on a graphical scale from *very strong* to *no accent at all*. The test was carried out using *Annotation System* [44]. The results showed that

utterances realized by speakers with L1 Korean were characterized by strong foreign accent, the accent of speakers with L1 German was assessed as moderate, and in native Polish no foreign accent was perceived. All differences are statistically significant. It can be expected that the strength of foreign accent will be reflected in metric scores and durational marking of prominences and phrase edges.

3.2. Phonotactic properties

In all accent groups **CV** syllables constituted about **50%** of all syllables. The second most common structure was **CCV** (PL: 16.9%, DE: 17.5%, KOR: 16.5%), the third one was **CVC** (PL: 12.2%, DE: 11.6%, KOR: 12.8%) and then **V** (PL: 7.9%, DE: 7.9%, KOR: 8%) and **CCVC** (PL: 7.9%, DE: 7.3%, KOR: 7.1%). Other syllable types – CCCV, VC, CCVCC, CVCC and CCCVC, were much less common (< 3%). Differences in the frequency and distribution of syllable (and interval) types between the three accent groups are below 1%. It can be assumed that mispronunciations that occurred in non-native speech did not affected the overall phonotactic structure of utterances. Comparison of phonotactic properties of native speakers' utterances with data from five languages analyzed in [45] showed that similarly to stress-timed languages, Polish has greater variation in complex syllable structure than rhythmically unclassified Czech (6 types vs. 3 types). But on the other hand, Polish complex syllables most often have the CCV structure (not CVC as in stress-timed languages) and frequency of CCVC syllables is higher than in English and German. These results show that in terms of phonotactic properties Polish, like Czech, does not fit into the binary classification into syllable- and stress-timed languages.

3.3. Quantitative description with rhythm metrics

3.3.1. The effect of speech rate

ANOVA results showed significant differences in the speech rate between native speakers and the two non-native accents ($F=31.2$, $p<0.01$). Speech rate measured in the number of C and V intervals per second (cf. [46]) was higher in native Polish than in German- and Korean-accented Polish (9.2 vs. 8.2 intervals/sec.). Therefore, in the utterances produced by non-native speakers, higher values of the raw consonantal metrics can be expected. In order to select the most stable rhythm metrics correlations between their values and speech rate were investigated (Table 1).

Table 1. *Correlations between metric scores and speech rate (* indicates p<0.05).*

Metric	PL	KOR	DE
%V	-0,01	-0,17	-0,23*
ΔV	-0,33*	-0,59*	-0,64*
ΔC	-0,30*	-0,51*	-0,60*
rPVI	-0,36*	-0,49*	-0,35*
nPVI	-0,07	-0,01	-0,57*
VarcoV	-0,07	-0,10	-0,18
VarcoC	0,13	-0,07	-0,20

It can be seen that with one exception (Varco C, PL) the values of the metrics increase with decreasing speech rate. In some cases (mostly in German-accented speech, DE) the inverse correlations are statistically significant. Varcos, and to lesser degree %V and nPVI, seem to be the most stable metrics. These results indicate need of using rate-normalized metrics instead of the raw ones.

3.3.2. The effect of accent

In order to determine whether there are significant differences in the temporal patterning between native and non-native Polish and between the two non-native accents, ANOVA was carried out with *accent* as predictor variable and *metric scores* as dependent variables (Table 2).

Table 2. *ANOVA results: The effect of accent on the metric scores.*

Rhythm metrics	ANOVA		Scheffe's test
	F test	p	
%V	43,6	0,00	all
ΔV	40,6	0,00	KOR x PL, KOR x DE
ΔC	3,1	0,04	PL x DE
nPVI	77,0	0,00	KOR x PL, KOR x DE
rPVI	0.9	0.4	---
VarcoV	27,2	0,00	KOR x PL, KOR x DE
VarcoC	5,0	0,01	PL x DE

Statistically significant differences in all metric scores except for rPVI can be observed between the three accent groups. The results of ANOVA and post-hoc comparisons (Scheffe's test) indicate that the effect of accent is stronger on vocalic than on consonantal metrics. Taking into account these results and correlations with speech rate (Table 1), it can be assumed that %V, nPVI and VarcoV will be the most robust metrics in discrimination between the three accents.

In native Polish the %V has higher value and nPVI lower value in the current study than in [5]: %V – 47.1 vs. 42.3 and nPVI – 42.1 vs. 46.6. The vocalic metrics place Polish far from stress-timed Dutch, German and English and closer to syllable-timed French and Catalan [5, 7]. Values of the consonantal metrics in the current study show lower variation in the C interval durations comparing to results in [5]: ΔC – 61.3 vs. 71.4 and rPVI – 68.2 vs. 79.1, but they still indicate high variation in structure and duration of C intervals. In a two dimensional space determined by vocalic and consonantal metrics, the latter place Polish far from most languages. These results indicate that Polish rhythm can not be classified using the binary distinction between syllable- and stress-timing.

In German-accented Polish, %V, nPVI and VarcoV have lower values than in native Polish, but the difference is statistically significant only in case of %V (Table 2). In Korean-accented Polish the scores of all the vocalic metrics are higher than in Polish and German: **%V** – 49.7 KOR, 47.1 PL, 45.1 DE, **ΔV** – 61.9 KOR, 45 PL, 45.6 DE, **nPVI** – 54.5 KOR, 42.1 PL, 40.7 DE, **VarcoV** – 51.3 KOR, 44.3 PL, 41 DE. The vocalic metrics take intermediate values between those reported for native Korean [37] and those obtained for native Polish, which indicates *transfer of temporal patterning from L1 Korean to Polish*. In the non-native accents consonantal metrics have higher scores than in native Polish, but these differences might result from different speech rates as indicated by significant inversed correlations (Table 1). Differences in the temporal patterning between native and non-native Polish are reflected mostly by the vocalic metrics: %V, nPVI and VarcoV. The results of quantitative analysis using rhythm metrics are in line with the results of the perception test, because they indicate greater differences in

temporal patterning between Korean-accented and native Polish than between German-accented and native Polish.

3.4. Phrasal properties of speech rhythm

3.4.1. Prominence level

Multivariate ANOVA results showed significant effect of prominence level on syllable duration ($F=354.8$, $p<0.01$) and small, but significant effect of speaker ($F=4.5$, $p<0.01$) and speaker and accent ($F=8.4$, $p<0.01$). Generally, in all accent groups syllable duration increases with the prominence level from unstressed to nuclear accented. The exception is durational marking of stressed syllables (S) in native Polish: these syllables are significantly shorter than all other syllables, including unstressed syllables (U) in Polish, but excluding U syllables in the two non-native accents (Tukey's HSD test). However, even though S syllables are generally longer than U syllables in Korean- and German-accented Polish, the difference is not statistically significant. Contrary to native Polish, in non-native accents the difference in duration between A (accented) and S syllables is not significant. There is also significant difference in durational marking of the highest level of prominence (NA) between Polish and Korean speakers. Similar patterns of variation due to prominence were found in vowel durations, except for S vowel durations in native Polish which this time were longer (but not significantly) comparing to U vowels.

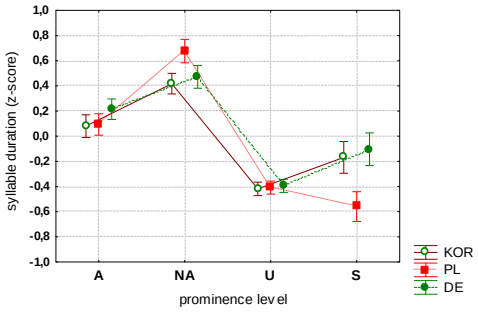


Figure 1: *Mean syllable duration depending on the prominence level (A – accent, NF – nuclear accent, U – no stress & no accent, S – stressed).*

3.4.2. Phrasal position

Similarly to prominence analysis, multivariate ANOVA results showed significant effect of phrasing level on syllable duration ($F=533.3$, $p<0.01$) and small, but significant effect of speaker ($F=15.2$, $p<0.01$), and speaker and accent ($F=19.1$, $p<0.01$). In all accent groups, non-phrase final syllables are significantly shorter than syllables at the edges of intermediate (iip) or intonational phrases (IP). Interestingly, duration of syllables at the edges of intermediate phrases is the longest, but in German-accented and native Polish it is significantly different only from duration of non-phrase final syllables (NF). IPs in native Polish are signaled by significantly increased duration of phrase-final syllables comparing to the non-native accents. Intermediate phrases are marked by significantly longer phrase-final syllables in native and Korean-accented Polish than in German-accented Polish.

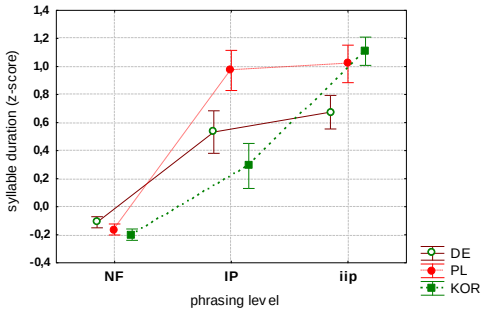


Figure 2: *Mean syllable duration depending on the phrasing level (NF – non-final, IP – intonational phrase, iip – intermediate phrase).*

4. Discussion and conclusions

The results of the analyses indicated significant differences in the rhythmic characteristics between native and non-native Polish. The strength of foreign accent in non-native Polish assessed in a perception study was reflected in more significant differences in the metric scores between native and Korean-accented Polish (strong accent) than between native and German-accented Polish (moderate accent). The results of quantitative analysis using vocalic rhythm metrics showed that Korean speakers transferred some temporal patterns from their L1 to Polish. Generally, Korean-accented Polish is characterized by more vocalic speech (higher %V) and more variation in the duration of V intervals than native Polish and German-accented Polish; the latter accent is characterized by less vocalic speech (lower V%) and less variation in the duration of V intervals than Korean-accented Polish and less variation in C interval duration comparing to native Polish. Generally, %V, nPVI and VarcoV are more stable as regards speech rate differences and more robust than the consonantal metrics. Small, but significant differences in the durational marking of various levels of prominence and phrasing can also be observed between the three accents and they also affect the temporal organization of utterances – this implies the need of including prosodic hierarchy in the analysis of speech rhythm. In the future, the analyses of phrasal properties of Polish rhythm will be complemented by analysis of distribution of prominences and phrasing e.g., position and number of pitch accents and phrase boundaries (see [47]). The phonotactic structure of Polish utterances and the quantitative description using rhythm metrics indicate that in terms of rhythm, Polish does not fit into the binary classification into syllable- and stress-timing. Phonotactically, it is most similar to "rhythmically mixed" Czech, whereas vocalic rhythm metrics provide some evidence of syllable timing and consonantal metrics put Polish aside from both syllable- and stress-timed languages. These results indicate that in order to better characterize Polish rhythm (and speech rhythm in general) it is necessary to go beyond the dimension of timing and to analyze also the contribution of pitch, intensity and tempo to production and perception of rhythmic structure and grouping.

5. Acknowledgements

Work presented in the paper was supported from grant DOBR/0008/R/ID1/2013/03 by the National Centre of Research and Development in Poland.

6. References

[1] Pike, K. L. (1945). The Intonation of American English. Ann Arbor: University of Michigan.

[2] Abercrombie, D. (1967). *Elements of general phonetics* (Vol. 203). Edinburgh: Edinburgh University Press.

[3] Dauer, R. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the XIth International Congress of the Phonetic Sciences*, volume 5, pages 447–450. Tallinn: Academy of Sciences.

[4] Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition 73*(3). 1-28.

[5] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(515-546).

[6] Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for∆ C. *Language and language processing*, 231-241.

[7] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351-373.

[8] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *J. Acoust. Soc. Am 127*(3). 1559-1569.

[9] Barry, W., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, *66*(1-2), 78-94.

[10] Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46-63.

[11] Mairano, P. and Romano, A. (2011). Rhythm metrics for 21 languages. In *Proceedings of the XVIIth International Congress of Phonetic Sciences*, 17-21 August 2011, Hong Kong, China. 1318-1321.

[12] Payne, E., Post, B., Astruc, L., Prieto, P. & del Mar Vanrell, M. (2012). Measuring child rhythm. *Language and Speech, 55*(2), 203-229.

[13] White, L. & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics 3*(5). 501-522.

[14] Mok, P. P., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Proceedings of the Speech Prosody 2008 Conference* (pp. 423-426).

[15] Tortel, A., & Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. In *Proceedings of Speech Prosody 2010*.

[16] Ordin, M., Polyanskaya, L. & Ulbrich, Ch. (2011). Acquisition of Timing Patterns in Second Language. In *Proceedings of INTERSPEECH 2011*, 27-31 August 2011, Florence, Italy. 1129-1132.

[17] Kinoshita, N., & Sheppard, C. (2011). Validating acoustic measures of speech rhythm for second language acquisition. In *Proceedings of the XIth International Congress of the Phonetic Sciences* (Vol. 17, pp. 1086-1089).

[18] Li, A. and Post, B. (2012). L2 rhythm development by Mandarin Chinese learners of English. In *Proceedings of Perspectives on Rhythm and Timing*, 19-21 July 2012, Glasgow, UK.

[19] Liss, J., White, L., Mattys, S., Lansford, K., Lotto, A., Spitzer, S. & Caviness, J. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research* Vol.52 1334-1352.

[20] Leemann, A., Dellwo, V., Kolly, M. J., & Schmid, S. (2012). Rhythmic variability in Swiss German dialects. In *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China.

[21] Jassem, W. (1962). Akcent języka polskiego (Accent of Polish). Wrocław: Ossolineum.

[22] Turk, A.E. & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics 27*, 171–206.

[23] Demenko, G. (1999). Analysis of Polish Suprasegmentals for needs of Speech Technology. UAM: Poznań.

[24] Wagner, A. (2009.) Analysis and recognition of accentual patterns. In *Proceedings of INTERSPEECH 2009*, 6-10 Sept., Brighton, UK.

[25] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *J. Acoust. Soc. Am. 91*, 1707–1717.

[26] Wagner, A. (2010). Acoustic cues for automatic determination of phrasing. In *Proceedings of Speech Prosody 2010*, 11-14 May 2010, Chicago, USA.

[27] Beckman, M. E. (1992). Evidence for speech rhythms across languages. *Speech perception, production and linguistic structure*, 457-463.

[28] Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, French, and English. *Journal of phonetics 19*(3-4), Jul-Oct 1991, 351-365.

[29] Prieto, P., Vanrell, M. D. M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication*, *54*(6), 681-702.

[30] Richter, L. (1987). Modelling of the rhythmic structure of utterances in Polish. *Studia Phonetica Posnaniensia 1*, 91–125.

[31] Klessa, K. (2006). Analysis of segmental duration for needs of speech synthesis in Polish. Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[32] Wagner, A. (2008). Comprehensive model of intonation for application in speech synthesis. Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[33] Malisz, Z., & Klessa, K. (2008). A preliminary study of temporal adaptation in Polish VC groups. In *Proceedings of the Speech Prosody 2008 Conference*, Campinas, Brazil.

[34] Malisz, Z. (2013). Speech rhythm variability in Polish and English: A study of interaction between rhythmic levels. Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[35] Gut, U. (2003). Non-native speech rhythm in German. In *Proceedings of the ICPhS conference* (pp. 2437-2440).

[36] Moon-Hwan, C. (2004). Rhythm typology of Korean speech. *Cognitive Processing*, *5*(4), 249-253.

[37] Mok, P., & Lee, S. I. (2008). Korean speech rhythm using rhythmic measures. In *Proc. of the 18th Intern. Congress of Linguists* (CIL18), Seoul, Korea.

[38] Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and speech*, *51*(4), 343-359. Cylwik et al. 2009

[39] Cylwik, N., Wagner, A., & Demenko, G. (2009). The EURONOUNCE corpus of non-native Polish for ASR-based Pronunciation Tutoring System. In *Proceedings of SLaTE*, Wroxall Abbey Estate, Warwickshire.

[40] Demenko G., Wypych M. & Baranowska E. (2003.) Implementation of Polish grapheme-tophoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. *Speech and Language Technology 7*:79-96

[41] Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. *Papers in laboratory phonology III: Phonological structure and phonetic form*, 7-33.

[42] Selkirk, E. O. (1995). Sentence prosody: Intonation, stress and phrasing. In Goldsmith, J., editor, Handbook of Phonological Theory. Oxford.

[43] Ostaszewska, D., & Tambor, J. (2000). Phonetics and phonology of contemporary Polish. Polish Scientific Publishers (PWN).

[44] Klessa, K., Karpiński, M. & Wagner, A. (2013). Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features [in:] Brigitte Bigi and Daniel Hirst (Eds.), *Proceedings of TRASP*. Aix-en-Provence, 30 August, 2013.

[45] Dellwo, V. (2008). The influence of speech rate on speech rhythm. Unpublished PhD thesis, Universität Bonn

[46] Beňuš, Š. & Šimko, J. (2012). Rhythm and tempo in Slovak. In *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China.

[47] Rosenberg, A., & Hirschberg, J. (2010). Production of English prominence by native mandarin Chinese speakers. In *Workshop on Prosodic Prominence: Perceptual, Automatic Identification.*

Long-term convergence of speech rhythm in L1 and L2 English

Hugo Quené[1], *Rosemary Orr*[12]

[1]Utrecht inst of Linguistics OTS, Utrecht University, Utrecht, the Netherlands
[2]University College Utrecht, Utrecht, the Netherlands
h.quene@uu.nl, r.orr@uu.nl

Abstract

When talkers from various language backgrounds use L2 English as a lingua franca, their accents of English are expected to converge, and talkers' rhythmical patterns are predicted to converge too. Prosodic convergence was studied among talkers who lived in a community where L2 English is used predominantly. Speech rhythm was operationalized here as the peak frequency in the spectrum of the intensity envelope, normalized to the speaking rate (in syll/s). Results indicate that talkers produced intensity contours with maximum periodicity at frequencies of about 0.32 times their syllable rates, i.e., peaks in intensity tend to occur every 1/0.32 syllables. These results were collected repeatedly, from 5 recordings conducted over 3 years with the same talkers. We found that variance between talkers in their rhythm decreases over time, thus confirming the predicted convergence in speech rhythm in L2 English. These findings show that speech rhythm in L2 English tends to converge, and that this prosodic convergence continues to proceed over several years, as well as over communicative settings.
Index Terms: speech rhythm; phonetic convergence; accommodation; L2;

1. Introduction

When talkers from different varieties or dialects of a language converse with each other, their dialects and accents tend to converge. This convergence has been reported on various time scales for varieties of L1 Dutch [11], L1 English [7, 5] and L1 French [4]. But does convergence also occur among varieties of an L2 language used as a lingua franca by talkers from various L1 backgrounds? It is widely assumed that a talker's accent will weaken or disappear in intensive contact with other talkers of L1 and L2 English. However, we have only limited insight in how and why L2 accents change over time in highly interactive environments. Do non-native speakers become more native-like, and does interference from their L1 decrease over time? And do native speakers also drift away from their native pronunciation standards? The international community of the University College Utrecht (UCU) provides an excellent environment to study phonetic convergence in L2: its students vary in their native languages (with Dutch being the majority L1), English is spoken as the lingua franca on campus, and the UCU is located in a Dutch-speaking country, so that any changes in English accents can be validly attributed to internal processes within the community.

The core hypothesis in this longitudinal research project is that the native and non-native accents of UCU students will gradually converge to a single common UCU English accent, in which properties of L1 British English, L1 American English, L2 Dutch English, and other varieties will be mixed. With regard to speech rhythm, we hypothesize that talkers' rhythmical patterns will converge. A key property of English is its lexical stress, which typically results in one stressed syllable of a word being more prominent (produced with more effort and longer duration) than the other unstressed ones. A salient rhythmical feature of English is that the unstressed syllables tend to be reduced more dramatically (both spectrally and temporally) than they are in other lexical-stress languages such as Dutch or German, or in non-stress languages such as Japanese.

In the present longitudinal study, we focus on long-term convergence in speech rhythm in L1 and L2 English, over a period of 3 years. Do L2 English speakers become more native-like, in that they will show more reduction of unstressed syllables? Or do L1 English speakers become less native-like, in that they will show less reduction of unstressed syllables? Does the variance between talkers in their preferred rhythmical pattern decrease over the years? These questions will be answered by means of the longitudinal speech corpus described below.

2. Corpus

The Longitudinal University College English Accents Corpus (LUCEA) is a corpus of speech recordings, collected at University College Utrecht (UCU) in Utrecht, the Netherlands. Four consecutive annual cohorts of students were or will be recorded, each at five "rounds" or time points over the three years of their undergraduate study, in longitudinal fashion. Talkers are recorded in Sept of year 1 of their programme (month 1), May of year 1 (month 9), Sept of year 2 (month 13), May of year 2 (month 21), and May of year 3 (month 33; most UCU students are absent on exchange visits in the first semester of year 3). Actual recording dates may differ by a few weeks from the nominal month of recording, due to students' availability. Recordings began in September 2010, and are still in progress.

2.1. Talkers

The talkers in the corpus are mostly full-time degree students, with some incoming exchange students and staff included who are native speakers of English. About 60% of the degree students and of the talkers in the corpus have Dutch as their L1. In addition to the Dutch speakers and the native speakers of English, over 30 other native languages are represented in the corpus. UCU students are required to use English in all their interaction with tutors and teachers. Students live on campus during their entire three-year degree programme, and in culturally and linguistically mixed social settings, English is also the language of choice. Thus the talkers use L1 or L2 English very intensively for interacting in the campus community. Of the student population of about 700, approximately one quarter (about 60 per cohort) participates as a talker contributing speech data to the LUCEA corpus.

The present study targets those talkers for which a full sequence of 5 recordings is available ($n = 18$); all these talkers were students of the first cohort (class of 2013) participating in the LUCEA corpus collection. In the first and last recording sessions, talkers also filled in a questionnaire about their language background, language use, study exchange experience, etc. In the entry questionnaire, 15 talkers declared themselves as native speakers of Dutch, and 1 talker each as a native speaker of Russian, Vietnamese, and German. In the exit questionnaire, 3 talkers (1 female, 2 male) also regarded themselves as L1 English speakers. These answers need not be contradictory, as many students at UCU have a complex language history with multiple native languages (e.g. different languages of mother, father, and country of residence).

2.2. Procedure

Each student talker receives €10 for each recording, with a bonus of €10 if all 5 rounds of recordings are completed. Recordings take place in a quiet office room on the UCU campus. The talker's speech is recorded by means of 7 microphones (Sennheiser ME64/K6p) placed around, behind, and above the talker, and also via a close-talking microphone (Sennheiser Headset HSP 2ew). All 8 channels are recorded digitally by means of a Saffire Pro 40 multichannel preamp and AD converter (at 44.1 kHz, 16 bits).

Talkers are asked to perform between 10 and 12 speech tasks, listed in Table 1, including texts to read aloud, monologues in both the native language and English (if English is not the native language), and a dialogue with the facilitator. Each recording session takes about 45 minutes, in which approximately 25 minutes of speech is collected.

Table 1: *Summary of tasks for talkers to be performed in each recording.*

Nr	Description
1	Say your name, and today's date and time
2	Short extract from the Rainbow Passage
3	*Please Call Stella* [13]
4	*The Boy who Cried 'Wolf'*†
5	Sentence sets for intelligibility testing
6	Five sentences for investigating prosody [14]
7	Extract from Declaration of Human Rights in L1 [12]
8	Extract from Declaration of Human Rights in English [12]
9	2 minute monologue in L1 (informal topic)
10	2 minute monologue in English (informal topic)
11	2 minute monologue in English (formal topic)
12	3 minute dialogue with facilitator

†Two different versions of this text have been used.

3. Method

The present study focused on a small part of the recordings, viz. the 5 English sentences taken from and studied by [14] (Table 1, task 6; see Appendix A). These sentences may be regarded as a semi-random sample in their distribution of stressed and unstressed syllables. Jointly the 5 sentences contain 26 stressed syllables out of a total of 81 syllables, a ratio of 0.321.

Talkers were always instructed to read each sentence without pausing or inhaling in mid-sentence, and without any errors. If necessary, a talker repeated a sentence until this was achieved. Only fluent realizations without error and without pausing were selected for subsequent analysis. Each sentence was excised and stored as a separate audio file. Incomplete recordings (due to errors in post-processing) were discarded, with $N = 423$ recordings by 18 talkers remaining.

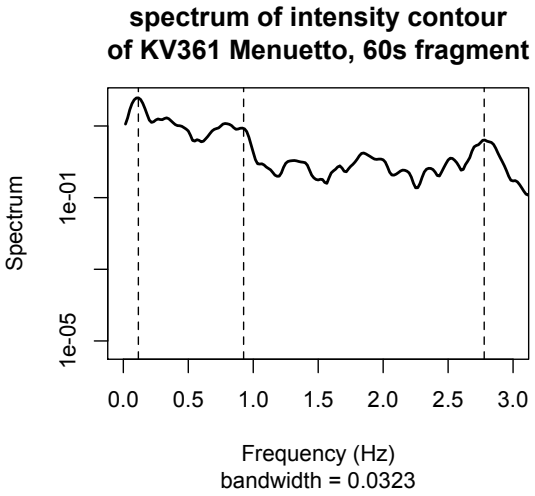


Figure 1: *Spectrum of intensity contour, with marking of salient frequencies.*

The intensity envelope of each sentence was measured by means of Praat [2], using a 25 ms window shift (40 Hz sampling frequency). The resulting intensity contour (in dB units) was exported. Subsequent analysis of the intensity envelope was inspired by [6] and was performed using R [10]. The intensity contour was converted to a spectrum, in order to bring out its periodic components. As an example, consider the spectrum of the intensity envelope of a musical fragment (by W.A. Mozart, KV361, Menuetto, in $3/4$ time) shown in Fig. 1. The peak at 2.77 Hz corresponds with the quarter notes (at a tempo of about 167 beats per minute, or 2.77 per s), and the peak at 0.92 Hz corresponds with the full measures (0.92 per s). The strongest peak at 0.116 Hz corresponds with an 8-measure or 24-beat periodicity of the intensity envelope.

From each spectrum, we then identified the frequency bin with the maximum intensity, i.e., the strongest periodic component of the intensity envelope. The windows used during preceding intensity analysis and spectral analysis and smoothing resulted in an eventual frequency resolution of 0.016 Hz (spacing between frequency bins).

The duration of each sentence was assessed from the duration of the intensity envelope, excluding silent intervals before and after the sentence. The articulation rate (sc. tempo, in syll/s) of each spoken sentence was computed using the syllable counts in Appendix A. (Note that sentences were always spoken without pauses.) The observed peak frequency in the spectrum of the intensity envelope was then normalized to this articulation rate. Thus the strong spectral component that corresponds to the syllable repetition frequency is removed, and the resulting measure represents the periodicity of the intensity envelope, expressed *relative* to the syllable-based periodicity of the intensity envelope. For the spectrum of Fig. 1 (with a base tempo of 2.77 Hz) this would yield a single dimensionless value of $0.116/2.77 = 1/24$, indicating a periodic or cyclical pattern in the intensity envelope that spans over 24 quarter-notes. The

relative strength of this normalized peak frequency was not assessed. Data from 2 sentence recordings were excluded from further analysis because of their unusually high articulation rate (> 8 syll/s), with $N = 421$ spoken sentences remaining.

The normalized peak frequencies of the intensity envelope of each spoken sentence were fed into a linear mixed-effects regression model (LMM) [8, 9], with talkers ($n = 18$) and sentences ($n = 5$) as two crossed random effects, using maximum likelihood estimation. LMM was done using package `lme4` [1] in `R` [10]. Fixed predictors were (i) the "round" or time of recording (coded as 4 contrasts between consecutive rounds), (ii) the talker's status as a native speaker of English (0=no, 1=yes), (iii) the interaction of predictors (i) and (ii), and (iv) the articulation rate (in syll/s, centered to its median). The rounds were also included in the random part at the talker level, i.e., we explicitly modeled the between-talker variance for each round separately.

4. Results and discussion

The coefficients and variances estimated by the LMM described above are listed in Table 2. The estimated normalized peak frequencies in the spectrum of the intensity envelope are also illustrated in Fig. 2.

Table 2: *Estimated coefficients of the LMM. For fixed effects,* r2r1 *refers to the contrast between round 2 and round 1, etc.;* native *indicates the talker's status as native speaker of English (yes=1); colons are used for interaction terms; asterisks mark fixed effects with* $p < .05$ *(based on 2.5% and 97.5% percentiles of bootstrapped estimates over 200 bootstrap replications). For random effects,* u *refers to talkers and* v *to sentences, and* e *to residual error; the 95% confidence intervals are based on percentiles of bootstrapped estimates over 200 replications.*

Fixed effects:	Estimate	Std.Error	t value
(Intercept)	0.311	0.036	8.674
tempo	−0.039	0.015	−2.609 *
r2r1	0.018	0.051	0.035
r3r2	0.032	0.059	0.548
r4r3	0.032	0.047	0.669
r5r4	0.037	0.036	1.025
native	0.082	0.029	2.822 *
r2r1:native	−0.004	0.122	−0.029
r3r2:native	−0.125	0.140	−0.892
r4r3:native	−0.195	0.112	−1.739 *
r5r4:native	−0.167	0.086	−1.939 *
Random effects:	Estimate	95%C.I.	
σ^2_{u1}	0.0082	(0.0020, 0.0167)	($n = 18$)
σ^2_{u2}	0.0048	(0.0009, 0.0107)	
σ^2_{u3}	0.0024	(0.0003, 0.0070)	
σ^2_{u4}	0.0024	(0.0005, 0.0076)	
σ^2_{u5}	0.0015	(0.0003, 0.0058)	
σ^2_{v}	0.0057	(0.0002, 0.0136)	($n = 5$)
σ^2_{e}	0.0165	(0.0137, 0.0182)	($n = 421$)

In the present sample of spoken English sentences, the intensity envelope shows a major periodicity at $0.31\times$ the syllable rate. This value matches with the proportion of stressed syllables (0.321) in the test sentences, suggesting that the peak frequency in the intensity envelope, or "normalized stress rate", corresponds with the rhythm of the spoken sentences. In English, unstressed syllables tend to be weaker in intensity (and shorter in duration) than stressed syllables, and this property is indeed captured by the intensity envelope showing periodicity at a rate corresponding with the average inter-stress interval. Thus this stress rate [6], here normalized relative to syllable rate, may be of interest for evaluating speech rhythm.

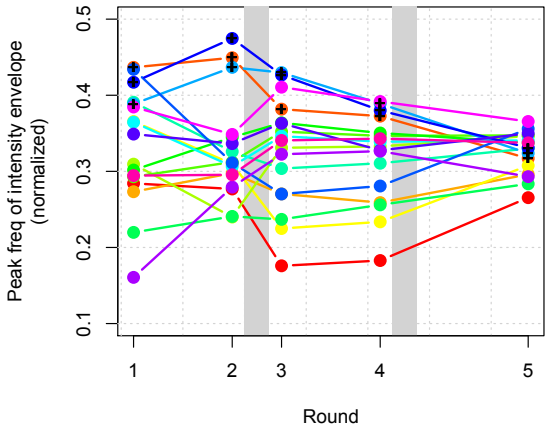


Figure 2: *Estimates of normalized peak frequencies in the spectrum of intensity envelope, broken down by round of recording (along abscissa, on approximate time scale) and by talker (with plussed symbols representing L1 English speakers). Shaded areas represent 2-month summer breaks during which talkers do not live on the UCU campus.*

Indeed, the normalized "stress rate" or peak frequency of the intensity envelope is significantly higher for the 3 native English talkers than for the 15 nonnative talkers in the present sample, as shown in Table 2. In our interpretation, this significant difference may be ascribed to the stronger reduction of unstressed syllables in English as compared to Dutch. This difference in reduction has been reported to extend to L1 English and L2 English spoken by Dutch native speakers [14, 3]. In our sample, 15 of the 18 talkers mentioned Dutch as one of their native languages. If these L2 talkers would reduce their unstressed vowels in their L2 English to a lesser degree than their native counterparts, then the intensity peaks of their stressed and unstressed syllables would also differ to a lesser degree than their native counterparts would produce. This would in turn result in a somewhat *lower* "normalized stress rate" for L2 talkers as compared to what native speakers would produce. For example, the native English speakers all pronounce *chairman* as [ˈtʃɛɚmən], whereas the native Dutch speakers of L2 English tend to pronounce this word as [ˈtʃɛɚˌmæn] with a less reduced second syllable, yielding a higher intensity peak for the unstressed syllable, and hence a lower "normalized stress rate". Thus the significant main effect of native-speaker status on normalized stress rate reflects this relatively subtle prosodic difference between native and Dutch-accented English in degree of syllable reduction [14, 3].

The significant interaction between native-status and recording round indicates that the 15 nonnative talkers do *not* change over time in their normalized stress rate, whereas the 3 native English talkers tend to converge to the somewhat lower, nonnative stress rate observed for the nonnatives. In other

words, the native speakers of English (who are a minority at UCU) tend to adapt their English speech rhythm to that of their nonnative peers (who are a majority), especially at the 4th and 5th recording session (end of year 2 and end of year 3, respectively). The convergence seems to extend into the students' third year on campus; this finding matches similar reports of long-term phonetic accommodation [11, 5]. Moreover, the phonetic accommodation is extended from conversational settings with peer students, and from classroom situations, to the interview setting of the corpus recordings. These findings supports the core hypothesis of long-term prosodic convergence among students in the UCU community.

The core hypothesis of phonetic convergence also predicts that the variance between talkers decreases over time. LMM allows this prediction to be tested, by modeling separate $\sigma^2_{u.}$ between-talker variances for each round of recordings. The resulting variance estimates in Table 2 seem to confirm this prediction to some extent (cf. Fig. 2), but the 95% confidence intervals of these variance estimates do overlap in bootstrap validation. The LMM reported in Table 2 also does not perform significantly better than a simpler model in which between-talker variances are pooled into a single estimate for all 5 rounds (i.e. in which homoskedasticity is assumed; likelihood ratio test, $\chi^2 = 7.62, \mathrm{df} = 14$, n.s.). Thus the decreasing variance between talkers is adequately captured by the significant interaction between native status and recording round (in the fixed part of the LMM, [9]), and there is no significant deviation from homoskedasticity beyond this interaction.

In conclusion, L2 English talkers do not show longitudinal changes in their rhythmical pattern. The native L1 English talkers however tend to move away from their native rhythmical patterns (observed initially), by decreasing the degree of reduction of unstressed syllables; hence they accommodate to the predominant variety of L2 English in the language community. These longitudinal changes confirm that members of this multilingual community, where English is used as the lingua franca, do converge in their speech rhythm of their L1 and L2 English accents.

5. Acknowledgements

We thank the UCU talkers for lending their voices, Roeland van Beek, Thari Diefenbach, Anne van Leeuwen, Lisa Teunissen, Kate Backhouse, Maria Koutiva and Kim Cruden for their assistance in conducting the recording sessions, and David van Leeuwen for technical assistance.

6. References

[1] Bates, D., Maechler, M., Bolker, B. and Walker, S. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5, 2013. Online: http://CRAN.R-project.org/package=lme4

[2] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer, version 5.3.51, 2013. Online: http://www.praat.org

[3] Braun, B., Lemhöfer, K. and Mani, N. Perceiving unstressed vowels in foreign-accented English, J. Acoust. Soc. Am., 129(1):376–387, 2011.

[4] Delvaux, V., and Soquet, A. "The influence of ambient speech on adult speech productions through unintentional imitation", Phonetica, 64(2–3):145–173, 2007.

[5] Evans, B.G. and Iverson, P., "Plasticity in vowel perception and production: A study of accent change in young adults", J. Acoust. Soc. Am., 121(6):3814–3826, 2007.

[6] Liberman, M.Y., "Speech rhythms and brain rhythms", Language Log, 2 Dec 2013.
Online: http://languagelog.ldc.upenn.edu/nll/?p=8116

[7] Pardo, J.S., "On phonetic convergence during conversational interaction", J. Acoust. Soc. Am., 119(4):2382–2393, 2006.

[8] Quené, H. and van den Bergh, H., "On Multi-Level Modeling of data from repeated measures designs: A tutorial", Speech Comm., 43(1–2):103–121, 2004.

[9] Quené, H. and Van den Bergh, H., "Examples of mixed-effects modeling with crossed random effects and with binomial data", J. Memory and Language, 59(4):413–425, 2008.

[10] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, version 3.0.2, 2013. Online: http://www.R-project.org/.

[11] Scholtmeijer, H. *Het Nederlands van de IJsselmeerpolders*. Kampen, 1992.

[12] Van Engen, K. J., M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow. "The Wildcat Corpus of Native- and Foreign-Accented English: Communicative efficiency across conversational dyads with varying language alignment profiles." Language & Speech, 53(4):510–540, 2010.

[13] Weinberger, S. Speech Accent Archive. George Mason University, 2013. Online: http://accent.gmu.edu

[14] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", J. Phonetics, 35(4):501-522, 2007.

A. Test sentences

Each test sentence (from [14]) is followed by its numbers of stressed and unstressed syllables (in typical readings in the present corpus) and its total number of syllables.

1 The supermarket chain shut down because of poor management (5, 10, 15). 2 Much more money must be donated to make this department succeed (6, 11, 17). 3 In this famous coffee shop they serve the best doughnuts in town (5, 10, 15). 4 The chairman decided to pave over the shopping center garden (5, 12, 17). 5 The standards committee met this afternoon in an open meeting (5, 12, 17).

Probing Theories of Speech Timing using Optimization Modeling

Andreas Windmann[1], *Juraj Šimko*[2], *Petra Wagner*[1]

[1]Faculty for Linguistics and Literary Studies, Bielefeld University, Germany
[2]Institute of Behavioural Sciences, University of Helsinki, Finland
[1]firstname.lastname@uni-bielefeld.de, [2]juraj.simko@helsinki.fi

Abstract

We implement two theories about the temporal organization of speech in an optimization-based model of speech timing and conduct simulation experiments in order to test whether both theories can account for the phenomenon of foot-level shortening (FLS) observed in English speech corpora. Results suggest that a model that induces compensatory timing relations between syllables and feet predicts empirical results very accurately. However, we also observe that the FLS effect can equally well be explained under the assumption that suprasegmental timing is confined to localized lengthening effects at the heads and edges of prosodic domains. Implications for theories of speech timing are discussed.

Index Terms: Speech timing, computational modeling

1. Introduction

In this paper, we shall test predictions made by two theories of suprasegmental speech timing. To this end, we will employ a computational optimization model [1], by implementing both theories in the model and evaluating modeling results against attested speech timing patterns of English. Results allow for comparing the empirical adequacy of different predictions and demonstrate the potential of our model as a test bed for different theories of suprasegmental speech timing.

The phenomenon under study is polysyllabic shortening, i.e. the property of syllables to shorten as a function of the number of syllables in some larger prosodic unit. We will concentrate on foot-level shortening (FLS), an alleged shortening effect triggered by the interval from a lexically stressed syllable onset to the next, possibly spanning word boundaries [2]. While not leading to true periodicity of stressed syllable onsets, FLS in English does seem to be well-supported by both experimental studies [3, 4, 5] and corpus analyses [6, 7, 8, 9]. Figure 1 shows two examples of FLS patterns found in English speech corpora, as evident from vowel rather than syllable durations. There is a marked shortening effect on stressed vowel durations which is, however, not linear but seems to be attenuated as syllable count in the foot increases. Similar patterns have been observed in experimental studies and could be interpreted as an effect of durations moving towards a compressibility threshold [12]. Figure 1 does not indicate FLS in unstressed syllables, which has been linked to incompressibility as well [13]. However, [10] do report consistent shortening effects of various prosodic constituents also on unstressed vowel durations.

Findings on FLS seem to support theories in which prosodic timing effects are distributed over larger prosodic domains, leading to inverse relationships between the number of syllables assembled in these domains and the durations of those syllables [14, 2]. This has been explicitly formalized in a class of computational models which assume that the temporal organization of speech is governed by oscillatory mechanisms at different levels of the prosodic hierarchy [15, 16, 17, 18, 9]. On this account, surface speech timing patterns emerge as a result of an interaction, or *coupling*, between the different oscillators, as a result of which they entrain to a stable frequency pattern. FLS would be explained to arise from hierarchical coupling between a dominant oscillator at the inter-stress interval or foot level and a more compliant syllabic oscillator, thus generating a tendency for stressed syllables to reoccur at regular intervals.

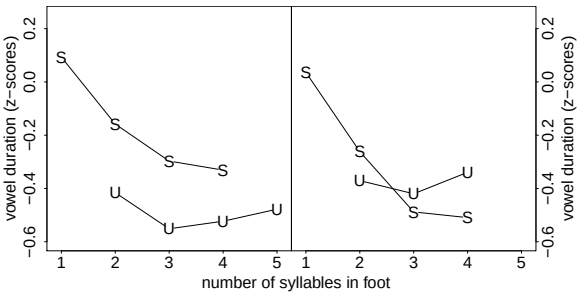


Figure 1: *Foot-level shortening effect on mean stressed (S) and unstressed (U) vowel/syllable durations in English. Graphs are reproduced from numerical results for two speakers from [8].*

In contrast to this, [19] and [20] propose a theory in which suprasegmental timing mechanisms are confined to localized lengthening at the heads and edges of prosodic domains. In this theory, there is little or no role for "domain-span effects", a term used by these authors to denote precisely the kind of inverse timing relationships that oscillatory approaches would predict. [19] and [20] base their claims on experimental findings suggesting that purported shortening effects at the word level in English can be largely accounted for by combinations of localized lengthening effects, such as accentual lengthening and word-final lengthening. However, [20] concede that their model may not necessarily account for FLS effects, which makes it a promising prospect to test for foot-level effects explicitly.

In this paper, we shall investigate whether both theories can account for FLS effects observed in English speech corpora. This will be done by implementing the mechanisms proposed by both theories, informally referred to as the "distributed" vs. the "localized" timing account, in our optimization-based model of speech timing. Predictions made by both theories will then be tested on input data derived from an authentic speech corpus and compared to published results. The paper is structured as follows: in section 2, we describe the general architecture of our model and the additions implemented in order to

accommodate the distributed and the localized timing account. Results of simulation experiments are presented in section 3 and discussed in section 4. Section 5 concludes the paper and addresses perspectives for further work.

2. Model Architecture

Our model is inspired by Hypo & Hyper-articulation (H&H) theory [21], implementing the assumption that speech patterns emerge from the resolution of conflicting demands related to minimization of effort and maximization of communicative success. It derives from an embodied optimization model of articulatory timing [22, 23]. The current model operates on specifications of sequences of lexically stressed and unstressed syllables, representing speech utterances. Given an input sequence, an optimization algorithm computes the vector S of syllable durations that minimizes the composite cost function C. C is a weighted sum of component functions that represent production and perception-related influences on constituent durations.

The basic architecture of the model includes three such components, D_S, T and P_S. D_S and T implement constraints associated with efficiency of information transmission. The durational cost component T captures the overall duration of the utterance, i.e., the time used for conveying the message encoded in the sentence of a part thereof. This component provides a control mechanism for overall speech tempo. D_S is proportional to individual syllable durations, based on the assumption that the syllable is a basic unit of information which speakers strive to transmit in an efficient manner. Motivation for having both T and D_S in the model comes from evidence that different mechanisms may be responsible for changes of local durations and overall speaking rate, cf. [24] and references therein.

Conversely, component P_S represents an impetus to maximize perceptual clarity, by imposing costs on the reciprocal of syllable durations. P_S thus decreases with syllable duration in a convex decaying fashion, assuming that very short durations impede perception while facilitation induced by durational lengthening will eventually reach a ceiling. Independent evidence for this modeling decision comes from gating studies, where subjects have to identify phonemes from acoustic syllable fragments of varying duration [25, 26].

Weighting factors allow for globally imposing premiums on the individual components in order to simulate requirements regarding efficiency (α_D), perceptual clarity (α_P) or overall speaking rate (α_T). The vectors δ_S and ψ_S, assigning weights to individual syllables, can be used to boost their relative perceptual clarity and simultaneously lower the premium on efficient information transmission. This mechanism is used to account for prosodic prominence in the model, assuming that speakers prioritize clarity over efficiency in prominent syllables [27]. We usually set δ_S to the reciprocal of ψ_S, in order to reduce the number of free parameters. Formally, the basic model is thus defined as

$$C = \alpha_D \sum_S \delta_S D_S + \alpha_P \sum_S \psi_S P_S + \alpha_T T \qquad (1)$$

In order to accommodate the distributed timing account, we implemented a version of the basic model with two additional component costs, D_F and P_F, together with respective weighting factors α_{DF} and α_{PF}. D_F and P_F are basically copies of D_S and P_S operating at the stress foot rather than the syllabic level. Thus, the two new components in combination impose a tendency to produce stress feet with a certain optimal duration, while D_S and P_S tend to converge to optimal syllable durations. This leads to an obvious trade-off and will trigger compensatory relationships for feet with different syllable counts. By setting the respective weighting factors, precedence can be given to either the foot or the syllabic level, simulating purported tendencies towards "stress timing" and "syllable timing", respectively. This architecture is conceptually very similar to the coupled oscillator models mentioned above. Independent motivation for such a design might be derived from findings on convergence of syllable and foot durations with certain temporal windows of cognitive processing [28]. Figure 2 visualizes the architecture of the distributed timing model.

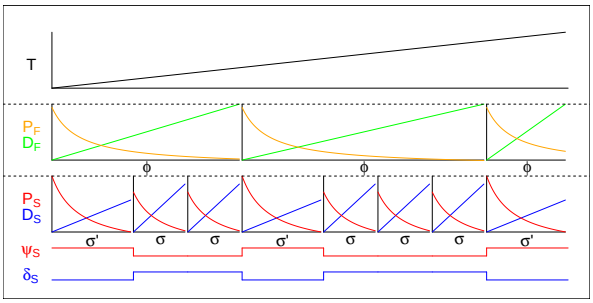


Figure 2: *Distributed timing model. Cost functions T (utterance level), D_F/P_F (stress foot level; ϕ) and D_S/P_S (syllabic level; σ; ' denotes lexical stress) as well as parameters δ_S and ψ_S are plotted as a function of respective constituent durations for a hypothetical utterance consisting of a trisyllabic, a tetrasyllabic and a monosyllabic foot. α_{DF} and α_{PF} are not shown.*

For the localized timing account, no additional cost functions had to be supplied. Its predictions were implemented by additionally enhancing ψ_S and thus boosting the perception-oriented component P for word-final syllables, leaving the model definition in (1) unchanged otherwise. This approach is in keeping with [19]'s reasoning that localized lengthening is utilized to increase the perceptual salience at important points in the speech signal, in this case word boundaries. There are thus no "domain-span" mechanisms in this version of the model (note that T does not induce compression as a function of utterance length if it is linear); only localized lengthening effects at the heads (stressed syllables) and edges (word-final syllables) of words are included. No attempt was made in either of the two versions of the model to incorporate utterance-final lengthening, since utterance-final syllables have usually been excluded in investigations of FLS. Effects of syllable structure and pitch accent were also ignored for the present purpose. Figure 3 visualizes the architecture of the localized timing model.

3. Simulations

Input data for the simulation experiments were derived from the Aix-MARSEC database, a corpus of English broadcast speech [29, 30, 31]. That is, we prepared input "utterances" for the model that were based on actual utterances from the corpus in terms of number of syllables and locations of lexical stress and word boundaries. The Aix-MARSEC database is ideally suited for this purpose because FLS and shortening effects in other domains have been documented in this corpus [6, 10].

Both versions of the model were implemented in R using the built-in optimization function *optim*. In order to keep computing time within reasonable limits, input data were restricted to 2000 utterances from the corpus, amounting to 7512 sylla-

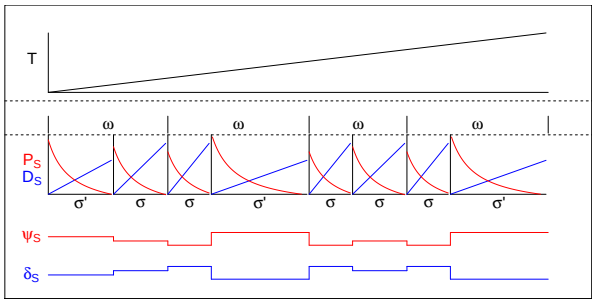


Figure 3: *Localized timing model. The utterance is the same as in Figure 2. The middle panel shows word boundaries (ω). Note the differentiation for word-final vs. non-final syllables in ψ_S and δ_S (in addition to stressed/unstressed differentiation).*

bles. Utterance-initial exametrical syllables, i.e., unstressed syllables not contained in a foot, were excluded. For the simulation with the distributed timing model, we set α_D and α_P to 0.5 and α_{DF} and α_{PF} to 1 in order to simulate the hypothesized dominant timing influence of the foot. ψ_S was set to 2 for lexically stressed and 1 for unstressed syllables, and δ_S was set to $1/\psi_S$, as explained above. All other parameters were set to 1. Simulation results from the distributed timing model are shown in Figure 4.

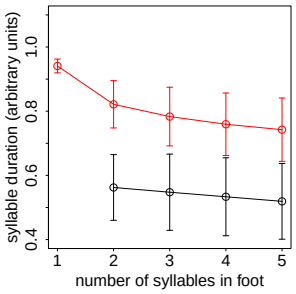


Figure 4: *Results from distributed timing model simulation for stressed (red) and unstressed (black) syllables.*

The distributed timing model reproduces the general pattern shown in Figure 1 remarkably well, particularly for the stressed syllables, which exhibit a marked asymptotic shortening tendency. The model does predict a consistent shortening effect in unstressed syllables as well, but it is quite weak compared to the effect in stressed syllables. Given that many other sources of durational variation are not taken into account in this version of the model, it may be assumed that the shortening effect on unstressed syllable durations is likely to vanish if simulations are carried out with a more full-blown model architecture. The fact that syllables shorten as a function of foot length is in itself of course rather trivial, given that the parameter settings impose temporal compression at the foot level. What is non-trivial about this result, however, is that the model (1) captures the asymptotic nature of the shortening effect in stressed syllables and (2) reproduces the finding of a weaker effect in unstressed syllables. We believe that both outcomes are indeed a consequence of incompressibility, which has been shown to emerge automatically from the architecture of our model [1].

The localized timing model was run on the same input data as the distributed timing model. In this simulation, ψ_S was increased by 0.5 for word-final syllables in order to simulate word-final lengthening. All other parameter settings were the same as in the distributed timing model simulation. Monosyllables were counted as final. Results are shown in Figure 5. Surprisingly, the localized timing model also captures the pattern of results shown in Figure 1 quite well, with asymptotic shortening in stressed and a weaker effect in unstressed syllables. The magnitude of the effect is somewhat smaller than in the distributed timing model simulation, but this of course depends on the exact numerical setting of the parameter values. The important result is that the localized timing model can account for the overall pattern of results. This is a striking finding, given that no explicit timing mechanism at the foot level is included in this version of the model.

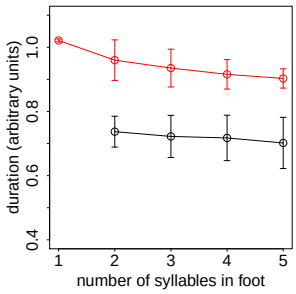


Figure 5: *Results from localized timing model simulation for stressed (red) and unstressed (black) syllables.*

On this account, the explanation for the FLS effect would be an entirely different one: rather than a genuine tendency towards temporal compression at the foot level, the phenomenon would be a mere statistical artifact, arising from an apparent tendency for word-final syllables to occur in shorter feet. An analysis of the whole MARSEC corpus was conducted in order to substantiate this correlation. We computed the probability of a syllable being in word-final position as a function of syllable count in the respective foot, by dividing the number of word-final syllables occurring in feet of a given length by the total syllable count for the respective foot length in the corpus. This was done separately for stressed and unstressed syllables. Results are shown in Figure 6. As can be seen, the probability of a syllable occurring in word-final position indeed decreases as a function of syllable count in the foot, in a fashion which is strikingly similar to the durational effect.

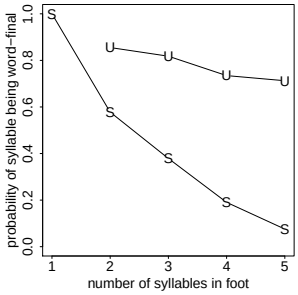


Figure 6: *Probability of syllables occurring in word-final position as a function of foot length in the MARSEC corpus.*

Upon closer inspection, this pattern of results is actually hardly surprising: if a stressed syllable is directly followed by

another stressed syllable and thus constitutes a monosyllabic foot, there is necessarily a word boundary intervening, since by definition, a word cannot contain more than one primary stressed syllable. The probability of being word-final therefore has to be one for monosyllabic stress feet. For bisyllabic feet, the probability of the stressed syllable being word-final is much lower – any bisyllabic word with initial stress followed by any word with initial stress will make for a non-word-final observation here – but there are still two frequent patterns that will generate final observations, 1) words with final stress followed by bisyllabic words with final stress, as in "[STAY a]WAKE", and 2) sequences of a word with final stress, a weak function word and a word with initial stress, as in "[JOHN the] BAPtist" (brackets mark target foot boundaries). For successively longer feet, the frequency of patterns that allow for word-final stressed syllables decreases – it is hard to imagine a pattern with a word-final stressed syllable followed by four unstressed syllables.

For unstressed syllables, the probability of occurring in word-final position decreases with foot length as well, but the effect is much weaker than in stressed syllables. A possible explanation is that longer feet are likely to involve polysyllabic words. In this case, some of the unstressed syllables in a long foot will be word-initial or medial, resulting in a weaker correlation between foot length and the probability to occur word-finally for unstressed syllables. This would explain the weaker or absent effect of foot length on unstressed syllables in the durational domain under a localized timing account.

4. Discussion

Our results show that both the distributed and, interestingly, also the localized account of speech timing can reproduce FLS patterns observed in English speech corpora. In the distributed timing account, the effect would be explained as a tendency towards periodicity of stressed syllable onsets, resulting from a trade-off between realizing optimal syllable and foot durations. Under the localized timing account, FLS would be classified as a mere by-product of language structure, i.e., the stronger tendency for syllables in shorter feet to occur word-finally, where they are subject to a localized lengthening effect.

If the distributed timing theory is correct, our implementation provides a promising explanatory platform for the FLS effect. Crucially, the model not only produces longer syllables in shorter feet, but also captures the precise nature of the shortening effect intriguingly well. We would therefore argue that our distributed timing model offers a more satisfactory account of the phenomenon than the oscillatory approach described in [17], who have to introduce ad-hoc assumptions in order to replicate the difference between stressed and unstressed syllables reported by [7]. [17] effectively "switch off" FLS in unstressed syllables, but it may well be that the effect was just masked by noise from other durational processes in [7], in keeping with our model's predictions. Indeed, [8] reports on the same data that unstressed durations do exhibit a significant shortening effect once the number of *phones*, rather than syllables per foot is used as the independent variable. Since [17] only compare bi- and trisyllabic feet, it is also not clear if their oscillatory model captures the non-linear nature of the shortening in stressed syllables. In our distributed timing model, both the weaker effect in unstressed syllables and the attenuation of the shortening effect in stressed syllables emerge automatically from the independently motivated property of durational incompressibility.

Results of the localized timing model simulations, however, show that it may be entirely unnecessary to invoke a timing mechanism at the foot level in order to reproduce the empirical results. This would be in keeping with [19]'s claim that "domain-span effects" have been falsely attributed to English due to ignorance of confounding influences such as final or accentual lengthening. Word-final lengthening in particular seems to be a well-attested effect in English [32, 33], although it is not entirely uncontroversial [34]. Of course, it cannot be decided based on our simulation results whether word-final lengthening is the trigger of a spurious FLS effect, or if, on the contrary, word-final syllables are longer than non-final ones *because* they tend to occur in shorter feet.

Thus, while our simulation results show that both accounts can generate the observed pattern, they do not allow for falsifying either theory. The predictions of the localized timing model converge with results by [19] and [20], as well as [33], who reports that in his large-scale corpus study, an apparent FLS effect on vowel durations disappears once a vowel's distance to the right word boundary is controlled. On the other hand, it seems that the localized timing model cannot fully account for *experimental* findings on FLS. For example, [5] report that a word-final stressed syllable is longer in a monosyllabic than in a bisyllabic foot, which is also acknowledged by [19] and [20]. There may be alternative analyses, such as a kind of stress clash effect here, however.

We have not tested for shortening effects in domains other than the stress foot. Results from [19, 20] and [33] are compatible with a domain-span effect at the *word rhyme* level, the unit that stretches from the onset of a stressed syllable to the next word boundary, alternatively referred to as *narrow rhythm unit* [35, 10]. However, [20] raise the possibility that this may in fact be a progressive word-final lengthening effect. Further empirical study is necessary in order to decide on these issues.

5. Conclusions

Using optimization modeling, we have shown that a model architecture that imposes distributed timing mechanisms can well account for patterns of foot-level shortening observed in English speech corpora. However, the effect may equally well be explained in a model of speech timing that only includes localized lengthening effects due to a tendency for shorter feet to contain mostly word-final syllables.

From a methodological point of view, results of our simulation experiments confirm that our optimization-based model provides a promising test bed for different theories of speech timing. The model itself is of course not theory-neutral, but it appears that the H&H assumptions it is based on are sufficiently general to accommodate other theories. Detailed studies on empirical data, including proper control for possible sources of durational variation, are required in order to determine which model architecture correctly captures the suprasegmental organization of English speech timing.

6. Acknowledgements

We would like to thank the present and former staff at the Laboratoire Parole et Langage at the University of Aix-Marseille, in particular Cyril Auran, Caroline Bouzon and Daniel Hirst, for making the MARSEC corpus publicly available. Further thanks go to three anonymous reviewers for helpful comments on an earlier draft of this paper. The first author gratefully acknowledges the Bielefeld graduate school of linguistics and literary studies for financial support.

7. References

[1] A. Windmann, J. Šimko, B. Wrede, and P. Wagner, "Modeling durational incompressibility," in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 1375–1379.

[2] D. Abercrombie, *Elements of general phonetics*. Edinburgh: Edinburgh University Press, 1967.

[3] T. P. Barnwell, "An algorithm for segment durations in a reading machine context." DTIC Document, Tech. Rep., 1971.

[4] M. Fourakis and C. Monahan, "Effects of metrical foot structure on syllable timing," *Language and Speech*, vol. 31, no. 3, pp. 283–306, 1988.

[5] B. Rakerd, W. Sennett, and C. Fowler, "Domain-final lengthening and foot-level shortening in spoken English," *Phonetica*, vol. 44, no. 3, p. 147, 1987.

[6] N. Campbell, "Foot-level shortening in the Spoken English Corpus," in *Proceedings of the 7th FASE Symposium*, Edinburgh, 1988, pp. 489–494.

[7] H. Kim and J. Cole, "The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase," in *Proceedings of Interspeech 2005*, Lisbon, 2005, pp. 2365–2368.

[8] H. Kim, "Speech rhythm in American English: A corpus study," Ph.D. dissertation, University of Illinois, 2006.

[9] J. Krivokapić, "Rhythm and convergence between speakers of American and Indian English," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013.

[10] C. Bouzon and D. Hirst, "Isochrony and prosodic structure in British English," in *Speech Prosody 2004, International Conference*, 2004.

[11] S. Shattuck-Hufnagel and A. Turk, "Durational evidence for word-based vs. prominence-based constituent structure in limerick speech," in *Proceedings of ICPhS 2011*, Hong Kong, 2011.

[12] D. Klatt, "Interaction between two factors that influence vowel duration," *The Journal of the Acoustical Society of America*, vol. 54, no. 4, pp. 1102–1104, 1973.

[13] C. Hoequist, "Durational correlates of linguistic rhythm categories," *Phonetica*, vol. 40, no. 1, pp. 19–31, 1983.

[14] K. Pike, *The Intonation of American English.* Ann Arbor: University of Michigan Press, 1945.

[15] M. O'Dell and T. Nieminen, "Coupled oscillator model of speech rhythm," in *Proceedings of ICPhS 1999*, San Francisco, 1999, pp. 1075–1078.

[16] P. Barbosa, "From syntax to acoustic duration: A dynamical model of speech rhythm production," *Speech Communication*, vol. 49, no. 9, pp. 725–742, 2007.

[17] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008, pp. 175–184.

[18] S. Tilsen, "Multitimescale dynamical interactions between speech rhythm and gesture," *Cognitive Science*, vol. 33, no. 5, pp. 839–879, 2009.

[19] L. White, "English speech timing: a domain and locus approach," Ph.D. dissertation, University of Edinburgh, 2002.

[20] L. White and A. E. Turk, "English words on the procrustean bed: Polysyllabic shortening reconsidered," *Journal of Phonetics*, vol. 38, no. 3, pp. 459–471, 2010.

[21] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech production and speech modeling*, W. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer, 1990, pp. 403–439.

[22] J. Šimko and F. Cummins, "Embodied task dynamics." *Psychological review*, vol. 117, no. 4, pp. 1229–1246, 2010.

[23] J. Šimko and F. Cummins, "Sequencing and optimization within an embodied task dynamic model," *Cognitive Science*, vol. 35, no. 3, pp. 527–562, 2011.

[24] K. S. Harris, "Vowel duration change and its underlying physiological mechanisms," *Language and Speech*, vol. 21, no. 4, pp. 354–361, 1978.

[25] W. Grimm, "Perception of segments of English-spoken consonant-vowel syllables," *The Journal of the Acoustical Society of America*, vol. 40, no. 6, pp. 1454–1461, 1966.

[26] M. Tekieli and W. Cullinan, "The perception of temporally segmented vowels and consonant-vowel syllables," *Journal of Speech, Language and Hearing Research*, vol. 22, no. 1, p. 103, 1979.

[27] K. J. De Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *The journal of the acoustical society of America*, vol. 97, no. 1, pp. 491–504, 1995.

[28] P. Wagner, *The rhythm of language and speech: Constraints, models, metrics and applications.*, Habilitation thesis, University of Bonn, 2008.

[29] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "Marsec: A machine-readable spoken English corpus," *Journal of the International Phonetic Association*, vol. 23, no. 1, pp. 47–53, 1993.

[30] G. Knowles, B. B. J. Williams, and L. Taylor, *A corpus of formal British English speech*. Longman, 1996.

[31] C. Auran, C. Bouzon, and D. Hirst, "The aix-marsec project: an evolutive database of spoken British English," in *Speech Prosody 2004, International Conference*, 2004.

[32] D. H. Klatt, "Vowel lengthening is syntactically determined in a connected discourse." *Journal of phonetics*, vol. 3, no. 3, pp. 129–140, 1975.

[33] J. P. Van Santen, "Contextual effects on vowel duration," *Speech Communication*, vol. 11, no. 6, pp. 513–546, 1992.

[34] A. E. Turk and S. Shattuck-Hufnagel, "Word-boundary-related duration patterns in English," *Journal of Phonetics*, vol. 28, no. 4, pp. 397–440, 2000.

[35] W. Jassem, *Intonation of Conversational English (educated Southern British)*. Nakl. Wroclawskiego Tow. Naukowego; skl. gl.: Dom Ksiazki, 1952, no. 45.

The influence of accentuation and onset complexity on gestural timing within syllables

Sandra Peters, Felicitas Kleber

Institute of Phonetics & Speech Processing, Ludwig-Maximilians-University, Munich, Germany
{sandra|kleber}@phonetik.uni-muenchen.de

Abstract

This paper presents results from a production experiment using electromagnetic articulography. The main aim of the study was to investigate how phrasal accent and the number of onset consonants influence the gestural timing of syllable constituents in German. Five speakers of German with sensors attached to the tongue tip, tongue body and lower lip were recorded reading sentences with either accented or unaccented target words that contained simplex (one consonant) and complex (two consonants) onsets. The nucleus was always /a/ and the coda consonant was always /p/. We analyzed acoustic segment duration and gestural overlap (in terms of lag measurements). Onset complexity influenced both CV and VC overlap and accentuation affected gestural overlap to a greater extent than acoustic vowel duration. However, the extent of overlap differed between segment sequences and accentuation patterns: while for CC and VC sequences trends for greater overlap in deaccented than in accented condition were found, CV overlap decreased with deaccentuation. Shorter plateau durations in this context explain the diminished CV overlap in a prosodically weak context. The findings are discussed with respect to the predictions made by articulatory phonology regarding gestural timing and with respect to timing stability in weak versus strong prosodic contexts.

Index Terms: Phrasal accentuation, incremental shortening, Articulatory Phonology, EMA

1. Introduction

This study addresses the challenge of investigating the timing of articulatory gestures between all syllable constituents (i.e. onset, nucleus, and coda) in two prosodic conditions and drawing the connection between gestural coordination and variation in acoustic duration.

Segments tend to be shorter in prosodically weaker contexts (i.e. unstressed syllables, deaccented words, at low prosodic boundaries) than in prosodically stronger contexts (i.e. lexically stressed syllables, accented words, at high prosodic boundaries). In particular, vowels were shown to differ in duration with respect to lexical stress and phrasal accent [1]. For consonants, lengthening has been found in the vicinity of high prosodic boundaries [2, 3]. Most studies showing prosody-dependent duration differences are based on acoustic measurements. Acoustic shortening, however, may not always be due to a reduction of the articulatory gestures, but instead to a greater degree of overlap between two gestures that occur in prosodically weak contexts [e.g. 3]. In addition, prosodic reduction effects may also be seen in spatial instead of temporal reduction. In particular, [4] showed that lax as opposed to tense vowels were spatially but not temporally reduced and thus not acoustically shortened.

Segmental duration is not only influenced by prosody but also depends on a couple of other factors. For example, vowels are shorter before obstruents than before sonorants. The duration of the vowel, which usually constitutes the syllable's nucleus, is also affected by the number of the adjacent consonants: vowels flanked by consonant clusters (henceforth complex) tend to be shorter than vowels flanked by single consonants (henceforth simplex). This phenomenon – commonly known as incremental compensatory shortening [5] – seems to be both language specific and dependent on syllable position. According to studies conducted in the framework of Articulatory Phonology (e.g. [6, 7]) vowels are shortened only by preceding but not by following clusters. This prediction arises from the assumption that onset clusters are globally timed, i.e. consonants are organized around a stable midpoint of the cluster – the c-center – causing vowel remote consonants to move away from the nucleus and vowel adjacent consonants to move towards the vocalic nucleus resulting in an increased overlap between the consonant and the following vowel (henceforth CV overlap) [8]. Coda clusters, on the other hand, are assumed to be locally timed, i.e. consonants are in sequential order, therefore not causing any increase in overlap between the vowel and the following consonant (henceforth VC overlap) and vowel shortening. But [9] found a trend towards more vowel shortening before coda clusters in sonorant, accented tokens as opposed to other contexts as well as an influence of accentuation on acoustic vowel duration. In addition, [10] found that onset manner had an influence on VC(C) timing, i.e. the vowel adjacent coda consonant shifted towards the vowel if the onset consonant was a lateral but not if the onset consonant was a nasal.

Previous studies using electropalatography (EPG) and electromagnetic (midsagittal) articulography (EM(M)A) found that prosodic strength affects duration and overlap of consonantal onset clusters. For example, [2] and [3] both found that overall cluster gestures in prosodically weaker contexts were shortened. With respect to overlap, [3] found that consonantal overlap decreased in stressed position and at higher prosodic boundaries. [11], however, reported more overlap when the target words were uttered in focus position than when produced in unaccented position. [11] explained their findings with a shortening of the gesture's plateau in prosodically weaker contexts (instead of indirect shortening due to more overlap).

The aim of this study was to further investigate the interplay between accentuation and articulatory timing patterns with respect to all syllable constituents, i.e. coda and onset consonants (C) as well as the vocalic nucleus (V). We extend our study to prosodic effects on CC, CV and VC overlap and plateau shortening. In particular we are interested in the effects of accentuation and cluster coordination on vowel shortening and VC organization. Three hypotheses were tested:

H1 There is more CV overlap in words with complex onsets than in words with simplex onsets.

H2 There is more overlap in the deaccented condition compared to the accented condition concerning all syllable constituents, i.e. CC, CV and VC sequences.

H3 If there is, however, less overlap in deaccented tokens, than plateau durations should be shortened.

2. Method

2.1. Speech Material and experimental set-up

The test items were non-existent monosyllabic words (representing monomorphemic words) that contained the lax vowel /a/. The vowel was preceded by either simplex (/n/) or complex (/kn/) onsets. The coda was kept simplex containing the labial stop /p/. All test words were produced in sentence medial position in the carrier phrase *"Melanie's Omi* [target word] *imitiert ein Lied." ("Melanie's grandma* [target word] *imitates a song.")* with only one intermediate (and thus one intonation) phrase. This test sentence was chosen as it allows for optimal tracking of the articulators. Each test word indicated the name of Melanie's grandma. The test sentence contained one pitch accent that was either on *Melanie's* (i.e. the target word was deaccented) or on the target word. In order to elicit the test words in accented and deaccented position, we first familiarized the participants with two questions that are each associated with one prosodic pattern. *"**Wessen** Omi* [target word] *imitiert ein Lied?"("**Whose** grandma* [target word] *imitates a song?")* was used to evoke the pitch accent on *Melanie's* in the deaccented condition. The question "***Welche** Omi von Melanie imitiert ein Lied?"* (*"**Which** grandma of Melanie imitates a song?"*) was used to trigger a pitch accent on the target word in the accented condition. In addition, during the recordings, blue coloured letters drew attention to the words that were to be produced with a pitch accent. The experimenter ensured that subjects corrected any incorrect prosodic patterns.

Articulatory recordings were made in a sound attenuated booth at the Institute of Phonetics in Munich. The movement of speech articulators was tracked using 3D Electromagnetic Articulography (EMA, AG 501) [12]. In total, twelve EMA receivers were attached to various parts of each speaker's head. The sensors included in the current analysis were those attached to the tongue tip (TT), tongue back (TB) and to the upper and lower lips (LA). Seven repetitions of each sentence were presented isolated in randomized order on a computer screen.

2.2. Participants

Five speakers (3 females, 2 males) of Southern Standard German aged between 19 and 25 were recorded. None of them reported any hearing or speaking disorders. Three out of the five participants were undergraduate students of phonetics, but they were naïve as to the purpose of the experiment. One participant was the first author of this paper.

2.3. Data analysis

The acoustic data was automatically segmented and labeled using MAuS [13]. Segment boundaries were hand-corrected using praat [14] where necessary and the target word's accentuation pattern (i.e. accented vs. deaccented) was manually labeled. For the present study, we only analyzed temporal articulatory measures, i.e. specific moments in time of vertical and horizontal movements of the relevant sensors. The physiological data was labeled using Emu [15]. The target words' components /k/ and /a/ labels were set based on the vertical movements of the tongue back. For /p/, the lip aperture was calculated as the Euclidean distance between upper and lower lips. For /n/ the tangential velocity of the tongue tip was used. The plateau's onsets and offsets were defined on the basis of changes in the articulators' velocity, which are interpolated values and represent the 20% threshold of the difference between two adjacent maxima in the velocity signal. Prior analyses have shown that vowels vary acoustically depending on the presence or absence of a pitch accent on the target word and also on the number of consonants within a cluster (see e.g. [9, 16]). Because of this (potential) variability, we did not use normalization methods usually applied in c-center studies as they normalize on anchor points using either the following coda consonant [7] or the acoustic vowel midpoint (e.g. [17]). Instead we used the normalization method described in [18]. Following this method, we first determined the lag between two neighboring segments and then normalized the lag on the entire duration of the two gestures. For measuring CC_{lag}, the first consonant's ($C1_{on}$) plateau offset (P_{off}) was subtracted by the second consonant's ($C2_{on}$) plateau onset (P_{on}) and divided by the entire gesture (G) as described in the following equation. Using the start and end of the gesture's plateau was found to be the most stable timing measure, i.e. the one with the lowest variation coefficient [19].

$$CC_{lag} = \frac{P_{on}[C2_{on}] - P_{off}[C1_{on}]}{G_{off}[C2_{on}] - G_{on}[C1_{on}]} \quad (1)$$

The same procedure was applied to the CV_{lag} and VC_{lag} with the exception of subtracting the consonant's ($C2_{on}$) plateau offset by the vowel's (V) plateau onset and the vowel's plateau offset by the coda consonant's ($C1_{off}$) plateau onset (cf. equations (2) and (3)), respectively. Thus, lower lag values indicate more overlap.

$$CV_{lag} = \frac{P_{on}[V] - P_{off}[C2_{on}]}{G_{off}[V] - G_{on}[C2_{on}]} \quad (2)$$

$$VC_{lag} = \frac{P_{on}[C1_{off}] - P_{off}[V]}{G_{off}[C1_{off}] - G_{on}[V]} \quad (3)$$

Plateau durations were calculated by subtracting the plateau onset by the plateau offset of each gesture. Figure 1 schematically displays the specific landmarks used for analysis.

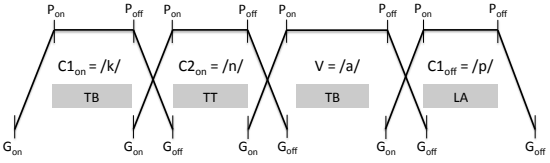


Figure 1: *Schematic illustration of gestures and landmarks used in the present analysis.*

For the statistical analyses we conducted repeated measures ANOVAs. CC_{lag}, CV_{lag}, VC_{lag} and plateau durations of all syllable constituents each served as the dependent variable. ONSET COMPLEXITY (simplex vs. complex) and ACCENTUATION (accented vs. deaccented) were the independent variables. Speaker was entered as a random factor.

3. Results

3.1. Acoustic vowel durations

Prior to the articulatory analyses, we tested whether onset clusters and different prosodic patterns acoustically influenced lax vowels. The results show that, especially in the accented condition, there was vowel shortening due to complex onsets (mean = 66.9ms) compared to simplex onsets (mean = 76.3ms). There was no vowel shortening due to deaccentuation. In spite of this result, there may be articulatory shortening (see [4] for discussion).

3.2. CC_{lag} and plateau durations

In order to account for the influence of ACCENTUATION on the temporal organization of onset clusters and plateau durations, we analyzed the CC_{lag} between /k/ and /n/ in accented and deaccented words (i.e. only words containing complex onsets were used).

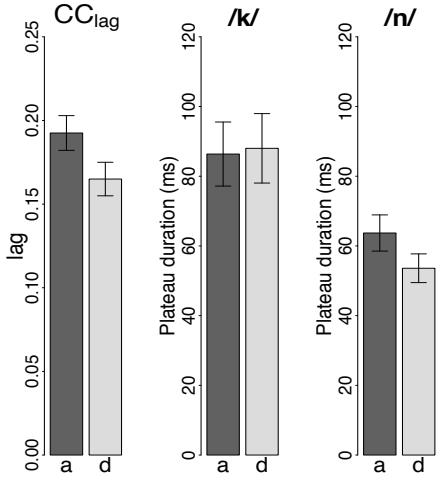


Figure 2: *CC_{lag} (left) between /k/ and /n/, plateau durations for /k/ (middle) and /n/ (right) in accented (a; dark grey) vs. deaccented (d; light grey) tokens.*

ACCENTUATION did not significantly affect the timing of onset clusters. However, a trend towards more overlap in terms of lower lag values can be observed in the deaccented condition in the left plot of Figure 2. This means that there was a tendency towards more overlap in deaccented than in accented words.

Concerning plateau duration of /k/ and /n/, there was no significant effect of ACCENTUATION. While the middle plot of Figure 2 shows that plateau durations of /k/ were about the same in accented and deaccented condition, the right plot of Figure 2 indicates tendencies for shorter plateau durations in deaccented words although, again, this was not statistically significant.

3.3. CV_{lag} and plateau durations

In order to investigate the influence of ONSET COMPLEXITY and ACCENTUATION both on the timing between the vowel adjacent onset consonant ($C2_{on}$) and the vowel and on plateau durations, we analyzed words containing both simplex (/n/) and complex (/kn/) onsets.

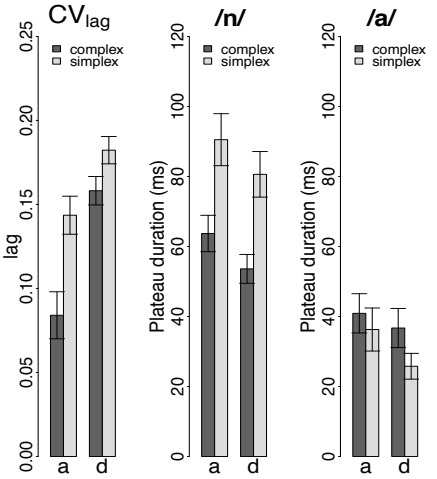


Figure 3: *CV_{lag} (left) between /n/ and /a/, plateau durations for /n/ (middle) and /a/ (right) in accented (a; dark grey) vs. deaccented (d; light grey) tokens.*

ACCENTUATION had a significant ($F[1,4]=18.6$, $p<0.05$) influence on the timing of $C2_{on}$ and the following vowel. There were higher lag values in the deaccented condition (cf. Figure 3.). This means, that, in general, there was less overlap in deaccented words compared to accented words.

The independent variable ONSET COMPLEXITY had no significant effect, although, again, a trend towards more overlap in terms of lower lag values can be observed in complex onsets compared to their simplex counterparts (cf. Figure 3), indicating a shift towards the vowel. This trend was much more pronounced in the accented than in the deaccented condition. That is, although there was no significant interaction, the effect of ACCENTUATION, tended to be greater in words containing complex onsets as opposed to simplex onsets (cf. left plot of Figure 3).

Concerning plateau duration of /n/, there was a significant influence of ONSET COMPLEXITY ($F[1,4]=12.3$, $p<0.05$). The plateau duration was shortened in complex onsets compared to simplex onsets in both prosodic conditions. There was no effect of ACCENTUATION.

For /a/, however, only the independent variable ACCENTUATION significantly influenced the plateau duration ($F[1,4]=67$, $p<0.01$). That is, the vowel's plateau was shortened in deaccented words although there was no shortening effect due to deaccentuation in the acoustics. ONSET COMPLEXITY did not effect the nucleus' plateau duration.

3.4. VC_{lag} and plateau durations

In order to investigate the influence of ONSET COMPLEXITY and ACCENTUATION both on the timing between the vowel and the following consonant ($C1_{off}$) and on plateau durations, again, we differentiated between simplex (/n/) and complex (/kn/) onsets.

Concerning VC sequences, the results have shown that neither ONSET COMPLEXITY nor ACCENTUATION had a significant influence on the timing between the vowel and $C1_{off}$. However, again, a trend towards more overlap in terms of lower lag values in the deaccented condition can be observed (cf. left plot of Figure 4). That is, there was a

tendency towards more overlap in words produced without a pitch accent compared to words produced with a pitch accent.

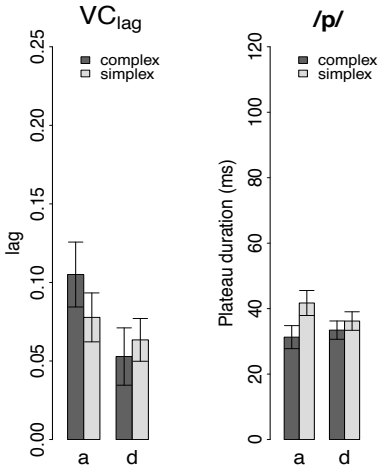


Figure 4: *VC_{lag} (left) between /a/ and /p/, plateau durations for /p/ (right) in accented (a; dark grey) vs. deaccented (d; light grey) tokens.*

The vowel plateau duration was significantly influenced by ACCENTUATION (cf. right plot of Figure 3), as described in the section above. Concerning /p/, there was a significant influence of ONSET COMPLEXITY ($F[1,4]=60.8$, $p<0.01$). This means that the coda's plateau was shortened as the number of onset consonants increased.

4. Discussion and Conclusion

The results of the present study are somewhat tentative as only some of the main and none of the interaction effects reached significance. This may be due to the general weak temporal compression effects in lax vowels (see for example [4]). In addition, there were only five speakers producing seven repetitions of four tokens (accented vs. deaccented; simplex vs. complex), i.e. the statistical power may have been too low. Nevertheless, clear trends showed that accentuation differently affects the gestural timing within a syllable depending on onset complexity. In particular, accentuation significantly influenced CV overlap and the vowel's plateau duration whereas onset complexity had a significant effect on the plateau duration of the vowel-adjacent consonants.

Concerning CC sequences, there was a trend towards more overlap in deaccented words. This finding partially supports hypothesis H1 and is in line with [2] and [3]. Despite this trend we found tendencies towards shorter plateau durations of $C2_{on}$ in the deaccented condition, although [2] found gestural lengthening of this consonant in lexically stressed position. Thus, contrary to the predictions made in hypothesis H3, plateau shortening may also co-occur with more overlap.

There were two findings regarding CV sequences: first, consonants preceding vowels are shifted into the vowel in complex onsets confirming hypothesis H1. This result supports the predictions made by Articulatory Phonology [8]. The second one is that the CV lag is greater in deaccented than in accented tokens, which can be explained by shorter plateau durations of both, /n/ and /a/. This result contradicts hypothesis H2 and confirms hypothesis H3 and thus is in line with findings described in [11].

In addition, we found that the cluster dependent shift tends to be greater in accented than in deaccented tokens. This may be explained by additional lengthening of the plateau duration in accented as opposed to deaccented tokens, which may then result in a greater shift into the vowel. This is not to say that articulatory lengthening in principal results in more overlap. Articulatory strengthening should cause greater coarticulatory resistance [20] resulting in less overlap (e.g. there was a tendency towards less overlap in accented CC sequences). In this particular case of incremental onset shortening, however, the shift towards the vowel seems to be stronger in the accented condition, because temporal overlap between longer plateau gestures (due to accentuation) is reached faster. This means that the predictions made by Articulatory Phonology depend not only on syllable position (i.e. onset vs. coda clusters) or the cluster's composition (cf. [16]), but also on prosodic structures.

Concerning VC sequences, there was a trend towards more overlap in deaccented words, which is again in line with [2] and [3] and partially supports hypothesis H1. Despite this trend, we found shorter plateau durations of the vowel in the deaccented condition. This again indicates that shorter plateau durations can also co-occur with increased overlap. In addition, the onset complexity influenced the VC timing. There was more CV overlap but less VC overlap in the accented condition. Moreover, the onset composition also affected the plateau duration of the coda consonant, i.e. there were shorter plateau durations of $C1_{off}$ as the number of onset consonants increases. This indicates that onsets and codas do not behave independently of each other within a syllable (see also [21]) and supports the findings presented in [10].

This study allows no direct comparison of plateau shortening between segments of different articulators, since the tongue tip is flexible resulting in short gestures and long plateau durations for nasals, whereas the tongue back moves slowly resulting in long vowel gestures and short plateau durations for vowels (cf. middle and right plot of Figure 3). For this reason, the results concerning plateau durations have to be extended by also taking into account gesture durations.

In summary, two major conclusions arise from this study. The first one is that lax vowels may be articulatorily compressed due to deaccentuation without showing acoustic shortening. The second one is that syllable constituents are affected differently by prosodic conditions and that onsets and codas do not behave independently of each other within a syllable. These findings need to be included in theories modeling syllable structure such as Articulatory Phonology.

5. Acknowledgements

This research was supported by European Research Council grant number 295573 'Sound change and the acquisition of speech' to Jonathan Harrington.

6. References

[1] Lehiste, I., "Suprasegmentals." MIT Press, Cambridge, 1970.

[2] Byrd, D. & Choi, S., "At the juncture of prosody, phonology, and phonetics — The interaction of phrasal and syllable structure in shaping the timing of consonant gestures." Papers in Laboratory Phonology 10. Mouton de Gruyter, 2010.

[3] Bombien, L., Mooshammer, C., Hoole, P., Rathcke, T. & Kühnert, B., "Articulatory strengthening in initial German /kl/ Clusters under prosodic variation." In Proceedings of the 16th International Conference of Phonetic Sciences. Saarbrücken, pp.457-460, 2007.

[4] Mooshammer, C. and Geng, C., "Acoustic and articulatory manifestations of vowel reduction in German." Journal of the International Phonetic Association, vol.28(2), pp.117-136., 2008.

[5] Katz, J., "Compression effects in English." Journal of Phonetics, vol.40(3), pp.390-402, 2012.

[6] Byrd, D., "C-centers revisited." Phonetica, vol.52, pp.263-282, 1995.

[7] Marin, S. and Pouplier, M., "Temporal organization of complex onsets and codas in American English: Testing the predictions of a gestural coupling model." Motor Control, vol.14, pp.380-407, 2012

[8] Browman, C. and Goldstein, L., "Some notes on syllable structure in articulatory phonology." Phonetica, vol.45, pp.140-155, 1988.

[9] Peters, S. and Kleber, F., "Compensatory vowel shortening before complex coda clusters in the production and perception of German monosyllables." Journal of the Acoustical Society of America, vol.134(5), p.4202, 2013.

[10] Peters, S. and Kleber, F., "Articulatory mechanisms underlying incremental compensatory vowel shortening in German." Abstract submitted to the 10th International Seminar on Speech Production, Cologne, Germany, accepted.

[11] Mücke, D., Grice, M. and Kirst, R., "Prosodic and lexical effects on German place assimilation." Poster Presentation at the 8th International Seminar on Speech Production, 8 - 12 December 2008, Strasbourg, France.

[12] Hoole, P., Zierdt, A. and Geng, C., "Beyond 2D in articulatory data acquisition and analysis. " Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, pp.265-268, 2003.

[13] Schiel, F., "Maus goes iterative.", Proceedings oft he IVth International Conference on Language Resources and Evaluation, pp.1015-1018, 2004.

[14] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer. Software program.", 1992.

[15] Harrington, J., "Phonetic Analyses of Speech Corpora." Chichester: Wiley-Blackwell, 2010.

[16] Marin, S., "The temporal organization of complex onsets and codas in Romanian: A gestural approach." Journal of Phonetics, vol.41, pp.211-227, 2013.

[17] Pouplier, M., "The gestural approach to syllable structure: Universal, language- and cluster-specific aspects." In: Fuchs S., Weirich M., Pape D. and Perrier P., (Eds.), Speech planning and dynamics. New York, Oxford, Wien: Peter Lang, pp.63-96, 2012.

[18] Bombien, L., "Segmental and prosodic aspects in the production of consonant clusters – On the goodness of clusters." Ph.D. Dissertation. Ludwig-Maximilians-University, Munich, 2011.

[19] Oliviera, L., Yanagawa, M., Goldstein, L. and Chitoran, I., "Towards standard measures of articulatory timing." The Journal of the Acoustical Society of America, vol.116(4), pp.2643-2644, 2004.

[20] Cho, T., "Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English." Journal of Phonetics, vol.32, pp.141-176, 2004.

[21] Hawkins, S. and Nguyen, N., "Predicting syllable-coda voicing from the acoustic properties of syllable onsets." In SWAP-2000, pp.167-170, 2000.

Head gesture timing is constrained by prosodic structure

Núria Esteve-Gibert[1], *Joan Borràs-Comes*[1], *Marc Swerts*[2], *and Pilar Prieto*[3,1]

[1]Universitat Pompeu Fabra, Spain
[2]Tilburg University, The Netherlands
[3]Institució Catalana de Recerca i Estudis Avançats

nuria.esteve@upf.edu, joan.borras@upf.edu, m.g.j.swerts@uvt.nl, pilar.prieto@upf.edu

Abstract

There is an increasing consensus to regard gesture and speech as parts of an integrated communication system, in part because of the findings related to their temporal coordination at different levels. In general, results for different types of gestures show that the most prominent part of the gesture (the apex) is typically aligned with accented syllables [6, 10-12, 14, 17]. The aim of the present study is to test for this coordination by focusing on head movements taken from a semi-spontaneous setting in order to look at the effects of upcoming phrase boundaries on their timing. Our results show that while apexes of head gestures are synchronized with accented syllables, upcoming phrase boundaries have an effect on the timing of three gestural points, namely the start, apex, and end time of head gestures. Crucially, these points are aligned differently with respect to the stressed syllable for trochees as compared with iambs/monosyllables, showing that head nods are retracted before upcoming phrase boundaries. This result corroborates previous results by Esteve-Gibert & Prieto [17] for pointing gestures in laboratory settings.

Index Terms: audiovisual speech, head gestures, prosodic structure, face-to-face communication, Catalan

1. Introduction

In face-to-face communication, meanings and intentions are conveyed by means of multimodal strategies, i.e., through both audio and visual channels. In fact, there is progressively more consensus on the idea that gesture and speech both form part of the same system of human communication [1-7]. McNeill [8] listed five main reasons to justify the tight relation between the two modalities: that they (1) co-occur in 90% of cases, (2) develop together in children, (3) are phonologically synchronous, (4) are semantically and pragmatically co-expressive, and (5) break down simultaneously in aphasia.

Several experimental studies have focused on the third of these reasons, namely that gesture and speech are synchronous from a phonological point of view. A number of these studies have found that temporal coordination can be observed between the phases of a gesture movement and related phonological events, in that the prominence in gesture and the prominence in speech co-occur in time [6, 9-14] [see 25 for an overview]. Yet, there is some debate as to what moments or events constitute the anchor points for this alignment between gesture and speech prominence. For the gesture movement, it is generally agreed that the prominence should be located in the *stroke phase*, i.e., the interval of time in which there is a peak of effort [8], or, even more precisely, at the *gesture apex*, i.e., the specific point in time (i.e., not an interval) in which the movement reaches its kinetic 'goal' [15].

There is less consensus on what constitutes the prominent part of speech that temporally coordinates with gesture. In several studies the prominent part of speech is understood as the focused word in the discourse, and they conclude that the stroke of the gesture coordinates with that word [9, 13, 16]. However, other studies take the lexically stressed syllable of (especially) that focused word as the key anchor for the gesture prominence, finding that the stroke of the gesture and the gesture apex coincide with the stressed syllable [10, 12]. And yet other studies integrate the two previous accounts and find that what aligns with the prominent part of the gesture is not simply the stressed syllable of the word in contrastive focus position, but more precisely the moment at which the pitch peak is produced within this contrastive stressed syllable [6, 11, 14, 17].

Esteve-Gibert & Prieto [17] analyzed the coordination between the gesture apex of a pointing gesture and the intonation peak in target words produced with different stress patterns (trochees, iambs, and monosyllables) by Catalan-speakers. Crucially, these words were produced in a contrastive focus condition in order to trigger different positions of the intonation peak within the stressed syllable. They found that the gesture apex and F0 peak co-occurred in time: whereas they were located at the end of the stressed syllable for trochees, they were associated with the middle of the stressed syllable for iambs and monosyllabic words. The main contribution of this article was to show that both intonation and gesture (pointing) movements were bound by prosodic phrasing, such that retracting effects occurred when there was an upcoming phrase boundary (as in monosyllables and iambs), while lagging effects occurred when there was no pre-tonic or post-tonic syllable after a preceding phrase boundary to contain part of the gesture prominence (i.e., in monosyllables).

It is important to point out that almost all the studies listed above described experimental research carried out in tightly controlled settings that hardly resemble natural interactions in face-to-face communication. Also, many of those who found that the gesture prominence coincides in time with the lexically stressed syllable analyzed deictic gestures. To our knowledge, only Loehr [18] investigated the temporal alignment between any kind of communicative hand movements (deictic gestures, iconic, and also beat gestures) and prosodic units (namely pitch accents and phrasing) in natural face-to-face interactions. Using the ToBI annotation system for American English (described in [19]), the author found that the gesture apexes coincided with pitch accents, and that the limits of the gesture phrase (defined as the combination of the gesture stroke and the preparation time needed for the gesture to reach the stroke) tended to coincide with the beginning of the intermediate phrases.

The aim of the present study is to investigate the role of two levels of prominence (accented syllables and prosodic boundaries) in the temporal coordination of head gestures and speech in semi-spontaneous face-to-face communication. Our specific purpose is to test the claim that the prosodic structure influences the timing of the gesture movement in the sense that the placement of the gesture apex within the stressed syllable depends on the metrical pattern of the target word. This has only been tested previously in laboratory settings in which participants did not have a specific communicative purpose while producing the speech and gesture signals. In the present study participants were engaged in a Guess Who game [20] and were not aware of the purpose of the study. An additional interest of the present investigation is that we focus our analysis on head gestures, which can have a potentially different behavior from other types of gestures like hand or eyebrow gestures (though some studies suggest that they show a behavior similar to that of arm and eyebrow movements; see [21] on beat gestures).

2. Methodology

2.1. Participants and procedure

Thirteen Central Catalan-speakers (1 male, 12 female), all of them undergraduates at the Universitat Pompeu Fabra in Barcelona, participated in a production task using two digital variants of the *Guess Who* board game as created by Suleman Shahid and colleagues at Tilburg University [22]. Participants played the game in pairs (i.e., with another native speaker), taking turns in adopting the different roles available. As Ahmad et al. [22] point out, the dynamic nature of games makes them a good tool for investigating human communication in different experimental setups, especially if the outcome of a game is controllable in a systematic manner.

In the *Guess Who* game, participants were presented with a board containing 24 colored drawings of human faces. These faces differed regarding various parameters, such as gender or the color of skin, hair, and eyes. Some faces were bald, some had beards or moustaches, and some were wearing hats, glasses, or earrings. As in the traditional version of *Guess Who*, the purpose of the game was to try to guess the opponent's mystery person before he or she could guess the participant's own. In this way, the game could be used to elicit in a naturalistic way target words with different metrical structures, namely trochees (e.g., *DOna* 'woman', *BARba* 'beard', *NEgre* 'black'), as well as iambs (e.g., *marRONS* 'brown', *barRET* 'hat', *verMELL* 'red') and monosyllables (e.g., *ROS* 'blond', BLAUS 'blue', NOI 'boy')[1].

During the game, participant A had to ask participant B questions to try to determine the mystery person on B's board. Players took turns asking questions about the physical features of their respective "mystery persons" in an effort to eliminate the wrong candidates. The winner was the player who guessed his/her mystery person first. In order to elicit not only questions but also statements, a variation of the game was designed. In this statement-elicitation variation of the game, participants took turns making statements about their mystery person, while the other player listened and eliminated all characters that did not exhibit a particular feature. Again, it was the player who guessed the identity of their "mystery person" first that won. Both participants within a pair took turns in the course of both variations of the game and therefore both provided examples of questions and statements.

Crucially for our goals, the types of simple questions and statements elicited with this procedure had the target words in focus position at the end of the prosodic phrase (e.g., *És una dona*? 'Is it a woman?', *És un home* 'It's a man', *Té bigoti?* 'Has he got a moustache?', *Porta un barret* 'She wears a hat', etc.).

Participants sat in the same room, facing each other across a table and in front of two laptop computers arranged so that they could not see each other's screen. Two camcorders were placed in such a way that they could record the upper part of each participant's body. Once the participants were seated, the camera was raised or lowered according to the participant's height. The experimenter then explained the game and gave instructions about the procedures to be followed for each of the two variations, which took place consecutively. Altogether each game lasted approximately twenty minutes, the time it took to play and win both variants in each set.

2.2. Coding

The relevant utterances (i.e., the questions about the mystery person and the statements used as cues) were annotated in terms of speech and gesture features. For speech, we used Praat [23] to mark the beginning of the opponent's responses and to indicate the duration of the word in focus position as well as the nuclear syllable. Then we imported the Praat label files into ELAN [24]. Figure1 shows an example of labeling with ELAN.

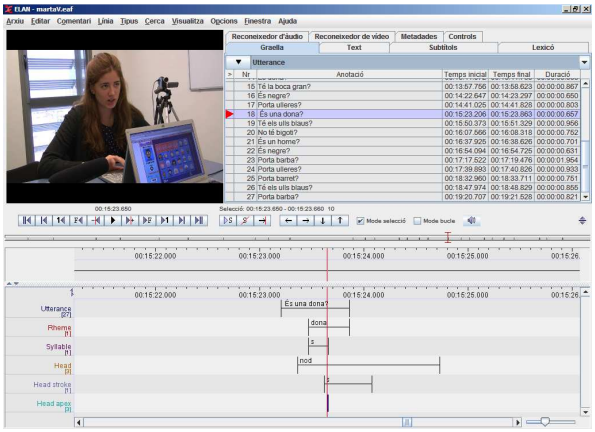


Figure 1: *Example of ELAN labeling.*

As for the gesture annotation, three tiers were created in ELAN, one to label the temporal limits of the head movement, another to locate the stroke of the head movement, and another to label the location of the gesture apex. Four possible head movement were taken into account: head nod, head upward, head tilt and other. *Head nod* referred to a downwards confirmation movement of the head; *upward* referred to a head movement directed upwards (in the opposite direction from

[1] Capital letters indicate the stressed syllable.

nodding); *head tilt* referred to a head inclination or sidewards movement; and *other* referred to any other movements timed with speech, e.g. negation gestures. Following the standard procedure, the annotation of the head movement timing consisted of locating the three gesture phases, namely the preparation phase, the gesture stroke, and the retraction phase. The head movement apex was located at the peak of effort of the head movement [4, 8].

3. Results

The total number of head gestures annotated was 114, consisting of 53 head nods, 42 head tilts, 15 head upwards, and 4 head gestures labeled 'other'. 67 of these gestures appeared in statements and 47 in questions.

3.1. Timing of the gesture apex

Figure 2 shows the temporal distance between the head gesture apex with respect to the end of the accented syllable. These results show that the gesture apex is aligned approximately with the end of the syllable for trochees (M = –67 ms), while it is aligned earlier in the case of monosyllables (M = –265 ms) and iambs (M = –404 ms). These results are consistent with the tendencies described in the literature for stress-final words.

A one-way ANOVA was run with the distance between the gesture apex and the stressed syllable end in milliseconds as the dependent variable and the stress pattern (three levels: trochees, monosyllables, and iambs) as the independent variable. Stress pattern was found to be significant ($F(2, 111) = 5.72$, $p = .004$, ${}_p\eta^2 = .09$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .05$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

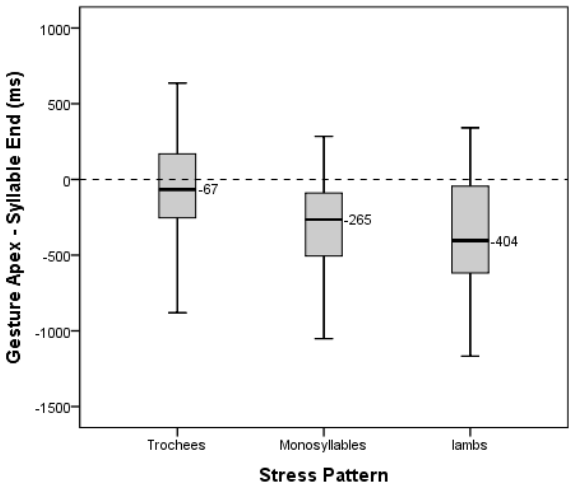


Figure 2: *Distance in time between the apex of the head gesture and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results in this section reveal that the temporal location of the apex of head gestures is significantly affected by the distance to the upcoming phrase boundaries, that is, the apex has to be retracted when the gesture associates with word-final nuclear syllables. Interestingly, this replicates [17]'s results for the coordination between pointing gestures and speech, as they found that the apex of the pointing gesture was retracted before an adjacent phrase boundary (as in monosyllables and iambs), in comparison with gestures associated with non-adjacent phrase boundaries (as in trochees).

3.2. Timing of the gesture start

Figure 3 shows the temporal distance between the start of the head gesture with respect to the end of the accented syllable. These results show that for trochees the head gesture start is aligned closer relative to the end of the syllable (M = –389 ms), than in the case of monosyllables (M = –670 ms) and iambs (M = –734 ms), where it is aligned much earlier.

A one-way ANOVA was run with the distance between the gesture start and stressed syllable end in milliseconds as the dependent variable and stress pattern as the independent variable. Stress pattern was found to be significant ($F(2, 111) = 7.30$, $p = .001$, ${}_p\eta^2 = .12$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .01$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

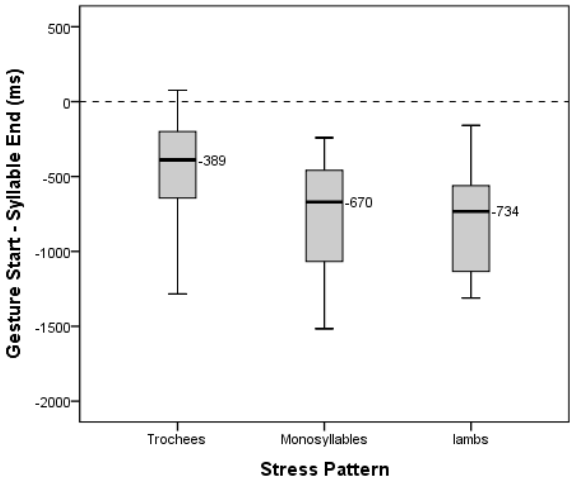


Figure 3: *Distance in time between the start of the head gesture and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results in this section again show an asymmetry between the temporal association of head gestures with trochaic words as compared to iambic and monosyllabic words. In our interpretation, the fact that head gestures start earlier in monosyllables and iambs with respect of the end of the accented syllable indicates that the scope of the head movement is the entire focused word, not only the accented syllable, although results in 3.1 indicate that the anchoring landmark in speech for the gesture apex is the accented syllable.

3.3. Timing of the gesture end

Figure 4 shows the distance in time between the gesture end and the end of the accented syllable. These results show that the gesture end is more closely aligned with the end of the syllable in the case of monosyllables (M = 111 ms) and iambs (M = –21 ms), than in the case of trochees (M = 344 ms).

A one-way ANOVA was run with the distance in time between the gesture end and the stressed syllable end as the dependent variable and stress pattern as the independent variable. Stress pattern was found to be significant ($F(2, 103) = 5.44$, $p = .006$, ${}_p\eta^2 = .10$). Bonferroni post-hoc tests revealed significant differences between trochees and monosyllables ($p < .05$), and also between trochees and iambs ($p < .05$), but not between monosyllables and iambs (p = n.s.).

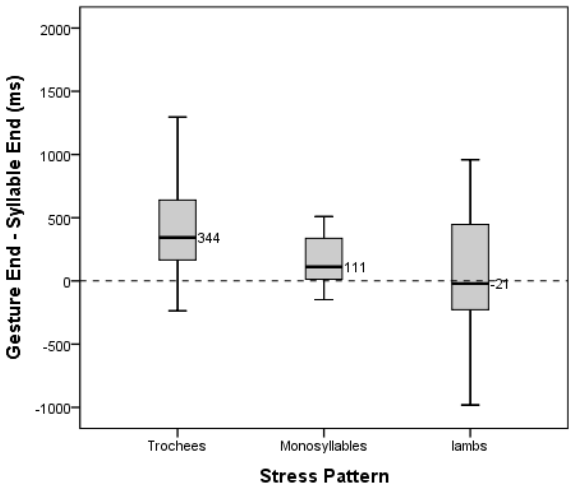


Figure 4: *Distance in time between the gesture end and the end of the accented syllable (in ms), as a function of the stress pattern of the word (trochees, monosyllables, and iambs).*

The results of this section show that the end time of head gestures is located later in trochees than in monosyllables and iambs, something that is related to the fact that this stress pattern has a final unstressed syllable available that can accommodate the retraction phase of the head gesture.

4. Discussion and conclusions

The aim of the present study was to test the claim that prosodic structure influences the timing of gesture movements in the sense that the location of the different phases of the gesture with respect to the stressed syllable depends on the metrical pattern of the target word. This has been tested with head gestures observed in a semi-spontaneous setting, while previous studies examined co-speech gestures produced in laboratory controlled settings.

Our results showed that head gestures are aligned differently with respect to the stressed syllable for trochees than they are for iambs and monosyllables. In trochees, the apex of the gesture is closely aligned with the end of the stressed syllable, the gesture start occurs together with the start of the stressed syllable, and the end of the gesture is located within the final unstressed syllable. By contrast, in iambs and monosyllables, the apex is located in the middle of the stressed syllable, the start of the head movement occurs well before the accented syllable and the ending time occurs right after the accented syllable. These results reveal that the scope of the entire head gesture movement operates on the entire focused word, since preceding and upcoming word boundaries determine the start and end time of the gesture. However, the timing of the gesture prominence (the gesture apex) is determined by the position of the stressed syllable: it occurs earlier when the stressed syllable is followed by a phrase boundary, and it occurs later with respect to the end of the stressed syllable when there is post-tonic material where to accommodate the retraction phase of the gesture. All in all our results show that the timing patterns of head gestures are constrained by prosodic structure and corroborate previous findings in the sense that the most prominent part of the head gesture (the apex) has been shown to be aligned with accented syllables [6, 10-12, 14, 17].

Further research is needed to investigate potential effects of sentence type on the temporal coordination of gesture and prosody. The prosodic structure of statement and questions is different, and this might have an impact on the temporal coordination. Our study does not have enough data to undertake these comparisons, so future studies are crucial to shed light on this issue. The study of the temporal coordination between gesture and speech (in particular, prosody) is important to understand how both modalities are entrained and if they are in fact part of the same system in communication, as proposed in the literature [1, 4, 8]. The fact that gesture and speech are produced in different physiological systems might impose biomechanical constraints in their coordination [25]. However, the evidence presented in this study and in previous work suggests that the analysis of the temporal coordination of both modalities is crucial to investigate the cognitive processes involved in speech and gesture production.

5. Acknowledgments

We would like to thank Suleman Shahid and Constantijn Kaland for their help with setting up the recording sessions, and Igor Jauk for his help with labeling the Catalan corpus. We are grateful to the participants in all the experiments for voluntarily giving us their time. This research has been funded by a research grant awarded by the Spanish Ministry of Science and Innovation (BFU2012-31995 “Gestures, prosody and linguistic structure”), by a grant awarded by the Generalitat de Catalunya (2009SGR-701) to the *Grup d’Estudis de Prosòdia*, and by a “Study abroad scholarship for research outside of Catalunya” 2010 BE1 00207, awarded by the Generalitat de Catalunya.

6. References

[1] Birdwhistell, R. L., “Introduction to kinesics: An annotated system for analysis of body motion and gesture”. Department of State, Foreign Service Institute, Washington DC, 1952.

[2] Birdwhistell, R. L., “Kinesics and context: Essays on body motion communication”, University of Pennsylvania Press, Philadelphia, 1970.

[3] Kendon, A., “Some relations between body motion and speech. An analysis of an example”, in W. Siegman & B. Pope (Eds.), *Studies in dyadic communication*, Pergamon Press, New York, NY,177-210, 1972.

[4] Kendon, A., “Gesticulation and speech: Two aspects of the process of utterance”, in M. R. Key (Ed.), *The relationship of verbal and nonverbal communication*, Mouton, The Hague, The Netherlands, 207-227, 1980.

[5] Kita, S., “How representational gestures help speaking”, in D. McNeill (Ed.), *Language and Gesture*, Cambridge University Press, Cambridge, 2000.

[6] De Ruiter, J. P., *Gesture and speech production*, unpublished doctoral dissertation). Katholieke Universiteit, Nijmegen, The Netherlands, 1998.

[7] McNeill, D., *Language and Gesture: Window into Thought and Action*, Cambridge University Press, Cambridge, 2000.

[8] McNeill, D., *Hand and mind*, University of Chicago Press, Chicago, IL, 1992.

[9] Butterworth, B. and Beattie, G., “Gesture and silence as indicators of planning in speech”, in R. Campbell & G. T. Smith (Eds.), *Recent advances in the psychology of language: Formal and experimental approaches*, Plenum Press, New York, NY, 347-360, 1978.

[10] Loehr, D. P., “Aspects of rhythm in gesture and speech”, *Gesture*, 7: 179-214, 2007.

[11] Nobe, S., *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/ threshold model of gesture production*, unpublished doctoral dissertation, University of Chicago, 1996.

[12] Rochet-Capellan, A., Laboissière, R., Galván, A., and Schwartz, J. L., “The speech focus position effect on jaw–finger coordination in a pointing task”, *Journal of Speech, Language, and Hearing Research*, 51: 1507-1521, 2008.

[13] Roustan, B. and Dohen, M., “Co-production of contrastive prosodic focus and manual gestures: Temporal coordination and effects on the acoustic and articulatory correlates of focus”, *Proceedings of Speech Prosody 2010 Conference*, 100110, 1-4, 2010.

[14] Rusiewicz, H. L., Shaiman, S., Iverson, J. M., and Szuminsky, N., “Effects of perturbation and prosody on the coordination of speech and gesture”, *Speech Communication*, 57,283-300, 2014.

[15] Kendon, A., *Gesture: Visible action as utterance*, Cambridge University Press, Cambridge, 2004.

[16] Ferré, G., “Timing relationships between speech and co-verbal gestures in spontaneous French”, *Proceedings of Language Resources and Evaluation, Workshop on Multimodal Corpora*, 86-91, 2010.

[17] Esteve-Gibert, N. and Prieto, P., “Prosodic structure shapes the temporal realization of intonation and manual gesture movements”, *Journal of Speech, Language, and Hearing Research* 56: 850-864, 2013.

[18] Loehr, D., “Temporal, structural, and pragmatic synchrony between intonation and gesture”, *Laboratory Phonology*, 3(1): 71-89, 2012.

[19] Beckman, M. and Ayers-Elam, G., “Guidelines for ToBI labeling”, ver. 3, Ohio State University, 1997.

[20] Borràs-Comes, J., Kaland, C., Prieto, P., and Swerts, M., “Audiovisual Correlates of Interrogativity: A Comparative Analysis of Catalan and Dutch”, *Journal of Nonverbal Behavior*, in press, published online: 05 October 2013.

[21] Krahmer, E. and Swerts, M., “The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception”, Journal of Memory and Language, 57(3): 396-414, 2007.

[22] Ahmad, M. I., Tariq, H., Saeed, M., Shahid, S., and Krahmer, E., “Guess who? An interactive and entertaining game-like platform for investigating human emotions”, in Jacko, J. A. (Ed.), *Human-computer interaction. Towards mobile and intelligent interaction environments*, vol. 3, *Lecture Notes in Computer Science*, 6763, Springer, Berlin, 543-551, 2011.

[23] Boersma, P. and Weenink, D., *Praat: Doing phonetics by computer*, version 5.3.04, computer program, 2012.

[24] Lausberg, H. and Sloetjes, H., “Coding gestural behavior with the NEUROGES-ELAN system”, *Behavior Research Methods, Instruments, & Computers*, 41: 841-849, 2009.

[25] Wagner, P., Malisz, Z., and Kopp, S. “Gesture and speech in interaction: An overview”. Sp. Comm 57: 209-232, 2014.

The ternary contrast of consonant duration in Inari Saami

Helen Türk[1] , Pärtel Lippus[1,2], Karl Pajusalu[1], Pire Teras[1]

[1] Institute of Estonian and General Linguistics, University of Tartu, Estonia
[2] Institute of Behavioural Sciences, University of Helsinki, Finland
helen.tyrk@gmail.com, {partel.lippus, karl.pajusalu, pire.teras} @ut.ee

Abstract

The three-way distinction of quantity occurs in several Finnic and Saami languages. The paper focuses on the length contrast of consonants in Inari Saami. Similarly to Estonian and other Finno-Ugric languages where three quantities are described, in Inari Saami the distinction between single consonants, short geminates or consonant clusters, and long geminates or consonant clusters appears only on the boundary of a stressed and unstressed syllable of a disyllabic foot. Our results show that in Inari Saami the duration of consonants is inversely related to the duration of both preceding and following vowels, and there is a tendency towards foot isochrony. The results are in line with previous studies on quantity opposition in Inari Saami and in other Finnic languages, showing the ternary distinction of consonant quantities as a foot-level feature of the language.

Index Terms: Inari Saami, geminates, three-way quantity

1. Introduction

The sound system of Inari Saami reveals three phonologically distinctive quantities. The ternary duration contrast occurs in several Saami languages, including North Saami which has a central position in the Saami language area [1]–[3]. Inari Saami is an eastern Saami language spoken by about 200 native speakers in northern Finland. The Inari Saami phonology is characterized by left-headed feet, word-initial primary stress, and a distinction between short and long vowels and consonants both in stressed and unstressed syllables. The three-way distinction of quantity appears only with consonants in primary stressed feet that are left headed. The ternary contrast is realized by the distinction of single consonants, short and long geminates (traditionally called half-long and long consonants), or consonant clusters on the boundary of the stressed syllable and the following unstressed syllable, e.g. *palo* [pɑlo] ‘fear, Gen/Acc.’, *paḷo* [pɑllo] ‘fear, Nom.’, *kallu* [kɑlːlu] ‘forehead, Nom.’, *ša'lde* [ʃɑlte] ‘bridge, Gen/Acc.’, *šalde* [ʃɑlːte] ‘bridge, Nom.’, *tääḷu* [tæːllu] ‘house, Nom.’, *täállun* [tæːlːlun] ‘house, Ess.’. In orthography, short geminates (or half-long consonants) are marked with a dot under a single letter, long geminates (or long consonants) are written with two letters. An apostrophe before a consonant cluster indicates that the cluster is short, see [4].

The phonological distinction between short and long geminates is a productive feature of Inari Saami word prosody, i.e. it occurs with all consonants. Also the consonant clusters are prosodically short and long; short geminates and consonant clusters occur in feet of the second quantity degree (Q2), long geminates and consonant clusters occur in feet of the third quantity degree (Q3).

Previous studies have indicated that segmental durations are interrelated in Inari Saami feet and there is a tendency to foot isochrony, which means that the length of the first and second syllables are inversely related [2], [5]. However, unlike Southern Finnic languages with ternary quantity opposition, the Saami languages including Inari Saami also preserved the quantity distinction of vowels in an unstressed syllable [1], e.g. *palloon* [pɑlːloːn] ‘fear, Ess.’.

The nature of temporal relations between consonants and surrounding vowels in Inari Saami is not completely clear. Earlier studies of Inari Saami quantities have focused on the relations of the consonants with the preceding vowel [2], [6]. Southern Finnic languages with a three-way quantity, on the other hand, have shown an inverse relation between consonants and the duration of the following vowel. Markus et al. found that this is relevant also in the case of Inari Saami [5]. In this paper we study the temporal features of all sounds in Inari Saami disyllabic feet with short consonants, short and long geminates and consonant clusters.

2. Materials and method

The data of this study were recorded using an Edirol R-09 digital recorder in 2013 from four male native speakers of Inari Saami. Two of the subjects were born in Inari, one in Syysjärvi and one in Ylivieska. Their parents were speakers of central and northern varieties of Inari Saami. Currently one speaker still lives in Syysjärvi, one has moved to Helsinki and two live in Ivalo. At the time of recording the age of the speakers was between 62 and 77 (average being 70.8). In addition to their native language, all subjects speak Finnish, three have a good knowledge of North Saami, and three also marked English or German as their foreign languages.

The total set of materials comprised 299 words with consonantal quantity embedded in 96 carrier sentences in Inari Saami. All test words were disyllabic with a phonologically short vowel as a syllable nucleus, while the intervocalic consonant was a short consonant (Q1; e.g. *sare* [sɑre] ‘blueberry, Gen/Acc.’, *kove* [kove] ‘picture, Gen/Acc.’), a short geminate (Q2; e.g. *saṛe* [sɑrre] ‘blueberry, Nom.’, *koṿe* [kovve] ‘picture, Nom.’), or a long geminate (Q3; e.g. *komme* [komːme] ‘ghost, Nom’, *hekki* [hekːki] ‘cage, Nom.’), a short consonant cluster (Q2; e.g. *pu'ško* [puʃko] ‘Esox, pike, Gen/Acc.’, *a'lge* [ɑlke] ‘boy, Gen/Acc.’) or a long consonant cluster (Q3; e.g. *puško* [puʃːko] ‘Esox, pike, Nom.’, *alge* [ɑlːke] ‘boy, Nom.’). Different vowels and syllable boundary consonants were selected to avoid the influence of the intrinsic duration on average segment duration. The analysed word structures were as follows: CVCV, (C)VCCV and (C)VC:CV (referred to as Q1, Q2 and Q3, respectively).

The test words were placed in phrase-medial and phrase- or sentence-final position of the carrier sentence, e.g. *Ohtâ **maṇe** lii taa*, *mut ohtâ **lodde** lii tobbeen* ‘One egg is here, but one bird is there’; *Must lii ohtâ **saṛe**, mut sust lii ohtâ **juŋŋâ*** ‘I have one blueberry, but you have one lingonberry’. The distribution of the analyzed tokens is shown in Table 1. Some utterances were left out from the analysis, mainly due to background noise or because they were misread.

Table 1. *Number of analyzed tokens produced by the four speakers.*

	Sp1	Sp2	Sp3	Sp4
Q1 short	6	7	6	7
Q2 geminate	14	14	14	14
Q2 cluster	8	8	8	8
Q3 geminate	24	22	23	24
Q3 cluster	24	22	22	24

Segment boundaries were labelled in Praat [7] and the duration of each segment was extracted with a script. Using the LME4 package in R, the log-scaled segment durations were tested with a mixed effects model for three factors: Position (levels: phrase-medial, phrase-final), Quantity (levels: Q1, Q2, Q3) and Consonantal (C2) structure (levels: geminate, consonant cluster).

3. Results and discussion

First, we present the segmental durations in disyllabic words, and then we compare the duration ratios of intervocalic consonants and their surrounding vowels (i.e. the ratios of V1/C2 and C2/V2).

3.1. Duration of segments

The average segmental durations are presented in Table 2. In the table, C1 marks the word-initial consonant. V1 is a short vowel in the first syllable, C2 a short consonant (Q1), the total duration of a short (Q2) and long (Q3) geminate, or the total duration of a short (Q2) and long (Q3) consonant cluster, and V2 a short vowel in the second syllable. The total duration of the disyllabic foot is also given. In the first, second and third part of the table respectively the average segment durations of words in the phrase-medial, phrase-final position and in both positions together are given. Figure 1 illustrates an overall average of segment durations (phrase-medial and phrase-final words pooled together).

It can be seen from Table 2 that the C1 duration is roughly 100 ms in all tested quantity degrees and phrasal positions. None of the tested factors were significant for the C1 duration.

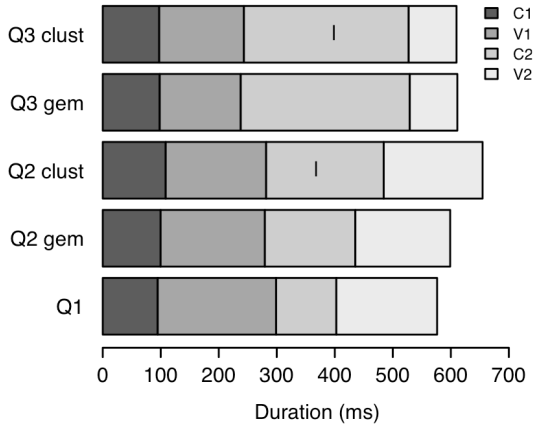


Figure 1: *Average segment durations in the analyzed word structures. The segment boundary in the consonant cluster is marked with a vertical bar.*

Table 2. *Average segment durations and standard deviations (in milliseconds) in the phrase-medial, phrase-final position and both positions analyzed together.*

Pos.	Structure	C1	V1	C2	V2	Total
Phrase-medial	Q1 short	97	181	82	163	522
	s.d.	15	28	14	21	47
	Q2 geminate	96	166	137	148	547
	s.d.	23	26	23	21	52
	Q2 cluster	110	161	185	148	577
	s.d.	17	31	26	26	60
	Q3 geminate	97	132	242	78	532
	s.d.	27	22	68	16	82
	Q3 cluster	96	135	238	75	527
	s.d.	16	19	50	13	65
Phrase-final	Q1 short C	93	224	122	183	623
	s.d.	16	46	32	50	88
	Q2 geminate	103	194	175	179	651
	s.d.	24	35	41	44	92
	Q2 cluster	107	186	220	192	679
	s.d.	17	29	37	39	91
	Q3 geminate	100	148	342	86	658
	s.d.	31	25	73	22	95
	Q3 cluster	99	158	334	91	664
	s.d.	17	21	56	20	75
Average	Q1 short C	95	204	104	174	576
	s.d.	16	44	32	40	87
	Q2 geminate	100	180	156	164	599
	s.d.	23	34	38	37	91
	Q2 cluster	109	173	203	170	628
	s.d.	16	32	36	40	92
	Q3 geminate	98	140	291	82	594
	s.d.	29	25	87	20	109
	Q3 cluster	97	146	284	82	593
	s.d.	16	23	72	18	98

Table 2 shows that the average duration of V1 is the longest in the case of Q1 (204 ms), somewhat shorter in the case of Q2 (173–180 ms), and the shortest in the case of Q3 (140–146 ms). The main effect of Position is significant [$\chi^2(df=1, N=299)=28.9, p<0.001$], but there are no interactions with Quantity and C2 structure. The effect of Position refers to the lengthening of segments in phrase-final position. There is also a significant main effect of Quantity [$\chi^2(df=2, N=299)=63.41, p<0.001$], and post-hoc test indicates that the duration of V1 varies significantly in the opposition of Q1 and Q2 vs. Q3 ($p<0.001$). The V1 duration in Q1 vs. Q2 is not significantly different. Additionally, there is no effect of C2 structure.

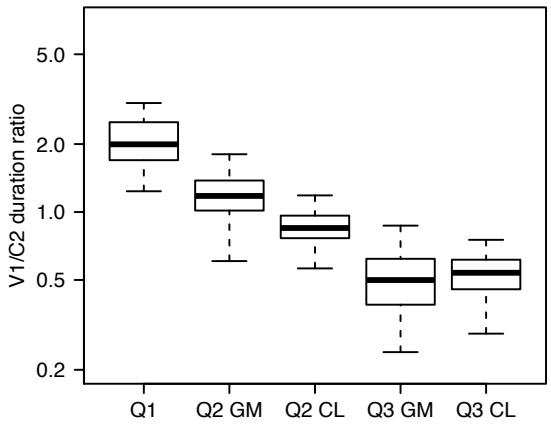


Figure 2: *Duration ratios of V1 to C2.*

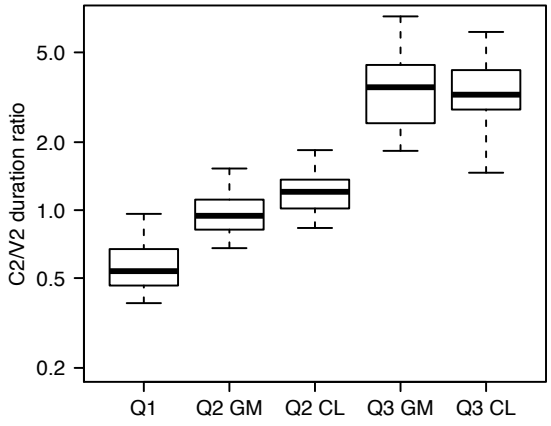


Figure 3: *Duration ratios of C2 to V2.*

As expected, the intervocalic short consonant has the shortest duration (104 ms). The short geminate is longer (156 ms) than the short consonant and the short consonant cluster is somewhat longer than the short geminate (203 ms). The long geminate and consonant cluster are the longest (291 ms and 284 ms respectively). Standard deviations show a greater variation in the duration of both long (Q3) geminate and consonant cluster. The Position has a main effect on the duration of intervocalic consonant(s) (C2) [χ^2(*df*=1, *N*=325)=71.965, p<0.001], which, as in the case of V1, indicates the phrase-final lengthening of a word. The duration of C2 in different quantities varies significantly [χ^2(*df*=2, *N*=325)=137.31, p<0.001], being the shortest in Q1 and the longest in Q3. There is also an interaction between Quantity and C2 structure [χ^2(*df*=2, *N*=325)=16.1, p<0.001]. Pairwise post-hoc testing shows a significant difference between all levels of Quantity (p<0.001), but C2 structure has a significant effect only in the case of Q2 (p<0.05) where a geminate is shorter than a consonant cluster. In Q3 a geminate and a consonant cluster are of similar duration.

V2 shows a similar pattern to V1, being the longest in Q1 (174 ms), shorter in Q2 (164–170 ms), and the shortest in Q3 (82 ms). Again, Position has a significant effect, but without any interactions [χ^2(*df*=1, *N*=299)=22.1, p<0.001]. As in the case of V1 and C2 it also points to the phrase-final lengthening. There is a significant main effect of Quantity [χ^2(*df*=2, *N*=299)=167.24, p<0.001] and post-hoc test shows the difference between Q1 and Q2 vs. Q3 (p<0.001), but no difference between Q1 vs. Q2.

The duration of the whole word varies significantly only between different phrasal positions [χ^2(*df*=1, *N*=299)=85.6, p<0.001], but quantities do not reveal a significant difference. This lack of difference between the word structures with different quantity degrees can be accounted for by a strong tendency to foot isochrony.

It can be concluded that the phrasal position influences the duration of all segments except C1. Segments are longer in the phrase-final position than in the phrase-medial position, but the phrasal position does not interact with the other tested factors. There is an interrelation between the durations of the intervocalic consonant(s) and the surrounding vowels: while the duration of C2 increases both the duration of V1 and V2 decreases. In consequence, the total duration of feet reveals a tendency to foot isochrony. The average foot durations are similar in all the quantities.

3.2. Duration ratio of segments

In Table 3 the duration ratios of V1 to C2, and C2 to V2 are presented. Figure 2 illustrates the duration ratios of V1 to C2 and the duration ratios of V2 to C2 are presented in Figure 3. As the phrasal position seems to have an overall lengthening effect and it does not interact with the different segmental patterns, the phrasal positions are pooled together in this section.

Table 3. *Average V1/C2 and C2/V2 duration ratios.*

Structure	V1/C2	C2/V2
Q1short	2.0	0.6
Q2 geminate	1.2	1.0
Q2 cluster	0.9	1.2
Q3 geminate	0.5	3.5
Q3 cluster	0.5	3.5

The average duration ratio of V1/C2 is 2 in Q1, 0.9–1.2 in Q2 and 0.5 in Q3. The first syllable vowel (V1) in Q1 words is twice as long as a single consonant (C2). Before the short (Q2) geminate and consonant cluster the vowel is shorter than before a single consonant, which in turn is shorter than a Q2 geminate and consonant cluster. In the case of Q2 the durations of the first vowel and the geminate or consonant cluster are almost equal. The V1 duration is the shortest and the C2 duration the longest in the case of Q3 and because of that the duration of a Q3 geminate and consonant cluster is twice as long as that of the first syllable vowel.

The duration of V2 is also strongly affected by the duration of the preceding consonant; there is an inverse relation. The duration ratios of C2/V2 are as follows: 0.6 in Q1, 1.0–1.2 in Q2, and 3.5 in Q3. The duration of the second syllable vowel is the longest after a short consonant: V2 is almost two times longer than C2. In the case of Q2 a short geminate and consonant cluster have almost same duration as the second syllable vowel. V2 is the shortest after the long

(Q3) geminate and consonant cluster that is more than 3 times longer than the following vowel.

Considering both duration ratios, the general correlation between the durations of neighbouring segments can be stated as follows. In the case of Q1: the duration of V1 > the duration of a short consonant < the duration of V2. In the case of Q2: the duration of V1 = the duration of a short geminate or consonant cluster = the duration of V2. In the case of Q3: the duration of V1 < the duration of a short geminate or consonant cluster > the duration of V2.

The duration ratio of the short consonant to the short geminate and long geminate is 1 : 1.5 : 2.8, and to the short consonant cluster and long consonant cluster it is 1 : 2 : 2.7. The duration ratio of the short geminate to the long geminate is 1.9 and of the short consonant cluster to the long consonant cluster it is 1.4. These ratios show that the duration of short geminates is closer to the duration of the short consonants than the duration of the short consonant clusters. The short geminate is one and a half times longer than the short consonant, and almost two times shorter than the long geminate. The short consonant cluster is two times longer than the short consonant and almost one and a half times shorter than the long consonant cluster. Long geminates and consonant clusters are almost three times longer than short consonants.

The results are in line with previous studies on quantity opposition in Inari Saami and other Finnic languages showing the ternary distinction of consonant quantities as a foot-level feature of the language. Bye et al. report that all their speakers make a ternary distinction in consonant duration after a short vowel [2]. However, in their data, the duration of V1 and V2 had a greater between-speaker variation, which is explained partly with a different language background of speakers. Similarly to the speakers of the current study, some of their speakers had a reverse relation of V1 and C2: a short V1 was longer before a short geminate than before a long geminate, and yet longer before a short consonant. However, for three of their speakers the difference of V1 before a short consonant and geminate was not significant. In line with this, the present study showed that the duration of surrounding vowels is significant only between Q1 and Q2 vs. Q3 but not between Q1 vs. Q2. Bye et al. report that only one speaker displayed a ternary inverse duration relationship between consonant and V2 but for other speakers there was no significant difference in V2 duration after a short consonant and geminate [2]. The latter applies to the pronunciation of our speakers, too.

Similar temporal ratio patterns between the intervocalic consonants and the surrounding vowels have been shown to be an efficient way to describe the quantity system of other languages that have a ternary contrast of consonant duration (e.g. [5]). In Estonian the variation of the intervocalic consonant duration does not affect V1 duration, but V2 duration is the longest after a short consonant and shortest after a long geminate. In Livonian, the variation of the intervocalic consonant duration does not affect V1 duration either, but there is a significant difference in V2 duration after a short consonant and geminate vs. a long geminate. Markus et al. have also reported duration ratios of consonants in different quantity [5]. In Estonian and Livonian the duration ratio of CC/C is respectively 2.19 and 1.49 and that of C:C/CC 1.4 and 1.52. The duration ratios of Inari Saami geminates seem to resemble that of Livonian, but the ratios of consonant clusters is more similar to that of Estonian long and short geminates.

4. Conclusions

Inari Saami has a ternary contrast of consonant quantity that occurs after a short vowel. Similarly to Estonian and some other Finno-Ugric languages where three-way quantities are described, in Inari Saami the distinction between short consonants, short geminates or consonant clusters, and long geminates or consonant clusters appears only on the boundary of a stressed and unstressed syllable of a disyllabic foot. In Inari Saami the duration of consonants is inversely related to the duration of both preceding and following vowels. The duration of V1 is significantly longer before a short consonant, short geminate and consonant cluster than before a long geminate and consonant cluster. The same difference concerns the duration of V2. Consequently there is a strong tendency towards foot isochrony. Duration ratios between consonants in different quantity indicate that the duration of the short geminate is closer to the duration of the short consonant than to the duration of the long geminate (the ratios 1.5 and 1.9). However, the duration of the short consonant cluster is closer to the duration of the long consonant cluster than to the duration of the short consonant (the ratios 2 and 1.4).

5. Acknowledgements

The authors are very grateful for the Inari Saami informants for participating in this study. Our special gratitude belongs to Hans Morottaja who helped us to find speakers. We would like to thank Eva Liina Asu for proofreading this paper. The second author would also like to thank Nele Salveste and Juraj Šimko for inspiring discussions. This research was partly funded by the Estonian Research Agency grant No. IUT2-37.

6. References

[1] P. Sammallahti, *The Saami languages: An introduction.* Karasjok, Norway: Davvi Girji, 1998.

[2] P. Bye, E. Sagulin, and I. Toivonen, “Phonetic Duration, Phonological Quantity and Prosodic Structure in Inari Saami”, *Phonetica*, vol. 66, no. 4, pp. 199–221, 2009.

[3] B. A. Bals Baal, D. Odden, and C. Rice, “An analysis of North Saami gradation”, *Phonology*, vol. 29, no. 2, pp. 165–212, Sep. 2012.

[4] P. Sammallahti and M. Morottaja, *Säämi-suomâ sänikirje. Inarinsaamelais-suomalainen sanakirja*. Ohcejohka: Girjegiisá Oy, 1993.

[5] E. Markus, P. Lippus, K. Pajusalu, and P. Teras, “Three-way opposition of consonant quantity in Finnic and Saamic languages”, in *Nordic Prosody. Proceedings of the XIth conference, Tartu 2012*, E. L. Asu and P. Lippus, Eds. Frankfurt am Main: Peter Lang, 2013, pp. 225–234.

[6] E. Sagulin, “Konsonant- och vokalduration I enaresamiska”, Examens arbete I matematisk statistik, Uppsala Universitetet, 2008.

[7] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. Computer program, 2013.

10 Wednesday 3

Slovak prosody in the phonetics-phonology debate: Yers and emergent prosodic breaks

Štefan Beňuš

Department of English and American Studies, Constantine the Philosopher University, Nitra, Slovakia
Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
sbenus@ukf.sk

Abstract

Prosody is central for understanding the cognitive system underlying human speech and relates to both more granular aspects of our phonological competence as well as more continuous aspects of observable articulatory movements and resulting acoustic characteristics. The understanding, and formal treatment, of the relationship between these two inter-related components of human speech is at the core of the cognitive approach to speech. In this presentation I contribute to this discussion by drawing links between two seemingly unrelated lines of my research on Slovak [1, 2], and argue that understanding the continuous prosodic nature of speech is critical for improving our understanding of cognitive competence underlying it. The first aspect concerns yer vowels as the proto-typical problem of Slavic phonology [1], the second involves the nature of prosodic boundaries [2].

In Slavic phonology, 'yer', or 'jer' is a widely used and recognized term for vowels that alternate with zero in many Slavic languages and can be traced to Old Church Slavonic and Indo-European. In Slovak, the presence of mid-vowels /e/ and /o/ in some words alternates with their absence as in (1).

(1) Alternations with yers in Slovak.

Nom.Sg.	Transcription	Gen.Sg.	Instr.Sg.	Gloss
pal**e**c	[palets]	palc-a	palc-om	'thumb'
lak**e**ť	[lakec]	lakť-a	lakť-om	'elbow'
p**e**s	[pes]	ps-a	psom	'dog'
kot**o**l	[kotol]	kotl-a	kotl-om	'cauldron'
pár**o**k	[pa:rok]	párk-a	párk-om	'sausage'

(2) Yer and non-yer vowels occur in similar environments

Yer	Gen.Sg.	Gloss	Non-yer	Gen.Sg.	Gloss
kábel	kábl-a	cable	Ábel	Ábel-a	name
palec	palc-a	thumb	balet	baleta	balet
párok	párk-u	sausage	nárok	nárok-u	requirement
smútok	smútk-u	sorrow	súťok	súťok-u	confluence

The patterns in (2), together with other features of Slovak, show that the yer vs. non-yer alternations cannot be treated as insertions or deletions since the environment could not be specified. Therefore, given well known reasons stemming from alternations as in (2), all phonological accounts assume that yer vowels are underlyingly different from non-yer vowels. This is common to all formal yer treatments despite important differences in theoretical machinery and predictions generated in phonological analyses using various formalizations [e.g. 3, 4, 5, 8, 9]. The second assumption, shared by all traditional accounts, is that the underlying difference between yer and non-yer vowels is neutralized in phonology (e.g. the traditional rule of Lower in [3]), and there should thus be no phonetic difference between the two vowel classes. I will review the phonetic evidence in [1] that analyzes the acoustic and articulatory aspects of producing words like in (2) and suggests that yer vowels are prosodically weaker than their non-yer counterparts and resemble vowels produced in faster speech rate. Hence, sub-phonemic prosodically-based differences might participate in modeling deep morpho-phonological alternations.

The second line of research investigates the spontaneous emergence of high-level prosodic boundaries induced by resolving low-level requirements for slower speech rate or more precise articulation in the vicinity of the syntactic affordance for such boundaries [2]. We investigate the patterns of temporal coordination among the bilabial gestures (m-opening, b-closing) and the tongue body gesture forming the vowel canonically following /b/ in real-word Slovak sequences in (3).

(3) Stimuli for emergence prosodic break; "#" denotes syntactic affordance

Slovak	IPA
Čítam (#) iba mu ...	[tʃi:ta**m**(#)i**ba**mu...]
Cítim (#) aby mu ...	[tsi:ci**m**(#)a**bi**mu]

We observed that the continuous variation of low-level tempo and hypo-hyper articulation resulted in continuous re-organization of the gestures reflecting the continuous variation in the boundary strength. The extrapolation between our findings and those reported in literature for linguistically planned brakes suggests a plausible hypothesis that both species of prosodic boundaries, i.e. 1) planned, top-down, traditionally phonological ones stemming from the interface between prosody and syntax/pragmatics and 2) emergent, bottom-up ones, realized though low-level phonetic variations, can stem from a single underlying mechanism, and can be implemented in the same way. Moreover, the observed systematicities can be accounted for using the formal optimization-based Embodied Task Dynamics model [6, 7] in which the variation in inter-gestural timing stemming from the prosodic characteristics arises through localized changes in relative demands on efficient perception, articulatory precision and temporal cohesion among the sequenced gestures.

Both lines of research thus suggest that exploring traditional phonological alternations as embedded in continuous prosodic substrate can provide novel insights into the cognitive mechanisms underlying our communicative competence.

Index Terms: phonetics-phonology, yers, prosodic boundaries, Slovak, articulation

References

[1] Beňuš, Š., "Phonetic variation in Slovak yer and non-yer vowels," Journal of Phonetics 31(1): 153-169, 2012.

[2] Beňuš, Š., Šimko, J., "Emergence of prosodic boundary: continuous effects of temporal affordance on inter-gestural timing," Journal of Phonetics, in press. [http://dx.doi.org/10.1016/j.wocn.2013.12.005].

[3] Lightner, Th. M, "Segmental phonology of contemporary standard Russian," Ph.D. Dissertation, MIT Press, 1965.

[4] Rubach, J., "The Lexical Phonology of Slovak," Oxford: Clarendon Press, 1993.

[5] Scheer, T., "How yers made Lightner, Gussmann, Rubach, Spencer and others invent CVCV," In: P. Bański, B. Łukaszewicz, M. Opalińska (Eds.), Studies in constraint-based phonology, pp. 133–207.Warsaw: Wydawnictwo Uniwersytetu Warszawskiego.

[6] Šimko, J., Cummins,F., "Embodied task dynamics," Psychological Review, 117(4): 1229–1246, 2010.

[7] Šimko, J., Cummins,F., "Sequencing and optimization within an embodied task dynamic model," Cognitive Science, 35(3): 527–562, 2011.

[8] Szpyra, J., "Ghost segments in nonlinear phonology: Polish yers. Language, 68: 277–312, 1992.

[9] Yearley, J., "Jer vowels in Russian. In: J. N. Beckman, L. Walsh Dickey, S. Urbanczyk (Eds.), Papers in optimality theory, pp.533–571. Amherst: GLSA, University of Massachusetts. Occasional papers in linguistics 18.

On the Origins of the Prosodic Word in Russian

Jaye Padgett

Department of Linguistics, University of California, Santa Cruz, USA
padgett@ucsc.edu

Abstract

The Prosodic Word (Pwd) is a foundational notion in phonological theories, being relevant for the statement of many phonological generalizations. In spite of their importance, there are basic open questions about prosodic words. Where do they come from? Can their structure in one language vs. another be predicted? In this paper I suggest a research program that attempts to address such questions by viewing prosodic words as emergent over time from the interaction of phonetics, phonologization, and syntactic structure.

Index Terms: prosodic word, domain generalization, Russian

1. Introduction

The Prosodic Word (Pwd) occupies a central position in the theory of Prosodic Phonology, e.g. [1], [2], [3], [4], [5]. A key motivation for the Pwd, as separate from the morphosyntactic word, is the lack of isomorphism between the two kinds of word. For example, word-final devoicing in Russian affects open-class lexical items but not prepositions, leading to contrasts like *sa**t** mʲiˈxailə* 'Mikhail's garden' vs. *pə**d** mɛsˈkvoj* 'near Moscow', in which the underlying /d/ of /sad/ 'garden' is devoiced (cf. *sada* 'garden (gen.)') but the /d/ of the preposition /pod/ is not. This contrast, among other facts, leads many researchers to conclude that the first phrase consists of two prosodic words, i.e. [sat]$_{\text{Pwd}}$ [mʲiˈxailə]$_{\text{Pwd}}$, while the second consists of one, i.e. [pəd mɛsˈkvoj]$_{\text{Pwd}}$, and to assume that devoicing affects consonants at the end of the prosodic word [6], [7], [8], [9], see discussion and references in [10].

The word is relevant to phonetic theories as well. For example, phonetic domain-initial strengthening and domain-final lengthening are most strongly observed at the highest level of the prosodic hierarchy such as the utterance and are more weakly present at the level of the word [11], [12]. There is also evidence that degree of coarticulation can depend on whether the relevant segments span a prosodic word boundary or are within a prosodic word [13].

In spite of their importance, there are many basic questions about prosodic words that remain unanswered (see discussion in [14]). How many prosodic categories are there? Are they innately given or emergent constructs, and if the latter, what explains their emergence and the precise form they take? What are the constraints on the structure of prosodic categories? It is much easier to ask these questions than to answer them, and this paper has the modest goal of suggesting a research program in which prosodic words (and possibly other higher prosodic constituents) are viewed as constructs that emerge over time through the interaction of phonetics, phonologization, and syntactic structure. A key component of this pursuit is something called *domain generalization*.

2. Domain generalization

Word-final devoicing as in Russian is attested in many unrelated languages. The account of it detailed in this section follows [15].

2.1. Phonetics-phonology mismatch

Many researchers have posited that final devoicing originates as a phonologization of utterance-final phonetic devoicing (e.g., [16], [17]). Gradient utterance-final devoicing occurs in many languages and can be attributed to a drop in sub-glottal pressure toward the end of an utterance [18] as well as spreading of the vocal folds in anticipation of non-speech breathing posture (e.g., [19], [20]). In addition, it has been argued that an obstruent voicing contrast might be hard to perceive unless the relevant obstruents directly precede a sonorant consonant or vowel [21]; since utterance-final consonants precede a pause, there may be perceptual as well as articulatory underpinnings to devoicing. These phonetic underpinnings are relevant to utterance-final position, but not word-final position (putting aside words that happen to be utterance-final). In phrasal contexts like *sa**t** mʲiˈxailə*, in which the word-final obstruent is utterance-medial and precedes a sonorant, there are no articulatory or perceptual underpinnings for word-final devoicing analogous to those described above. Yet many languages have phonologized final devoicing specifically at the level of the word all the same.

2.2. Domain generalization

The idea of *domain generalization* is that language learners, even while encountering a generalization about utterance-final position, are predisposed to learn them as word-final. Suppose that phonological generalizations are built from a store of lexical representations [22], [23]. It is plausible to assume also that we store many more words than phrases or utterances. First, we encounter many more words than utterances (since words make up utterances). Second, words are also easier to remember, because they tend to be shorter than utterances, and a given word is reinforced in memory more often by repeated exposure than a given utterance. Words are therefore a more likely source of generalization.

Domain generalization implies that word-final devoicing comes about in the following way. At an initial stage of a given language (here we entertain the scenario using Modern Russian forms), devoicing begins as a phonetically motivated utterance-final process, as in Stage 2 below. At this stage words like /sad/ 'garden' are realized with final devoicing when they occur in utterance-final position but not elsewhere. But under the influence of the many stored devoiced variants like [sat] of words like /sad/, the learner generalizes devoicing to *all* words. This is Stage 3. This process can be seen underway in Polish, where some dialects maintain utterance-final devoicing and others have innovated word-final devoicing [24].

Stage 1	/**sad** vixodʲit v drugoj **sad**/
(No devoicing)	[**sad** vixodʲit v drugoj **sad**]
	'The garden lets out onto another garden'
Stage 2	/**sad** vixodʲit v drugoj **sad**/
(Utterance-final devoicing)	[**sad** vixodʲit v drugoj **sat**]
Stage 3	/**sad** vixodʲit v drugoj **sad**/
(Word-final devoicing)	[**sat** vixodʲit v drugoj **sat**]

2.3. Artificial grammar experiment

In order to test the hypothesis that learners are biased toward word-based generalizations, as domain generalization implies, two artificial grammar experiments were carried out in [15]. Participants were exposed to constructed languages in which both voiced and voiceless obstruents occurred in syllable onset position but word-final obstruents were only observed in utterance-final position and were only voiced or voiceless (depending on experimental condition). Participants were thus implicitly given information about the voicing of word-final obstruents in utterance-final position, but no information about word-final obstruent voicing otherwise, a poverty of stimulus design ([25], [26], [27]). Results showed learning of the utterance-final devoicing (or voicing) generalization, and also extension of the learned pattern to word-final position even for words in utterance-medial position, supporting domain generalization.

3. The Russian prosodic word

3.1. The prosodic word as emergent

What facts motivate the Russian Pwd? As noted above, an important motivation comes from the facts of final devoicing and prepositions. According to this diagnostic, a preposition (or string of prepositions) plus one open-class lexical item constitute a prosodic word, e.g., [pəd mɐsˈkvoj]$_{\text{Pwd}}$, 'near Moscow'. Devoicing is Pwd-final, accounting for the lack of devoicing in /pod/ 'near'. Voicing assimilation among obstruents also occurs within the Pwd, as in [pɐt ˈpapəj]$_{\text{Pwd}}$ 'under papa' from /pod ˈpapa/.

Russian also has various enclitics, and these also trigger voicing assimilation, as in [ˈsog ʐʸi] 'juice (emphatic)' from /sok ʐʸe/, cf. [ˈsok tə] 'juice (topical)'; or [ˈsat tə] 'garden (topical)' from /sad to/, cf. [ˈsad ʐʸi] 'garden (emphatic)'. Such voicing assimilation does not occur as readily across the boundaries of open-class lexical items, suggesting that enclitics are also incorporated into the Pwd. However, final devoicing applies before these enclitics, as can be seen whenever an enclitic begins with a sonorant, e.g. [ˈsok lʲi] 'juice (interrogative)' and (crucially) [ˈsat lʲi] 'garden (interrogative)'. If devoicing is Pwd-final, then enclitics cannot be within the prosodic word in such examples.

Such considerations lead researchers to posit two 'word-like' prosodic levels for Russian, which we might call the 'prosodic word' and 'clitic group' or the 'minimal' and 'maximal' prosodic word, or which we might distinguish in some other way. One structure argued for is shown in Figure 1 (see [10], [28] for detailed arguments). In this structure the preposition /iz/ and noun /knʲig/ group together as a Pwd (notated ω). However, the interrogative enclitic /lʲi/ is outside of this Pwd (incorporated directly into the prosodic phrase), accounting for the final devoicing of /knʲig/.

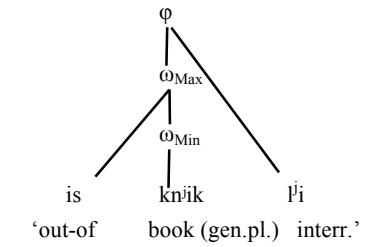


Figure 1: *Prosodic structure for proclitics vs. enclitics*

Though this structure succeeds in capturing the necessary distinctions, it raises the question: what explains the structure? From the point of view of prosodic theory the structure could just as easily be as in Figure 2.

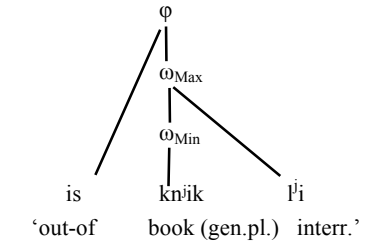


Figure 2: *Prosodic structure for proclitics vs. enclitics*

In the conventional approach to this problem, phonological facts like final devoicing are seen as (partly) determined by prosodic structure. As an alternative approach to answering the question above, we might instead try to derive the prosodic structure from the phonological facts. Final devoicing, as we have seen, arises historically when utterance-final devoicing (which is phonetically motivated) is generalized to the ends of all words. A word class that stands apart in not undergoing final devoicing, and which therefore motivates the Pwd, is the class of prepositions. Yet prepositions are a class of word that can never appear in utterance-final position in Russian, because they cannot be stranded (with marginal exceptions, see [29]).

The idea, then, is that domain generalization, assumed here to be the source of word-final devoicing, did not affect the class of prepositions, because Russian speakers had no experience of utterance-final prepositions and therefore no experience of devoiced prepositions. It is the array of facts this scenario engendered that leads the phonologist to posit the Pwd.

It does not follow from this idea that notions like the Pwd are imaginary or relevant only to linguists. Russian learners might well posit an organizational unit like the Pwd in response to the Russian facts, especially if this unit is useful in other ways. (See below.) However, the suggestion here does imply an understanding in which the Pwd is not, for example, an innate category provided by a universal grammar (see also [30] on this point). Rather, it *emerges* from a complex interaction of factors, including phonetic facts (providing the underpinning of utterance-final devoicing), phonologization (with domain generalization a key component of

phonologization, extending the generalization to open-class words in any position, and syntactic structure (explaining the exceptionality of prepositions).

3.2. Other evidence for the Russian prosodic word

What other facts motivate the Russian Pwd? There are at least two other noteworthy lines of evidence.

First, the Pwd is traditionally held to be the domain of lexical stress in Russian. Put another way, prepositions are part of the lexical stress domain: they do not carry stress independently; more importantly, a stress that 'belongs to' a following noun sometimes retracts onto the preposition itself, as in [ˈpod ruku] 'by the arm', compare [pəd ruˈkoj] 'at hand' [6], [7].

Second, the Pwd is relevant to the statement of vowel reduction facts [31], [28]. The vowels /o/ and /a/ reduce to [ə] when unstressed – compare [ˈ**ɡo**t] 'year' to [ɡədɐˈvoj] 'annual' (from /godoˈvoj/) and [ˈpr**a**f] 'law (gen.sg.)' to [prəvɐˈvoj] 'legal' (from /pravoˈvoj/. An exception is when these vowels immediately precede the stressed syllable of a word; in such cases the relevant syllable is much longer and the vowel is realized as something like [ɐ] [32], also seen in the examples above. This exception only applies within words, however: the word-final /o/ of /ˈmalo/ in /ˈmalo ˈskazano/ 'little said' reduces completely even though it precedes a stressed syllable in the following word: [ˈmal**ə** ˈskazənə]. It is significant, therefore, that pretonic reduction is to [ɐ] also for prepositions, e.g., [pɐt ˈpapəj]$_{\text{Pwd}}$ 'under papa' from /pod ˈpapa/, further supporting the analysis of such sequences as involving single Pwds.

That fact that at least three independent phonological processes – final devoicing, stress, and vowel reduction – apparently converge on the same Pwd analysis for preposition + word complexes presents a challenge for the view that Pwds emerge from the interaction of syntactic, phonetic, and phonological factors, as suggested here. If a domain such as Pwd is not given in advance but emerges as suggested earlier, how do these independent processes converge on the same domain? One possible answer is that speakers indeed posit Pwds based on facts like those of devoicing (or one of the other processes mentioned above), but that once posited, the Pwd can become relevant for, or even trigger, other phonological processes. In such a view, though not innate, Pwds are real, grammaticized organizational units, we might hold to the expectation that there are few such categories. However, this understanding of Pwds would be very hard to distinguish from the view that they are innate.

The alternative possibility is that different phonological phenomena lead to Pwd-like behavior independently, so that what counts as a 'Pwd' will depend on what phenomenon is in question; they need not converge on one answer. This is the view advocated in [30], for example, which argues that "prosodic domains are language-particular, intrinsic and highly specific properties of individual phonological rules or constraints" (though [30] allows that if enough processes appear to target the same domain they will have "a gravitating effect within the system, attracting phonological patterns which evolve in the course of sound change"). The discussion of Russian final devoicing here envisions one way that such language-specific organizational units might come about.

Further research is required to understand best how the Russian facts bear on these questions, but some evidence is already at hand that what we call a Pwd in Russian depends on which phenomenon we look at. One example comes from facts analyzed in [28]. A certain kind of compound can take stress in each member, e.g., *ˌbombə-uˈbʲeʐʸiʃːə* 'bomb shelter' and *ˌmʲedʸ-instʲiˈtut* 'medical institute'. This fact suggests an analysis of such compounds as involving two Pwds: [ˌbombə]$_{\text{Pwd}}$-[uˈbʲeʐʸiʃːə]$_{\text{Pwd}}$. Yet a word-final /o/ or /a/ in the first member of such compounds does not reduce to [ə] but to [ɐ] e.g., *ˌsaxərɐ-ˈvarnʲɪ* 'sugar refinary' from /ˌsaxar**o**-ˈvarnʲa/. The vowel reduction facts therefore suggest an analysis of such compounds as involving *one* Pwd: [ˌsaxərɐ-ˈvarnʲɪ]$_{\text{Pwd}}$. (Cf. [ˈmal**ə**]$_{\text{Pwd}}$ [ˈskazənə]$_{\text{Pwd}}$, discussed above.) Likewise final devoicing does not target the first word of such compounds, as the example *ˌmʲedʸ-instʲiˈtut* shows. Of course, these inconsistencies are a problem only if we expect all phonological processes to point to one and the same 'Pwd'.

4. Conclusions

The sources of evidence for something like the Pwd are diverse, including word-edge segmental phonology like final devoicing, but also facts about rhythm or stress, tone, apparent reference to morphosyntactic features, and effects of frequency. The discussion here has had nothing to say about the potential origins of phenomena other than word-final devoicing and the means by which they also converge on something like the prosodic word. But the point of this paper is that we might begin to make sense of the sometimes conflicting evidence about prosodic words, and explain aspects of their structure, if we view them as organizational constructs that emerge over time from the interaction of independently posited properties of a language. The hope is that this kind of thinking can be applied to these other sources of evidence as well.

5. Acknowledgements

The author thanks Gillian Gallagher, Bruce Hayes, Junko Ito, Pat Keating, Armin Mester, and Rachel Walker for discussion that benefitted this paper.

6. References

[1] Selkirk, E., "Prosodic domains in phonology: Sanskrit revisited", in Aronoff, M., and Kean, M.-L. [Eds.], Juncture, 107-129, Anma Libri, 1980.

[2] Selkirk, E., Phonology and Syntax: The Relation between Sound and Structure, MIT Press, 1984.

[3] Selkirk, E.: "On derived domains in sentence phonology", Phonology, 3, 371-405:1986.

[4] Nespor, M., and Vogel, I., "Prosodic domains of external sandhi rules", in Hulst, H.v.d., and Smith, N. [Eds.], The Structure of Phonological Representations, 222-255, Foris, 1982.

[5] Nespor, M., and Vogel, I., Prosodic Phonology, Foris, 1986.

[6] Gvozdev, A.N., O fonologicheskikh sredstvakh russkogo iazyka: sbornik statei, Izdatel'stvo akademii pedagogicheskikh nauk RSFSR, 1949.

[7] Jakobson, R., "Die Verteilung der stimmhaften und stimmlosen Geräuschlaute im Russischen", in Woltner, M., and Bräuer, H. [Eds.], Festschrift für Max Vasmer, Harassowitz, 1956.

[8] Halle, M., The sound pattern of Russian, Mouton, 1959.

[9] Vinogradov, V.V., ed., Grammatika russkogo iazyka, Izdatel'stvo Akademiia nauk SSSR, 1960.

[10] Padgett, J., "The role of prosody in Russian voicing", in Borowsky, T., Kawahara, S., Shinya, T., and Sugahara, M. [Eds.], Prosody matters: essays in honor of Lisa Selkirk, 181-207, Equinox, 2012.

[11] Keating, P., Cho, T., Fougeron, C., and Hsu, C.-S., "Domain-initial articulatory strengthening in four languages", in Local, J., Ogden, R., and Temple, R. [Eds.], Phonetic interpretation: papers in laboratory phonology vi, 145-163, Cambridge University Press, 1998.

[12] Cho, T., and Keating, P.: "Articulatory and acoustic studies on domain-initial strengthening in Korean", Journal of phonetics, 29.2, 155-190:2001.

[13] Varis, E.E., The Spanish feminine *El* at the syntax-phonology interface, Ph.D. dissertation, University of Southern California, 2012.

[14] Revithiadou, A., "The phonological word", in Van Oostendorp, M., Ewen, C., Hume, E., and Rice, K. [Eds.], The Blackwell companion to phonology, volume 2, Blackwell, 2011.

[15] Myers, S., and Padgett, J., "Domain generalization in artificial language learning", Ms., UT Austin and UC Santa Cruz, 2014.

[16] Wackernagel, J., Altindische grammatik. Band I: lautlehre (reprinted from the 1896 edition), Vandenhoeck and Ruprecht, 1957.

[17] Hyman, L., "Word demarcation", in Greenberg, J. [Ed.], Universals of human language, volume 2: phonology, 443-470, Stanford University Press, 1978.

[18] Westbury, J.R., and Keating, P.A.: "On the naturalness of stop consonant voicing", Journal of linguistics, 22, 145-166:1986.

[19] Lisker, L., Abramson, A., Cooper, F., and Schvey, M.: "Transillumination of the larynx in running speech", Journal of the Acoustical Society of America, 45, 1544-1546:1969.

[20] Shadle, C., "The aerodynamics of speech", in Hardcastle, W., and Laver, J. [Eds.], The handbook of phonetic sciences, 33-64, Blackwell, 1997.

[21] Steriade, D., "Phonetics in phonology: the case of laryngeal neutralization", Ms., UCLA, 1997.

[22] Pierrehumbert, J., "Probabilistic phonology: discrimination and robustness", in Bod, R., Hay, J., and Jannedy, S. [Eds.], Probabilistic linguistics, 177-228, MIT Press, 2003.

[23] Edwards, J., Beckman, M., and Munson, B.: "The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition", Journal of speech, language, and hearing research, 47, 421-436:2004.

[24] Jassem, W., and Richter, L.: "Neutralization of voicing in Polish obstruents", Journal of phonetics, 17, 317-325:1989.

[25] Wilson, C., "Experimental investigation of phonological naturalness", in Garding, G., and Tsujimura, M. [Eds.], Proceedings of WCCFL 22, 101-114, Cascadilla Press, 2003.

[26] Wilson, C.: "Learning phonology with substantive bias: an experimental and computational study of velar palatalization", Cognitive science, 30.5, 945-982:2006.

[27] Finley, S., and Badecker, W.: "Artificial language learning and feature-based generalization", Journal of memory and language, 61, 423-467:2009.

[28] Gouskova, M.: "The phonology of boundaries and secondary stress in Russian compounds", Linguistic review, 27, 387-448:2010.

[29] Gribanova, V., Composition and locality: the morphosyntax and phonology of the Russian verbal complex, Ph.D. dissertation, UC Santa Cruz, 2010.

[30] Schiering, R., Bickel, B., and Hildebrandt, K.A.: "The prosodic word is not universal, but emergent", Journal of linguistics, 46, 657-709:2010.

[31] Gribanova, V., "Phonological evidence for a distinction between Russian prepositions and prefixes", in Zybatow, G., Lenertová, D., Junghanns, U., and Biskup, P. [Eds.], Studies in formal Slavic phonology, morphology, syntax, semantics and information structure: proceedings of the 7th European Conference on Formal Description of Slavic Languages, Leipzig, 2007, 383-396, Peter Lang, 2009.

[32] Padgett, J., and Tabain, M.: "Adaptive Dispersion Theory and phonological vowel reduction in Russian", Phonetica, 62, 14-54:2005.

Local and Global Acoustic Correlates of Information Structure in Bulgarian

Bistra Andreeva [1], *Jacques Koreman*[2], *William Barry*[1]

[1] Computational Linguistics & Phonetics, Saarland University, Germany
[2] Department for Language and Literature, NTNU, Norway
{andreeva,wbarry}@coli.uni-saarland.de, jacques.koreman@ntnu.no

Abstract

In this study the local and global prosodic exponents of information structure are examined in the production of six Bulgarian question-answer elicited sentences under different focus conditions (broad focus and non-contrastive and contrastive narrow focus). Local cues are the phonetic properties of the nuclear accented syllables, while global cues reflect broader phonetic patterns in the intervals before and after the nuclear accented syllable, which in some cases vary independently of the tonal accent. Results show that speakers consistently discriminate broad and narrow focus by both local and global acoustic cues. Contrastive and non-contrastive accents are differentiated exclusively by local cues, but only when the focus is early in the sentence.

Index Terms: information structure, prosody, local and global cues, Bulgarian

1. Introduction

Most languages employ prominence-giving mechanisms to mark the relative informational importance of particular words in a phrase, often combined with word order and special lexical items or syntactic constructions. It is common to distinguish three elements of information structure (IS, e.g. [19]): 'topic' (the subject matter, on which new information is to be offered), 'focus' (the new information offered) and the 'given information' (information given previously or assumed to be known). These elements can be realized prosodically by means of a 'topic accent', a 'focus accent' or by 'de-accentuation'. At some basic production level, the speaker invests more effort in accentuated words compared to the words conveying given information, with the consequent acoustic effects of greater duration and intensity, higher or changing fundamental frequency (F0) and in some way more distinct spectral properties [10, 17, 24, 33]. However, there is evidence that languages differ in the amount each of the acoustic dimensions changes under accentuation [4, 25, 26] and there is considerable debate about which properties are used by the listener to identify prominent words or syllables. Pitch (measured as F0) is often seen as dominant [28, 14, 18], but duration [9, 20], intensity [9, 24, 35] and even voice quality [33] have also been singled out as important if not dominant determinants of perceived prominence.

Depending on the information provided by the pre-context, the focused part of a phrase can be restricted to one word – 'narrow focus' – or extend over much of the phrase – 'broad focus'. Within 'narrow focus', there is considerable disagreement in the literature about whether 'contrastive' and 'non-contrastive' focus are two distinct IS categories. Clearly, the context may or may not specify a semantic entity to which the focused word is in explicit contrast, providing a textual basis for a distinction. However, e.g. Rooth [31] sees an implicit contrast in any narrow focus; any expression has two semantic representations: the meaning of the expression itself and a set of alternatives. In the case of explicit contrast the alternative is known, but for Rooth the meaning of the expression does not change if the alternatives are not explicit.

Clear prosodic evidence for or against a contrastive – non-contrastive distinction is not apparent from the literature. Some have argued that there is no difference [16, 12, 34], while others have found evidence and argued that some acoustic features differ between contrastively vs. non-contrastively focused elements [13, 27, 7, 23].

Somewhat surprisingly perhaps, there is also disagreement about the reliability of the broad – narrow focus distinction. Of course, acceptance of the same utterance following both a pre-context cueing narrow focus and one cueing broad focus can only occur when the narrow focus is on the final lexical item. But given this condition, equal acceptance has been shown in several studies [11, 21, 37], while others claim that their subjects have consistently been able to make a distinction [8, 32].

In this paper, the prosodic exponents of broad focus and of non-contrastive and contrastive narrow focus are examined in the Sofia variety of Contemporary Standard Bulgarian.

Important factors in the realization of the information structure in Bulgarian utterances are:

- *word order*, remarkably flexible and discourse conditioned, as in all Slavic languages;
- morphological category of *definiteness*, unusual in the Slavic language family;
- *clitic replication* of nominal material;
- intonation.

Avgustinova [5] models the IS of Bulgarian utterances as interplay of the first three factors, while Miševa [29] and Nikov & Miševa [22] address the role of intonation. They experimentally investigated the regularities of F0 changes expressing phonetic prominence presented in terms of the traditional theme-rheme partitioning of the sentence. They conclude that the linguistically relevant phonetic characteristic of the given material (theme) is simply the absence of accentual prominence, i.e. de-accentuation. New material (rheme) shows the same intonational pattern in narrow and broad focus, but the accentual contrast between the prominent and the surrounding syllables is greater in narrow than in broad focus.

Andreeva et al. [1], Andreeva [2] and Avgustinova & Andreeva [6] adopted the terminology used in the Information Packaging approach in [36], where the basic focus-ground (cf. rheme and theme) articulation of the utterance is further refined by dividing the ground into link (what the focus is about) and tail (how the focus fits in the context). Contrary to the findings in [22, 29] they report that the underlying (phonological) pitch accent pattern for the thematic material is L*+H. Differences in the particular phonetic realizations depend on how the theme is realized on the surface, i.e. as a link (non-final in the intonational phrase) or a tail (final). In the link (pre-nuclear) the underlying pattern is realized phonetically as a gliding (slow) F0 rise from a low target within the accented syllable up to the next syllable (if there is enough syllabic ma-

terial), otherwise only within the syllable itself. In the tail (post-nuclear) the underlying pattern is not realized phonetically, i.e. there is a phonological rule deleting all pitch accents after the nuclear tone. In the opposition narrow vs. broad focus the underlying H* for the nucleus is realized with an emphasis [+raised peak] in the marked member of the opposition (i.e. narrow focus). In the case of a contrastive narrow focus, the phonetic realization of the shape of the underlying H* is also different, namely H*> [+raised peak; +delayed peak].

In this article, the question we would like to address is whether Standard Bulgarian distinguishes between different types of focus: a) non-contrastive and contrastive narrow focus, and b) broad and narrow focus. We first investigate the local acoustic cues in the nuclear syllable in terms of duration, F0 and intensity. Bruce [13] claimed that the focus domain is larger than the focused constituent and can affect the prosodic-acoustic realization of the whole sentence. Therefore, we shall also investigate the global effects of the IS on duration, F0 and intensity in the part of the utterance preceding and following the nuclear accent.

2. Material and Methods

The Bulgarian data that were used in this study were taken from an existing speech corpus consisting of read speech for several languages [4]. The stimulus material consisted of sentences with a fixed, canonical word order *subject* < *verb* < *direct object* < *indirect object* < *oblique*. This increases the role of prosody as an information-structuring factor, allowing us to focus on the acoustic correlates of different focus types. There were two critical words (CWs) in the sentence which could be realized with prominence, one early (CW1) and one late in the sentence (CW2). For each sentence, a number of questions were devised to elicit a) a *broad-focus* response, b) a response with a *non-contrastive narrow focus* on the early and c) on the late CW and d) a *contrastive focus* on the early and e) on the late CW. The sentences (with the critical words underlined) are:

1. Димо Данев гледа две деца.
   Dimo Danev gleda *dve* detsa.
   Dimo Danev looks after two children.
2. Бате Стефан взе седем книги.
   Bate Stefan vze sedem knigi.
   The elder Brother Stefan has taken seven books.
3. Играх на дама без кака ти.
   Igrax na dama bez kaka ti.
   I played draughts without your older sister.
4. Бате Мани пи тъмна бира.
   Bate Mani pi tămna bira.
   The elder Brother was drinking dark beer.
5. Дим Данев пя три пъти.
   Dim Danev pja tri păti.
   Dim Danev has sung three times.
6. Кака Нина търси черен хляб.
   Kaka Nina tărsi čeren xljab.
   The elder sister Nina is looking for dark bread.

The Bulgarian data corpus consists of 1080 sentences in total (6 speakers x 6 sentences x 5 focus conditions x 6 repetitions). In this article we present analysis results for local measurements in the CW for all sentence repetitions. We also present a more detailed analysis of the global prosodic patterns in the entire sentence for the first three of the six available repetitions.

2.1. Recordings and processing

Six regionally homogeneous speakers of Contemporary Standard Bulgarian as spoken in Sofia (3 female, 3 male) were recorded in a sound-treated studio. They read aloud each of the above sentences from a PowerPoint presentation in response to pre-recorded questions. The sentences and the questions eliciting different focus responses were pseudo-randomized and offered to the informants in six blocks, resulting in six repetitions of each sentence for each focus condition. The subjects were paid for participation.

The recordings were made using an AKG C420IIIPP headset on a Tascam DA-P1 DAT recorder and transferred digitally via the optical channel to a PC using the Kay Elemetrics MultiSpeech speech signal processing program.

Segmentation, labelling with SAMPA and further processing were done using the Kiel XASSP speech signal analysis package. Six labelling assistants were allocated different sentences (to maximize labelling consistency across conditions within each sentence) and segmentation problems were regularly discussed and decided with the authors at group level. In addition to the segmental labelling the pitch accents were also labelled by the first author, using BG-ToBI [2], with the peak alignment of the L(ow) and H(igh) targets explicitly specified. The positions of the F0 maxima and minima were double-checked by an automatic procedure for which the Praat pitch tracker was used.

2.2. Acoustic measurements

Local and global acoustic measures were calculated using praat scripts and operationalized as described in the two following subsections.

2.2.1. Local measurements

Local measurements of duration, F0 and intensity were made in the CWs in all the sentences read aloud by the informants.

a) Duration

Durations were measured for the stressed syllables of the CWs. The durations of the vowels in these syllables were also measured. Since all analyses and comparisons were carried out on individual sentences spoken in different focus conditions, it was possible to normalize all durational measurements as a percentage of the mean duration of the corresponding unit in the sentence.

b) Fundamental frequency

F0 was calculated as the mean fundamental frequency [Hz] across the syllable nucleus (vowel or syllabic sonorant) of the lexically stressed syllable of the CW. These values were also normalized by expressing them as percentages of the mean overall F0 of the sentence.

As a measure of peak alignment, the above absolute temporal distance from the F0 peak to syllable onset and rhyme onset were calculated. In order to compensate for the varying segmental durations on peak alignment, the above absolute measures were converted to relative measures, taken as a proportion of syllable and rhyme durations.

c) Energy

Energy was measured in two ways. First, as the mean intensity [dB] of the stressed vowel in the CW. These intensity values were normalized by subtracting the sentence intensity. Second, energy was measured as the spectral balance in the vowel.

This was computed as the difference in energy between the 70-1000 Hz and 1200-5000 Hz frequency bands.

2.2.2. Global measurements

Global measurements of duration, F0 and energy were made for the first three sentence repetitions by each speaker in each condition (focus type x sentence).

a) Duration

Durations were measured for the beginning of the sentence (sb) up to the focused syllable and for the sentence end (se) starting from the end of focused syllable. The values were normalized for speaking rate by calculating the percentage of the total sentence duration.

Since the number of syllables in the sentences varies, the tempo of sb and se were computed by dividing their duration by the number of syllables in the interval.

b) Fundamental frequency

In addition to mean F0 and peak alignment (section 2.2.1), the minimum F0 value preceding (L) and following the peak (Lpost) was measured, and the pitch excursion between the preceding F0 minimum and the peak (LH) and between the peak and the following F0 minimum (HLpost) was computed (s. Figure 1).

Individual F0 differences were removed by converting the obtained measurements to semitones by means of the following formula:

39.863 * log10(Maximum/Minimum)

Mean F0 values for sb and se were also computed and normalized by converting them to percentages of the sentence mean.

c) Energy

The intensity of sb and se were also measured and normalized using the same procedure as for the stressed vowels.

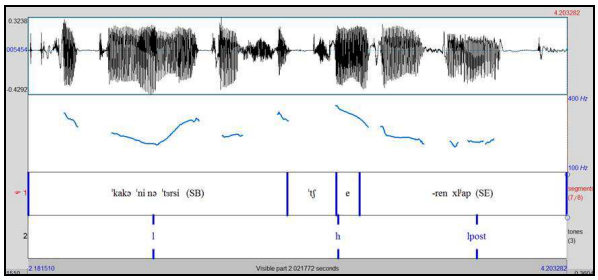


Figure 1: *Labeling example (sentence 6, narrow non-contrastive focus on CW2).*

2.3. Statistical analysis

The effect of focus condition was analyzed as a within-subjects factor separately for the local and global measurements in a mixed between-within MANOVA, with subject as a between-subjects factor. We report univariate tests with Greenhouse-Geisser estimates of F. These were verified with the multivariate Pillai's trace statistic; cells were equal in size. Separate Bonferroni post-hoc tests were carried out, if appropriate. The confidence level was set at α=0.05.

3. Results

In our data, the narrow focus is realized as (L+)H* pitch accents (except by speaker SP6). When the focus is realized on CW1 the H target is mostly reached close to the end of the accented syllable (93 %); when the focus is realized on CW2 die H target is reached close to the beginning of the accented syllable (85%). In broad focus condition, we observe H+!H*/L* with early peak alignment. Speakers vary as to their preferred choice of phonologically specified accent types and their phonetic realization. SP6 exclusively uses downstepped nuclear accents (H+!H*) in the narrow focus condition regardless of the position within the sentence, SP5 has a strong preference for downstepped nuclear accents (H+!H*) in the narrow focus condition on CW2 and speakers 1, 2, 4 and 5 show a preference for late peak alignment in the contrastive focus on CW1. The number of the pitch accents with early and late peak alignment used in the different focus conditions is summarized in Table 1.

Table 1: Distribution early versus late peak alignments (left-hand column) per focus condition and speaker (note: only 2 sentences were analyzed for broad focus).

		speaker						total
		SP1	SP2	SP3	SP4	SP5	SP6	
EARLY	CW2 broad	6	6	6	6	6	6	36
	CW2 nc	18	18	17	17	18	18	106
	CW2 c	16	17	18	13	18	18	100
	CW1 nc	7	5	10	2	0	16	40
	CW1 c	2	4	12	4	1	16	39
total		49	50	63	42	43	74	321
LATE	CW2 broad	0	0	0	0	0	0	0
	CW2 nc	0	0	1	1	0	0	2
	CW2 c	2	1	0	5	0	0	8
	CW1 nc	11	13	8	16	18	2	68
	CW1 c	16	14	6	14	17	2	69
total		29	28	15	36	35	4	147

In Bulgarian broad focus sentences, each content word is accented. In our data, CW2 is the last content word in sentences 3 and 5, while it is followed by another content word in the remaining sentences. To determine whether the acoustic realization of sentences in which only the object (CW2) carries a narrow focus differs systematically from those in which the entire event is focused (broad focus), we analyze sentences 3 and 5 separately from the other sentences. To investigate whether speakers prosodically differentiate non-contrastive and contrastive narrow focus we analyze all sentences, excluding the broad focus conditions. Sentences containing an early focus (on CW1) are analyzed separately from those containing a late focus (on CW2).

3.1. Local acoustic correlates of IS

The results from the statistical analysis of the local acoustic measurements (see section 2.2.1) for all focus conditions are summarized in Table 2.

3.1.1. Broad vs. Narrow

When the nuclear accent falls on CW2, both focus condition ($F[2, 90] = 29.739$, $p<0.001$) and speaker ($F[5, 90] = 35.307$, $p<0.001$) have a significant effect on the peak alignment. Moreover, there is a significant interaction between the two factors ($F[10, 90] = 5.012$, $p<0.00$). Speakers 1–4 align the F0 peak substantially earlier in the broad focus condition than in narrow focus, while speakers 5 and 6 do not differentiate between the two focus conditions.

Broad focus differs from narrow focus in that it has shorter syllable durations ($F[1,724; 51.717] = 211.658$, $p<0.001$), lower vowel intensity ($F [1.458; 41.117] = 539.372$, $p<0.001$), greater spectral tilt in the vowel ($F [2.467; 69.081] = 32.807$,

p<0.001) and a lower F0 in the vowel (F [1.720; 43.005] = 340.662, p<0.001). This F0 difference reflects the use of a different nuclear accent types: H+!H*/L* for broad and (L+)H* for narrow focus.

Table 2: *Main effects for focus condition and subject on local acoustic measurements and interactions (*** p<0.001)*

parameter	focus cond.	subject	interaction
nuclear accent on CW2 (broad vs. late)			
peak alignment	***	***	***
syll. duration	***	n.s.	n.s.
vowel intensity	***	***	***
vowel SpecTilt	***	***	***
vowel F0 mean	***	***	***
nuclear accent on CW1 (contr. vs. non-contr.)			
vowel duration	***	n.s.	n.s.
syll. duration	***	n.s.	n.s.

3.1.2. Contrast vs. Non-Contrast

Contrastive and non-contrastive focus is realized identically on CW2. When the focus is realized on CW1, the pitch accents are identical, but speakers produce a systematically longer vowel (F [2,218; 66.526] = 42.542, p<0.001) and syllable durations (F [2,281; 60.636] = 267.788, p<0.001) in the contrastive focus condition.

3.2. Global acoustic correlates of IS

Table 3 summarizes the results from the statistical analysis of the global acoustic measurements for all focus conditions (see section 2.2.2).

Table 3: *Main effects for focus condition and subject on global acoustic measurements and interactions (*** p<0.001)*

parameter	focus	subject	interaction
pre-nuclear accent on CW1 (broad vs. late)			
vowel intensity	***	n.s.	n.s.
nuclear accent on CW2 (broad vs. late)			
excursion LH	***	***	***
excursion HLpost	***	***	***
tempo pre-nuclear (sb)	***	***	n.s.
intensity pre-nuclear (sb)	***	***	***
intensity post-nuclear (se)	***	***	n.s.

3.2.1. Broad vs. Narrow

Considering non-local (global) effects, we also investigated the realization of CW1 for broad- vs. narrow-focus differences when the nucleus is on CW2. Although the CW1 was not deaccented in narrow focus (compare also [1, 2, 6]), and pitch accents realized in broad and narrow focus conditions were identical (L*+H), we observe a difference in *intensity* in the pre-nuclearly accented CW1 vowel, with a higher vowel intensity in broad than in narrow focus (F [1.717; 51.507] = 631.053, p<0.001), i.e. a measureable weakening of the pre-context in narrow focus on CW2.)

With respect to the pitch excursion (see section 2.2.2), a main effect was found for focus condition for the L-H excursion (F[2, 89] = 33.948, p<0.001) as well as for the H-Lpost excursion (F[2, 89] = 12.607, p<0.001), with larger excursions for narrow focus. There was also a main effect for speaker, both for the L-H excursion (F[5, 89] = 106.959, p<0.001) and for the H-Lpost excursion (F[5, 89] = 122.041, p<0.001). Focus and speaker also interacted significantly (F[10, 89] = 4,932, p<0.001), with only speakers 1-4 differentiating between the broad and narrow focus.

In the broad focus condition, the tempo in the pre-nuclear interval (sb) is lower (F[2, 90] = 6.662, p<0.01) and the intensity is higher than in narrow focus (F[2, 90] = 1.562, p<0.01), while intensity for the post-nuclear interval (se) is lower than in narrow focus (F[2, 90] = 12.582, p<0.001).

Speakers also differed significantly (tempo sb: F[5, 90] = 14.868, p<0.001; intensity sb: F[5, 90] = 55.567, p<0.01; intensity se: F[5, 90] = 11.909, p<0.001). An interaction between speaker and focus condition is only found for intensity in the pre-nuclear interval sb (F[10, 90] = 3.113, p<0.001). Again, speakers 5 and 6 do not differentiate between broad and narrow focus.

3.2.2. Contrast vs. Non-Contrast

No differences were found between the global measurements for contrast versus non-contrast, independent of the position of the nuclear accent.

4. Discussion and Conclusions

We investigated the prosodic realizations of information structure categories in Bulgarian. With regard to the difference between non-contrastive and contrastive focus, we observed that contrastive focus was marked more prominently than non-contrastive focus only locally and only in terms of vowel and syllable duration, when the CW occurs in the first half of the utterance. These results are not captured by a standard ToBI annotation.

With regard to the difference between broad focus and narrow focus on CW2, it was found that both local and global parameters were used. More specifically, narrow-focused syllables in CW2 were consistently realized with a longer duration, later peak alignment (but still early in the syllable), greater F0 excursions and higher energy than syllables with broad focus (local measures). This finding is not surprising, since all subjects but one used different pitch accent types to signal narrow vs. broad focus: (L+)H* vs. H+!H*/L*, respectively.

More important for the issue addressed in this study are the differences found in the global measurements. In agreement with results of previous research [1, 2, 6] no deaccentuation was found for narrow focus on CW2 in pre-nuclear position. Broad and narrow focus are not distinguished by accent type on CW1 (L*+H for both), nor is there a difference in global F0. However, the thematic, pre-nuclear interval (sb) in the narrow focus condition differs from the rhematic, pre-nuclear interval in the broad focus condition in terms of global measures. In responses with broad focus the interval preceding a focused syllable (sb) has a longer duration and a higher intensity than responses with narrow focus on CW2. This intensity difference is also found for the pre-nuclear CW1 vowel alone. Also, the interval following the nuclear accent has a lower intensity in responses with broad focus than responses with narrow focus on CW2. This finding is consistent with the observed post-nuclear vowel devoicing in broad focus conditions observed in [3].

To conclude, the all-important function of intonation, namely to transmit the relative weighting of information in speech communication, cannot be captured by a purely phonological description of realized accent types. Crucially, the IS-related patterns of phonetic prominence which are revealed in this study show a complex interplay between phonological categories and the local and global phonetic signal properties.

5. References

[1] Andreeva, B., Avgustinova, T. and Barry, W.J. (2001). Link-associated and focus-associated accent patterns in bulgarian. Gerhild Zybatow, Uwe Junghanns, Grit Mehlhorn and Luka Szucsich (eds.), Current Issues in Formal Slavic Linguistics, 353-364, Peter Lang: Frankfurt am Main.

[2] Andreeva, B. (2007). *Zur Phonetik und Phonologie der Intonation der Sofioter-Varietät des Bulgarischen*, PHONUS 12, Saarbrücken: Institute of Phonetics, University of the Saarland, PhD theses.

[3] Andreeva, B. and Koreman, J. (2008). The status of vowel devoicing in Bulgarian: phonetic or phonological? T. Zybatow et al. (eds.), *Formal Description of Slavic Languages: The Fifth Conference*, 81-91. Peter Lang: Frankfurt am Main.

[4] Andreeva, B., Barry W. and Koreman J. (accepted). A Cross-language Corpus for Studying the Phonetics and Phonology of Prominence. *The 9th edition of the Language Resources and Evaluation Conference* (LREC 2014), 26-31 May, Reykjavik, Iceland.

[5] Avgustinova, T. (1997). *Word Order and Clitics in Bulgarian*. Saarbrücken Dissertations in Computational Linguistics and Language Technology. Volume 5.

[6] Avgustinova, T. and Andreeva, B. (1999). Thematic Intonation Patterns in Bulgarian Clitic Replication. *Proc. of The XIVth International Congress of Phonetic Studies* (ICPhS'99), San Francisco, 1501-1504.

[7] Bartels, C., Kingston, J., (1994). Salient pitch cues in the perception of contrastive focus. Boach, P., Van der Sandt, R. (eds.), *Focus & Natural Language Processing, Proc. of J. Sem. conference on Focus. IBM Working Papers*. TR-80, 94–106.

[8] Baumann, S., Grice, M., and Steindamm, S. (2006). Prosodic Marking of Focus Domains - Categorical or Gradient? In *Proc. of Speech Prosody*, Dresden, Germany, 301-304.

[9] Beckman, Mary E. (1986). *Stress and Non-Stress Accent*. Netherlands Phonetic Archives, Series No. 7. Foris.

[10] Bertinetto P. M. (1981). *Strutture prosodiche dell'italiano.* Firenze: Accademia della Crusca.

[11] Birch, S. and Clifton, C. (1995) Focus, accent, and argument structure: effects on language comprehension. *Language and Speech*, 38 (4), 365-391.

[12] Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language*, 37, 83-96.

[13] Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: CWK Gleerup.

[14] Cooper, W., Eady, S. & Mueller, P. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of Acoustical Society of America*, 77(6), 2142-2156.

[15] Couper-Kuhlen, E. (1984). A new look at contrastive intonation., *Modes of Interpretation: Essays Presented to Ernst Leisi*, Watts, R., Weidman, U. (Eds.) Gunter Narr Verlag, 137–158.

[16] Cutler, A. (1977). The Context-Independence of "Intonational Meaning". *Chicago Linguistic Society (CLS 13)*, 104-115.

[17] Dauer, R. (1987). Phonetic and phonological components of language rhythm. *Proc. of the 11th International Congress of Phonetic Sciences*, Vol. 5, 447-450. Tallinn: Estonian Academy of Sciences.

[18] Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *Journal of the Acoustical Society of America*, 80, 402-415.

[19] Féry, C. and Krifka, M. (2008). Information Structure: Notional Distinctions, Ways of Expression. P. v. Sterkenburg (ed.), *Unity and diversity of languages*, Amsterdam: John Benjamins, 123-136.

[20] Fry, D. B. (1955). Duration and Intensity as Physical Correlates of Linguistic Stress. *Journal of the Acoustical Society of America*, 27, 765–768.

[21] Gussenhoven, C. (1983*). Testing the reality of focus domains. Language and Speech*, 26, 61–80.

[22] Hirst, Daniel, Di Cristo, Albert (Eds.). (1998). *Intonation Systems: A Survey of 20 Languages.* Cambridge University Press, Cambridge.

[23] Ito, K. Speer, S. R. and Beckman, M. E. (2004). Informational status and pitch accent distribution in spontaneous dialogues in English, In *Proceedings of the International Conference on Spoken Language Processing*, Nara: Japan, 279-282.

[24] Kochanski, G., E. Grabe, and J. Coleman (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America* 118, 1038-1054.

[25] Koreman, J., Andreeva, B. and Barry, W. (2008). Accentuation cues in French and German, P. A. Barbosa, S. Madureira, & C. Reis (eds.), *Proceedings of the 4th International Conferences on Speech Prosody* , 613-616, Campinas: Editora RG/CNPq.

[26] Koreman, J., Andreeva, B., Barry, W., van Dommelen, W., Sikveland, R.-O. (2009). Cross-language differences in the production of phrasal prominence in Norwegian and German, Martti Vainio, Reijo Aulanko, and Olli Aaltonen (eds.), *Nordic Prosody*, Proceedings of the Xth Conference, Helsinki 2008, Frankfurt: Peter Lang, 139-150.

[27] Krahmer, E., & Swerts, M. (2001). On the alleged existence of contrastive accents. *Speech Communication*, 34, 391-405.

[28] Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32(4), 451-454.

[29] Miševa, A. (1991). *Intonacionna sistema na bălgarskija ezik*. Sofija: Bălgarska Akademija na Naukite.

[30] Peperkamp, S., E. Dupoux, and N. Sebastia´n-Galle´s (1999). Perception of stress by French, Spanish, and bilingual subjects. In: G. Olaszy & V. Orbán (eds.). Proceedings of Eurospeech 99, Vol. 6, 2683–2686. Budapest: ESCA.

[31] Rooth, M. (1992). A theory of focus interpretation. *Natural Language Semantics*, **1**, 75 –116.

[32] Rump, H. H., and Collier, R. (1996). 'Focus conditions and the prominence of pitch accented syllables. *Language and Speech*, 39, 1–17.

[33] Sluijter, A., V. van Heuven (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100, 2471-2485.

[34] 't Hart, J. Collier, R. & Cohen, A. (1990). *A perceptual study of intonation*. CambridgeUniversity Press, Cambridge.

[35] Turk, A. & Sawusch, J. (1996) The processing of duration and intensity cues to prominence. *Journal of the Acoustical Society of America*, 99, 3782-3790.

[36] Vallduví, E. and Engdahl, E. (1995). Information Packaging and Grammar Architecture. J. N. Beckman (ed.) *NELS* 25. Vol. 1. University of Pennsylvania: 519-533

[37] Welby, P. (2003). Effects of pitch accent position, type, and status on focus projection.*Language and Speech*, 46, 53 – 81.

Description of Polish speech rhythm using rhythm metrics and the time-delay approach: A comparative study

Agnieszka Wagner

Institute of Linguistics, Adam Mickiewicz University in Poznan, Poland

wagner@amu.edu.pl

Abstract

The goal of this study is to provide a multidimensional description of rhythmic structure of Polish utterances. For this purpose a time-delay approach proposed in [1] is applied and results of qualitative and quantitative analyses based on time-delay plots are compared with results obtained with selected rhythm metrics. The study shows that description that relies on a combination of rhythmic scores is inconclusive and difficult to interpret, because it does not account for rhythmic structuring nor grouping. The time-delay approach, on the contrary, appears to be very efficient in exploring short-time and long-term timing variability that determine Polish speech rhythm.

Index Terms: Polish speech rhythm, time-delay approach, meter, grouping, prominence, rhythm metrics

1. Introduction

1.1. Analysis and description of speech rhythm

For many years, research on speech rhythm was based on *isochrony* paradigm [2, 3] and concept of *rhythmic classification of languages* (i.e. the distinction between syllable-timed, stress-timed and mora-timed languages). The fact that instrumental studies failed to bring evidence for stress- and syllable-based isochrony caused that in rhythm research the focus moved from duration measurements of syllables and feet to investigation of phonemic and phonological factors which affect the timing of syllables and feet. Observed differences in vowel reduction, stress-based lengthening and syllable complexity between stress- and syllable-timed languages motivated the development of *rhythm metrics* – formulas that measure durational variability in consonantal (C) and vocalic (V) intervals. The most widely used metrics include: %V-ΔC [4], PVIs [5] and Varcos [6]. They have been applied in studies on rhythmic classification of languages [7], acquisition of timing patterns in L1 [8] and L2 [9, 10, 11, 12, 13, 14], detection of speech impairments [15] or dialect discrimination [16]. The proponents of rhythm metrics provided experimental results showing that they can be regarded as acoustic correlates of perceived speech rhythm and supporting the hypothesis of rhythmic classification of languages. However, rhythm metrics have been the subject of considerable criticism. First of all, it can be argued that what metrics describe is not rhythm, but timing of utterances [17]. Secondly, metric scores vary considerably within languages, because they are sensitive to a number of factors such as tempo, speech style or method of measurement [18, 19]. Thirdly, rhythmic classification based on the stress-timing vs. syllable-timing dichotomy applies to *some languages in some conditions* (for example, different metrics provide different classifications of the same language [18]) and there are languages, for example Polish, that can not be assigned to any class on the basis of the metric scores and are consequently labelled as "intermediate". At the same time, metrics are incapable of giving information on distinctive features of such intermediate rhythm. In the end, metrics such as PVIs or %V-ΔC describe single aspect of rhythm, i.e. variability in time domain, but "rhythm cannot be described as a one-dimensional property of speech, e.g. as more or less variable or more or less stress timed" ([1], p.145). Rhythm can be regarded as a perceptual impression of a *structure* consisting of more or less prominent (strong, weak) events (beats or syllables) which are *grouped* in a particular manner to form *perceptually distinct patterns* such as iambs or trochees. It is doubtful whether processes underlying rhythmic structuring and grouping can be explained by means of rhythm metrics, but there is evidence from [1] that they can be successfully explored using a multidimensional approach which takes into account short-term and long-term timing variability including, among others, relative timing of functionally distinct transitions (to stressed, to unstressed and to phrase final syllables), acceleration and deceleration tendencies, compensatory shortening and time shrinking (i.e. a psychoacoustic phenomenon which results in perception of decelerating sequences as isochronous). A multidimensional account of rhythm incorporates various levels of the prosodic hierarchy and dimensions other than duration/timing, because intensity, F0 and spectral features also constitute important correlates of prominence. The approach taken in this study makes it possible to analyze and describe speech rhythm in such a multidimensional manner. It is based on *time-delay plots* which provide "a useful tool in order to explore timing relations that are perceived as typical rhythms in speech" and which are "directly interpretable along similar rhythm-related dimensions as have been detected in typological analyses" ([1], p. 155). The time-delay approach seems to be particularly useful to investigate fine-grained timing differences related to rhythmic structure and grouping in non-prototypical – rhythmically "intermediate" or unclassified – languages, such as Polish.

1.2. Polish speech rhythm

As regards phonological properties, Polish can be regarded as rhythmically "mixed" – it has fixed lexical stress, no vocalic reduction in unstressed syllables and subtle stress-related and accentual lengthening, which are considered typical features of syllable-timed rhythm, whereas high phonotactic complexity and presence of compensatory shortening points to stress-timing. Former studies provided evidence for accentual lengthening of vowels [20, 21, 22] and some support for isochrony within narrow rhythm units [23]. The results of recent corpus-based analyses [24, 25] showed that accentual lengthening is limited to vowels and syllables associated with major prominences (phrase accents), whereas durational marking of minor prominences (which coincide with word stress) is very subtle, if any. In [26] overall intensity was regarded as the main acoustic correlate of stress and in [27] – pitch movements. In [25] intensity and F0 features correlated significantly only with major prominences. Prosodic phrase

boundaries in Polish are signaled most of all by increased duration of the phrase-final and the penultimate syllable and vowel (associated most of the time with phrase/nuclear accent), but F0 features also play a significant role [22, 29]. The results of a multidimensional analysis of Polish rhythm in [1] showed only subtle lengthening of stressed syllables, tendency for foot final lengthening, deceleration throughout the foot, some compensatory shortening effects (comparing binary and longer feet) and potential evidence for time shrinking phenomenon that can contribute to impression of isochrony. These results clearly show that Polish can not be easily classified based on the stress-timed/syllable-timed dichotomy. The accounts of Polish rhythm based on rhythm metrics are inconclusive: According to PVIs [5], Polish is close to syllable-timed languages, but, according to %V – ΔC, it is grouped with stress-timed English and Dutch [4].

1.3. Objectives of the study

The objective of the study is to compare two approaches to analysis of speech rhythm in Polish: rhythm metrics [4, 5, 6] and the time-delay approach [1], and to provide explanation of factors underlying the "mixed rhythm effect" in Polish, because existing descriptions are neither informative nor satisfactory in this respect. For this purpose the study explores short-term timing variability in the realization of functionally different prosodic transitions: to stressed, to unstressed, to phrase final syllables, and long-term timing characteristics of rhythmical grouping at various levels of prosodic hierarchy – feet (of different sizes) and prosodic phrases. Rhythmic characteristics of Polish are also compared cross-linguistically by referring to results reported in the literature.

2. Methodology

2.1. Speech data

The speech material includes recordings of a literary fairy tale "The teapot" (by H. Ch. Andersen), read by five speakers – all coming from Poznan and presenting the Poznan-Cracow pronunciation. The text consists of 19 phonetically and prosodically rich sentences (491 syllables). Recordings were carried out in a sound-treated booth, directly to a disk and with a sampling frequency of 16 kHz. The subjects were asked to read the text once (sentence after sentence), at their own pace. Sentences containing disfluencies or mispronunciations were re-recorded. The recorded material constitutes part of a speech database created for the purpose of studying speech rhythm in native and non-native Polish [30]. The whole speech material was subject to automatic phonetic transcription and alignment [31] which were verified and manually corrected following standard segmentation criteria. Syllable boundaries were determined as in [32]. Prosodic annotation consisted in labeling four levels of prominence (unstressed, stressed but unaccented, accented and nuclear accented) and two levels of phrasing (intermediate and intonational phrase) [33, 34, 35]. Annotation and duration measurements were carried out in *Praat*. For statistical analyses *Statistica 10* was used.

2.2. Rhythm metrics

Segmentation into vocalic and consonantal intervals was based on the phonetic transcription and alignment. All vowels were marked as vocalic intervals and all consonants (except for post-vocalic glides) – as consonantal intervals. A vocalic interval could contain a single vowel or 2-3 subsequent vowels, or a vowel followed by a glide. Intervals could span across syllable and word boundaries. As in [18] prepausal intervals were not excluded from measurements and segments separated by a pause were treated as two distinct intervals. For each sentence we calculated the following metrics:

- **%V** – the proportion of vocalic intervals, **ΔV** and **ΔC** – the standard deviation of the duration of vocalic and consonantal intervals respectively [4]
- **rPVI-V** (raw Pairwise Variability Index) and **nPVI-V** (vocalic normalized Pairwise Variability Index): the mean of the duration differences between successive C intervals and the mean of the duration differences between successive V intervals divided by the sum of the same intervals respectively [5]
- **VarcoV/VarcoC**: standard deviation of vocalic/consonantal interval duration divided by mean vocalic/consonantal duration [6]

2.3. Time delay approach

This approach relies on time delay plots which are used to visualize relative timing of *functionally different transitions*, e.g. to stressed, to unstressed and to phrase final syllables. In time-delay plots, the duration of syllable$_i$ is plotted on the x-axis against the duration of syllable$_{i+1}$ plotted on the y-axis. Time delay plots can be applied to explore both short-term (syllable-level) and long-term (foot- and phrase-level) timing variability, and offer the possibility of interpreting continuous data in both continuous and categorical manner. By subtracting the duration of syllable$_i$ from the duration of syllable$_{i+1}$, we get information whether the transition is locally *accelerating*, *decelerating* or *isochronous* (continuous description). Decelerating transitions are plotted above the diagonal, accelerating transitions below it and the isochronous ones – along the diagonal (Figure 1).

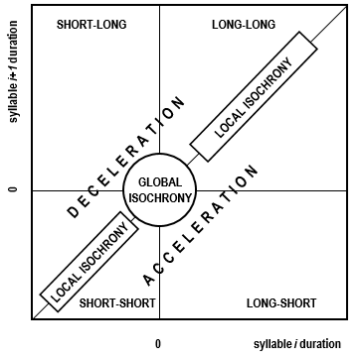


Figure 1: *Interpretation of a time-delay plot (based on [1]).*

Transitions can also be grouped into one of the four categories: short-short, long-long (locally isochronous transitions), short-long and long-short (an alternating rhythm). For example, if durations of *syllable$_i$* and *syllable$_{i+1}$* are both above/below the mean, the transition is categorized as long-long/short-short. A tendency towards global isochrony is indicated by concentration of data points in the center of the time-delay plot. The plots can be interpreted quantitatively, e.g. by performing one-factorial ANOVA with the duration difference *syllable$_{i+1}$* - *syllable$_i$* as dependent variable and transition type as predictor variable (see also [1]). On a higher

level of rhythmic-prosodic organization, i.e. foot level, time-delay plots can be applied to visualize relative timing relations within feet and to observe *compensatory shortening* which contributes to impression of isochrony in stress-timed rhythm.

3. Results

3.1. Quantification using rhythm metrics

Comparison of the metric scores obtained in the current study with those reported in [18] (both studies used the same method of consonantal and vocalic interval measurements) shows that from among six rhythmically different languages, i.e. German, English, Spanish, Italian, Greek and Korean, Polish is characterized by the least variability in vocalic interval duration – it has the lowest nPVI and VarcoV. In terms of amount of vocalic speech (%V) Polish is ranked in between languages traditionally considered stress-timed (German, English) and those regarded as rhythmically unclassified (Greek, Korean) and syllable-timed (Italian, Spanish). High values of the raw consonantal metrics, i.e. ΔC and rPVI, indicate similarity to stress-timed German and English. On the contrary, VarcoC ranks Polish in between Greek, syllable-timed Italian and Spanish on the one hand (low VarcoC – low variability in consonantal interval duration) and stress-timed German and English, and Korean on the other (high VarcoC).

As regards the effect of speech rate (measured in the number of C and V intervals per second, cf. [36]) on the metric scores, significant inverse correlations were found for ΔV, ΔC and rPVI, indicating instability of these metrics. For the purpose of quantification of the actual distances between Polish and the six rhythmically different languages in the rhythm space determined by the most stable metrics, i.e. VarcoV-VarcoC and %V-nPVI (the latter, contrary to %V-VarcoV, were not significantly correlated with each other), Euclidean distances were calculated (Table 1). As they show, according to %V-nPVI, Polish is close to syllable-timed Italian and Spanish, but VarcoV-VarcoC place Polish the closest to German.

Table 1. *Euclidean distances between Polish and six rhythmically different languages.*

rank	VarcoV-VarcoC	%V- nPVI
1	German (7,3)	Spanish (7,4)
2	Spanish (9,3)	Italian (7,6)
3	Italian (10,7)	Greek (11,2)
4	English (10,8)	Korean (12,4)
5	Korean (14,2)	German (13,6)
6	Greek (14,3)	English (17,9)

Unlike VarcoV-VarcoC, the combination of %V-nPVI provides some support for the rhythm typology – values of these metrics (Figure 2) and the Euclidean distances based on them separate syllable-timed Spanish and Italian from stress-timed German and English. The metrics do not provide strong evidence for an "intermediate" rhythm in Polish: while values of %V and VarcoC are indeed intermediate between languages traditionally regarded as stress- and syllable-timed, nPVI and VarcoV place Polish in a very different region from all the other languages (Figure 2).

Generally, it can be assumed that low nPVI and VarcoV in Polish reflect lack of reduction of unstressed vowels on the one hand, and lack of significant stress-based lengthening of vowels on the other (see also discussion in section 1.2).

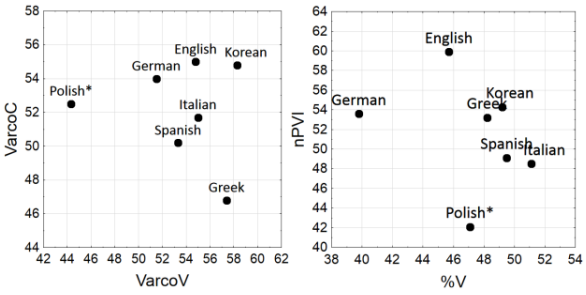


Figure 2: *Rhythm space determined by VarcoV-VarcoC (left) and %V-nPVI (right) for the six languages in [18] and Polish (*current study).*

As for VarcoC in Polish, its values (Figure 2) reflect the phonotactic structure of the utterances: high percentage (about 50%) of simple CV syllables (as in syllable-timed languages), high variation in complex syllable structures (indicating a tendency towards stress-timing) and frequency of very complex syllables even higher than in stress-timed languages.

3.2. Analysis with time-delay plots

3.2.1. Short-term timing variability

Table 2 presents concentrations of functionally different transitions in the four relative timing quadrants: Transitions to stressed syllables are concentrated mainly in the short-long and short-short quadrants, transitions to unstressed syllables – in the short-short quadrant, and transitions to phrase final syllables – in the long-long and short-long quadrants.

Table 2. *Concentrations of functionally different transitions in the four relative timing quadrants (based on z-score normalized syllable durations).*

transition	short-long	long-long	long-short	short-short
to stressed	37%	15%	16%	32%
to unstressed	21%	7%	23%	49%
to phrase final	29%	57%	3%	11%
total:	28%	18%	19%	35%

Generally, it seems that Polish, unlike English or French, *favors short-short sequences* the most. The predominance of short-short sequences in transitions to unstressed syllables, and long-long sequences in transitions to phrase final syllables, indicates a tendency towards *local isochrony* (like in syllable-timed Italian, [1]). At the same time, the high count of short-long transitions to stressed syllables, and to a lesser extent to phrase final syllables, indicates a tendency towards *alternation* (like in stress-timed English, [1]): Such timing patterns may contribute to the impression of an *intermediate* or *mixed* rhythm in Polish. Final lengthening is confined to the syllable at the edge of the phrase, but very often previous syllable is also lengthened (as indicated by high percentage of long-long transitions, see also Figure 3) due to co-occurrence with stress. One-factorial ANOVA showed significant *effect of transition type on relative timing of the sequence of syllables* expressed by the difference $syllabledur_{i+1}$ - $syllabledur_i$ (indicating local acceleration or deceleration): F=50.9, df=2, p<0.01. Post-hoc

comparisons revealed that this effect is due to a strict trend towards *deceleration in transitions to stressed* (mean=0.5, σ=1.3) *and phrase final syllables* (mean=0.45, σ=1.5) on the one hand, and *acceleration in transitions to unstressed syllables* (mean=-0.07, σ=1.1) on the other. Figure 3 shows relative timing of transitions from unstressed to stressed syllables, with a distinction between lexical stress, pitch accent and nuclear accent, and transitions to phrase-final syllables.

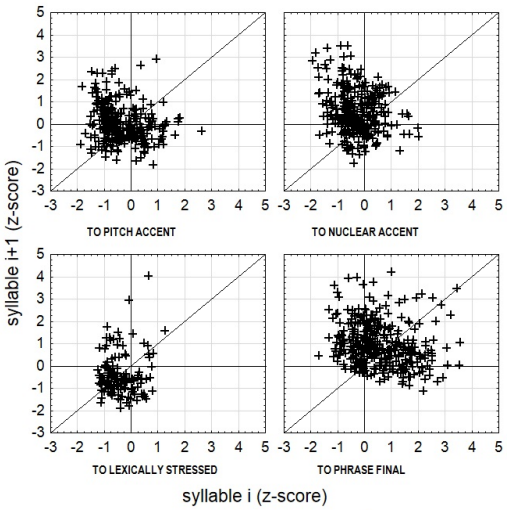


Figure 3: *Relative timing in functionally different transitions.*

Transitions to pitch accents have very similar distribution to the general one (when all prominent syllables are classified as stressed, see Table 2), whereas transitions to nuclear accents and lexically stressed syllables are concentrated mostly in the short-long and short-short quadrants respectively. These distributions indicate that durational marking is reserved mostly for nuclear accents – major prominences. One-factorial ANOVA showed significant differences in the relative timing between the three transition types, i.e. to lexically stressed, pitch accented and nuclear accented syllables: F=23, df=2, p<0.01. There is a strong *deceleration trend that increases with the level of prominence* from lexical stress (mean=0.02, σ=1) to nuclear accent (mean=0.86, σ=1.3), with pitch accents in between (mean=0.4, σ=1.2).

3.2.2. Long-term timing variability

The goal of the analysis of long-term characteristics of feet of different sizes is to "detect timing regularities that listeners can learn in order to form certain expectations concerning upcoming rhythmical events. Such long-term expectations are what we defined as *meter*" ([1], p. 162). It can be seen in Figure 4a, that in binary non-final feet, the stressed and the following unstressed syllable are almost identical in length, which indicates a tendency towards local isochrony. In this respect, Polish differs from both stress-timed English (-> increased duration on foot-initial stressed syllables) and syllable-timed French (-> a strong tendency towards foot-final lengthening) [1]. Binary phrase-final feet in Polish have a very similar timing pattern to that observed in French i.e., foot-final lengthening [1]. Ternary and quaternary feet have a tendency of acceleration after the foot-initial, stressed syllable and deceleration at foot boundary. Foot internal syllables (syl2 in ternary and syl3 in quaternary feet) are significantly shorter than foot-initial and foot-final syllables.

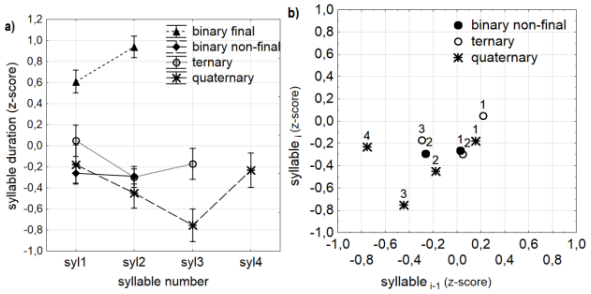


Figure 4: *Long-term timing patterns across different feet.*

These timing patterns are similar to those found in German longer feet [1]. Deceleration at foot boundary may cause the effect of *time shrinking* of the pre-final syllable and may lead to the *impression of isochrony and a lack of variation* in longer feet. Time shrinking may also concern stressed syllables in binary phrase-final feet. What can be traced in Figure 4b is the *compensatory shortening* – distribution of average relative durations of foot-internal syllables in binary non-final, ternary and quaternary feet shows tendency to shorten syllables with increasing foot length, but foot-final syllables are not affected by this phenomenon.

4. Discussion and conclusions

The results of the analysis with rhythm metrics indicated very low variability in vocalic interval duration (VarcoV and nPVI) in Polish comparing to six rhythmically different languages. In terms of variability in consonantal interval duration (VarcoC) and the amount of vocalic speech (%V) Polish can be regarded as *intermediate* between syllable- and stress-timed languages. Euclidean distances showed that according to %V-nPVI, Polish is close to Italian and Spanish, but VarcoV-VarcoC place Polish the closest to German. The description of Polish rhythm provided by the metrics is thus inconclusive and hard to interpret. On the contrary, time-delay approach appeared to be a very efficient method for visualization and analysis, both quantitative and qualitative, of the contribution of short-term and long-term timing variability to rhythmic structure and rhythmic grouping at different hierarchical levels (syllable, foot and phrase level). The time-delay analysis provided some evidence for "mixed" rhythm in Polish which is characterized by *as much local isochrony* (-> syllable-timing) *as alternation* (-> stress-timing). Speech rhythm in Polish is also determined by *deceleration* in transitions to *phrase-final and stressed syllables* (the trend increases with the level of prominence), and *acceleration* in transitions to *unstressed syllables*. The analysis of long-term timing variability brought some evidence of *time shrinking* and *compensatory shortening*, which may lead to the impression of syllable- and stress-based isochrony respectively. The analyses showed that rhythmic grouping is determined by *deceleration* at the foot and phrase boundary. As regards rhythmic structure in Polish, it is necessary to go beyond the dimension of timing, because the current study showed that durational marking of prominence is very weak, if any.

5. Acknowledgements

Work presented in the paper was supported from grant DOBR/0008/R/ID1/2013/03 by the National Centre of Research and Development in Poland.

6. References

[1] Wagner, P. (2008). The rhythm of language and speech: Constraining factors, models, metrics and applications. Habilitationsschrift, University of Bonn.

[2] Pike, K. L. (1945). The Intonation of American English. Ann Arbor: University of Michigan.

[3] Abercrombie, D. (1967). *Elements of general phonetics* (Vol. 203). Edinburgh: Edinburgh University Press.

[4] Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition 73*(3). 1-28.

[5] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, *7*(515-546).

[6] Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient forΔ C. *Language and language processing*, 231-241.

[7] Mairano, P. and Romano, A. (2011). Rhythm metrics for 21 languages. In *Proceedings of the XVIIth International Congress of Phonetic Sciences*, 17-21 August 2011, Hong Kong, China. 1318-1321.

[8] Payne, E., Post, B., Astruc, L., Prieto, P. & del Mar Vanrell, M. (2012). Measuring child rhythm. *Language and Speech, 55*(2), 203-229.

[9] White, L. & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics 3*(5). 501-522.

[10] Mok, P. P., & Dellwo, V. (2008). Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Proceedings of the Speech Prosody 2008 Conference* (pp. 423-426).

[11] Tortel, A., & Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. In *Proceedings of Speech Prosody 2010*.

[12] Ordin, M., Polyanskaya, L. & Ulbrich, Ch. (2011). Acquisition of Timing Patterns in Second Language. In *Proceedings of INTERSPEECH 2011*, 27-31 August 2011, Florence, Italy. 1129-1132.

[13] Kinoshita, N., & Sheppard, C. (2011). Validating acoustic measures of speech rhythm for second language acquisition. In *Proceedings of the XIth International Congress of the Phonetic Sciences* (Vol. 17, pp. 1086-1089).

[14] Li, A. and Post, B. (2012). L2 rhythm development by Mandarin Chinese learners of English. In *Proceedings of Perspectives on Rhythm and Timing*, 19-21 July 2012, Glasgow, UK.

[15] Liss, J., White, L., Mattys, S., Lansford, K., Lotto, A., Spitzer, S. & Caviness, J. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language, and Hearing Research* Vol.52 1334-1352.

[16] Leemann, A., Dellwo, V., Kolly, M. J., & Schmid, S. (2012). Rhythmic variability in Swiss German dialects. In *Proceedings of Speech Prosody 2012*, Shanghai, China.

[17] Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66*(1-2), 46-63.

[18] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*(3), 351-373.

[19] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *J. Acoust. Soc. Am 127*(3). 1559-1569.

[20] Richter, L. (1978). Wpływ pozycji w zestroju akcentowym na czas trwania głosek. (The effect of the position in the accent group on the phoneme duration). *Lingua Posnaniensia*, vol. 21.

[21] Imiolczyk, J., Nowak, I., Demenko, G. (1994). High intelligibility text-to-speech synthesis for Polish. *Archives of Acoustics 19* (2), pp. 161-172.

[22] Demenko, G. (1999). Analysis of Polish Suprasegmentals for needs of Speech Technology. UAM: Poznań.

[23] Richter, L. (1987). Modelling of the rhythmic structure of utterances in Polish. *Studia Phonetica Posnaniensia* 1, 91–125.

[24] Klessa, K. (2012). Polish segmental duration: selected observations based on corpus data. *Speech and Language Technology*, vol. 14/15, pp. 95-104.

[25] Malisz, Z., & Wagner, P. (2012). Acoustic-phonetic realisation of Polish syllable prominence: a corpus study. *Speech and Language Technology*, vol. 14/15, pp. 105-114

[26] Dłuska, M. (1933). Próba badań nad trwaniem spółgłosek polskich w zależności od brzmienia. (An attempt at investigating duration of Polish consonants depending on their quality). *Slavia Occidentalis* vol. 12, pp. 288-297.

[27] Jassem, W. (1962). Akcent języka polskiego (Accent of Polish).

[28] Klessa, K. (2006). Analysis of segmental duration for needs of speech synthesis in Polish. Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[29] Wagner, A. (2010). Acoustic cues for automatic determination of phrasing. In *Proceedings of Speech Prosody 2010*, 11-14 May 2010, Chicago, USA.

[30] Wagner, A. (2012). Speech rhythm in native and non-native Polish, In *Proc. of ISCA Workshop on Experimental Llinguistics*, Athens, 27-29 August 2012.

[31] Demenko G., Wypych M. & Baranowska E. (2003.) Implementation of Polish grapheme-tophoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis. *Speech and Language Technology*, vol. 7, pp. 79-96.

[32] Ostaszewska, D., & Tambor, J. (2000). Phonetics and phonology of contemporary Polish. Polish Scientific Publishers (PWN).

[33] Prieto, P., Vanrell, M. D. M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication, 54*(6), 681-702.

[34] Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. *Papers in laboratory phonology III: Phonological structure and phonetic form*, 7-33.

[35] Selkirk, E. O. (1995). Sentence prosody: Intonation, stress and phrasing. In Goldsmith, J., editor, Handbook of Phonological Theory. Oxford.

[36] Beňuš, Š. & Šimko, J. (2012). Rhythm and tempo in Slovak. In *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China.

11 Wednesday 4

Event-Related Potential investigation of Initial Accent processing in French

Marion Aguilera[1], *Radouane El Yagoubi*[2], *Robert Espesser*[3], *Corine Astésano*[1,3]

[1] U.R.I Octogone-Lordat (E.A. 4156), Université de Toulouse, UTM, Toulouse, France
[2] Laboratoire CLLE-LTC (UMR 5263), Université de Toulouse, UTM, Toulouse, France
[3] Laboratoire Parole & Langage (UMR 7309), Aix-Marseille Université, Aix-en-Provence, France

marion.aguilera@univ-tlse2.fr, yagoubi@univ-tlse2.fr, Robert.Espesser@lpl-aix.fr, astesano@univ-tlse2.fr

Abstract

This study investigates stress processing through the Event-Related brain Potential (ERP) technique. It aims at evaluating whether French listeners can perceive and discriminate the Initial Accent (IA) and whether IA is encoded in the phonological representation. Participants listened to trisyllabic words in two stress-pattern conditions, with (+IA) or without (-IA) initial accenting, in an oddball paradigm. The EEG was recorded in both a passive and an active listening task, and in two different oddball versions: one where standard stimuli were +IA words and deviants -IA words, and the reverse for the other version (-IA standard, +IA deviant). Behavioral results show faster processing and less errors for +IA stimuli. ERP results show larger MisMatch Negativity component for -IA words, pointing out 1) that French listeners are sensitive to *f0* manipulation, and 2) that +IA is the preferred stress template in French. Altogether, our results indicate that French listeners not only discriminate stress patterns but that IA is encoded in long-term memory, hence phonologically relevant.

Index Terms: stress-pattern processing, initial accent, MisMatch Negativity, French

1. Introduction

French prosody presents some peculiarities, which makes its phonological description difficult to fully apprehend and which has consequences on effective propositions for speech processing models. Traditionally, descriptions of the French accentuation system account for a primary Final Accent (FA) co-occurring with intonational boundaries, and a secondary, optional Initial Accent (IA), essentially seen as an extra-metrical phenomenon [1; 2; 5]. Because accentuation (FA) in French is post-lexical and not lexically distinctive, it has been described as a 'language without accent' [1] or a 'boundary language' [2; 3]. In this view, it is also said that French listeners are 'deaf' to accentuation and have no long-term memory representation of stress patterns [4; *inter alia*]. Although FA is now undisputedly seen as a pitch accent [5], IA's phonological status is still unclear. While some models describe it as a pitch accent [6], most models in the frame of the AutoSegmental Metric approach consider IA as a 'loose boundary marker', which peak can be aligned to the first, second or even third syllable of the content word or accentual phrase in some cases [5; 7]. As such, IA is not clearly phonologically implemented in the French prosodic system: its role is that of a secondary, rhythmical balancing device, yielding to FA in case of tonal crowding [5; 8].

However, diverse studies point out systematic use of IA in prosodic structure marking and speech segmentation. Recently for example, IA has been shown to be a more reliable cue to prosodic structuring than FA in the marking of content words and accentual phrases [9]. Namely, IA very commonly marks lexical words inside larger prosodic units, more so than FA, and is a marker of left prosodic boundaries even when in close vicinity to FA and not rhythmically necessary. Other studies indicate that the cohesive prosodic units formed by IA and FA, also described as 'accentual arches' [10] strongly enhance the encoding and segmentation of linguistic units, over units marked by FA alone [11; 12]. Finally, IA ('early rise') is used to segment lexical units in the speech flow [13]. Despite potential 'looseness' of this 'early rise', naïve French listeners systematically perceive IA *on the first syllable* of content words and accentual phrases, even when its peak reaches its maximum further in the unit [14]. This may indicate that IA is strongly linked to the representation of lexical words. Altogether, these production and perception findings point out a more important role of IA than described in prosodic models. Namely, they are in keeping with a possible strong phonological role of IA as a left marker of small units, i.e. accentual phrases and possibly the level of the lexical word.

It is thus interesting to further investigate whether IA is *encoded* at some level of the linguistic representation. The Metrical Segmentation Strategy [15] states that the mental lexicon is accessed through pre-lexical, language-specific stress templates. Finding neural bases for IA processing in French would give further insights as to its phonological role in French. Very few studies have to date investigated the neural bases of accentuation, and mostly do so through the exploration of Event-Related Potential (ERP) components. Stress pattern violations in Dutch in an AX paradigm elicited a N325 component indicating pre-lexical processing of stress [16]. In French, violation of metrical patterns induced by artificial lengthening of the medial syllable in trisyllabic words induced delayed lexical decision. Indeed, lexically *congruent* words in the sentence elicited a larger N400 component when metrically incongruent, indicating that French listeners do have pre-lexical stress template representations [17]. Similar results were found for Chinese [18]. Other results using the oddball paradigm allow investigation of short-term and long-term memory processing of stress patterns, through the MisMatch Negativity ERP component [19; 20; 21].

The oddball paradigm allows presenting listeners with deviant stimuli in a background of standard stimuli, which processing elicits an early negativity in the context. The MisMatch Negativity (MMN) is thus a reconstructed ERP component obtained by subtracting deviants' elicited brainwaves from standards'. The oddball paradigm allows measuring the mnemonic trace that the repetition of standard stimuli leaves in short-term memory [22]. A MMN component emerges on the deviant stimuli when the standard stimuli leave a strong mnemonic trace in short-term memory. This effect is

emphasized by linguistic experience of phonological representations in the native language, i.e. by comparing the auditory stimuli with those representations in the long-term memory. Thus, while the oddball paradigm helps determining whether an auditory stimulus is perceptually discriminated, it can also help reveal whether this discrimination goes beyond lower processing levels [23]. In its pre-attentive version (*passive listening condition*), the MMN allows examining how the brain processes linguistic information online, without requiring listeners to make conscious decisions about the speech stimuli. Hence, results can be interpreted independently from other linguistic levels.

While most studies using MMN manipulate stress patterns' legality in words and pseudo-words [20; 21], the first step of our investigation concerns IA processing at the word level. By manipulating presence or absence of IA on words, our experiment is designed to test 1) whether French listeners *can* perceive and discriminate IA (acoustical low-level processing or salience discrimination), and 2) whether IA is encoded at the word-level (phonological, higher cognitive level or stress template processing). Indeed, although words can sometimes be pronounced without systematic IA in the speech flow, we hypothesize that IA stress patterns are the preferred, expected stress templates in French. Hence, we predict that ± IA stimuli are discriminated (short-term memory) and that a larger MMN will be elicited by -IA deviant patterns in a +IA standards background because they are less expected stress patterns (long-term memory stress pattern representation).

2. Method

2.1. Speech stimuli

The stimuli consisted of two words (*candidat* ('candidate') and *diffusion* ('broadcast')) that were naturally uttered with **IA (+IA)** in a sentence context by a naïve speaker. The words were placed in the sentence so as to appear at the beginning of a major intonation phrase (eg. *Le principe de cette nouvelle émission, dit-elle, et sa* ***diffusion****, pourraient être très mal accueillis par le public* ('the concept of this new program, she said, and its **broadcast,** may not be well accepted by the audience')), hence reinforcing the probability of clear marking of the target word by IA and FA [9].

The words were then extracted and resynthesized *without* **IA (-IA).** FA was kept constant and natural (see Figure 1).

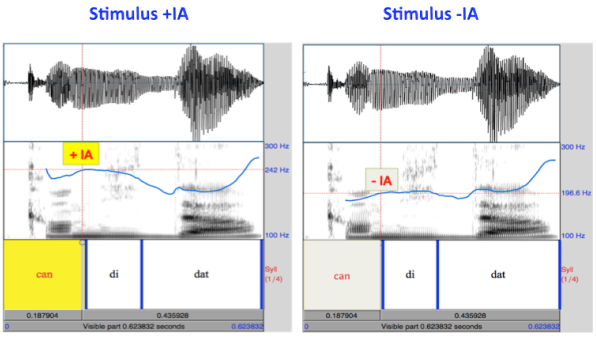


Figure 1: *Example of f0 resynthesis with (+IA) and without (-IA) on the word 'candidat', with quadratic interpolation from the f0 value of the preceding determinant to the f0 value at the beginning of the last stressed syllable for -IA words.*

The *f0* value of the first vowel was lowered near the *f0* value of the preceding (unaccented) determinant ('*le'* or *'sa')*. A quadratic transformation modified the *f0* values to reach progressively the *f0* value at the beginning of the last (accented) vowel, in order to maintain naturalness (micro-prosodic variations and some features of the original *f0* pattern). The +IA stimuli were forward and back transformed to equalize the speech quality between +IA and -IA stimuli.

The duration of the target words was held constant in both stress conditions (+IA; -IA), since only the *f0* parameter was manipulated (*candidat*: 624 ms; *diffusion*: 741 ms).

2.2. Participants and experimental tasks

30 healthy right-handed French native speakers, aged 18-35 (mean age 23,8; 27 females), participated in 2 versions of the Oddball paradigm in the same sequential order: a Passive *then* an Active listening task. Task order was not randomized throughout participants because doing the Active task before would have drawn attention to our stress manipulation. During the Passive task, subjects were watching a silent movie while listening to the repetition of the stimulus *'candidat*', in either of the two stress conditions (+IA; -IA). A total of 1092 words were presented with a pseudo-random combination of 106 deviant words and 986 standard words. The Inter Stimuli Interval (ISI) was 576 ms. The total experiment duration was 23 mns, in one single block. After a 10 mns' break, participants performed the Active task, where they were told to press a button as soon as they detected a deviant stimulus. In this task, the stimulus was '*diffusion*' in either of the two stress conditions (+IA; -IA), with the same total number of words and the same proportion of deviant and standard stimuli as in the previous task. The ISI was 459 ms and the experiment lasted 28 mns in total, throughout two blocks of 14 mns each with a 5 mns' break between the two blocks.

The participants were divided into two groups according to the Oddball version presented (see Figure 2): the first group (n= 14) performed the passive and the active task with -IA words as deviants and +IA words as standards (Version1); the reverse was true for the second group (n= 16) who performed both tasks with +IA as deviants and -IA as standards (Version2).

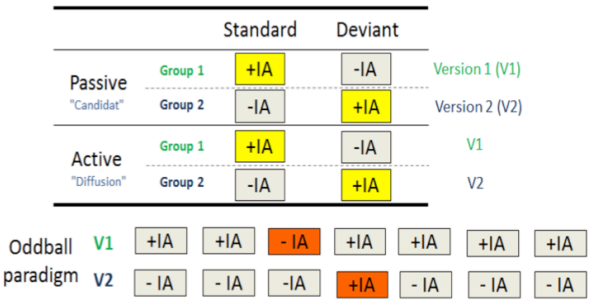


Figure 2: *Two versions of the Oddball Paradigm, where either -IA or +IA is the deviant. Participants are in 2 groups, each group having the same deviant condition in the Passive and the Active condition.*

2.3. EEG recordings

The EEG was recorded in both passive and active tasks, with 32 Ag/AgCl-sintered electrodes mounted on an elastic cap and located at standard left and right hemisphere positions over

frontal, central, parietal, occipital and temporal areas (International 10/20 System; Jasper, 1958) at : Fz, Cz, Pz, Oz, Fp1, Fp2, AF3, F3, AF4, F4, C3, C4, P3, P4, PO3, PO4, P5, P6, O1, O2, F7, F8, T3, T4, T5, T6, FC5, FC6, CP1, CP2, CP5 and CP6. The Horizontal ElectroOculoGram (HEOG) was recorded from a bipolar montage with electrodes placed 1 cm to the left and right of the external canthi; the Vertical ElectroOculoGram (VEOG) was recorded from a bipolar montage with electrodes placed beneath and above the left eye, to detect blinks and vertical eye movements. The EEG and EOG were amplified by BioSemi amplifiers (ActiveTwo System) with a band-pass filter of 0.01-30 Hz and was digitized at 512 Hz. Trials containing ocular artefacts, movement artefacts, or amplifier saturation were corrected from averaged ERP waveforms. The data were analysed using Brain Vision Analyser software version 2.0 (Brain Products, Munich, Germany). Each electrode was re-referenced off-line to the algebraic average of the left and right mastoids. Continuous recordings were segmented into 1100 ms duration starting 200 ms before stimulus onset (baseline). Trials were averaged within each of the two stress conditions. Finally, data were averaged across participants to obtain the grand-averages. Subject-averages were aligned to the 200 ms baseline preceding the auditory target. EEG results are presented in the Passive task, while only behavioural data (error rates and RTs) are presented for the Active task.

ERP data were statistically analysed by computing the mean ERP amplitude relative to a 80 ms window centred at the peak latency [100-250 ms] on 9 electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz and P4). For each comparison (within and between participants) ANOVAs were conducted with Stimulus Type (standard *vs* deviant) and Stress Condition (+IA *vs* -IA). All *p*-values were adjusted with the Greenhouse–Geisser epsilon correction for nonsphericity.

3. Results

Because of an abnormal error rates (>40%), data from five participants were excluded from the analyses. Moreover, one more participant was excluded due to a large number of artifacts on the ERP data. Thus, a total of 24 participants (version -IA: n=12; version +IA: n=12) were included on the behavioral data analyses and the ERP grand averages.

3.1. Behavioural results from the Active Task: Reaction Times and Errors Rates by deviant type

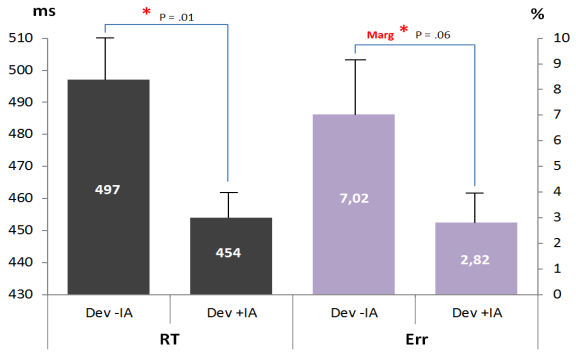


Figure 3: *Reaction Times (left; in ms) and Error Rates (right; in %) for -IA and +IA deviants. * indicates statistic significance.*

Paired two-tailed t tests revealed that RTs were significantly faster for detecting +IA deviants than for -IA deviants (454 ms vs. 497 ms; $t(23) = 3.27$, $p < .01$). Moreover, error rates differences between the two stress conditions were marginally significant ($t(23) = 1.1$, $p = .06$), -IA deviants being slightly more prone to errors than +IA deviants (7.02% *vs* 2.82%; see Figure 3).

3.2. Event-Related Potential results from the Passive Task

3.2.1. Within participants' MMN comparison

Results show a significantly larger MMN amplitude for -IA deviants than for +IA deviants ($F(1, 23) = 6.52$, $p < .01$), indicating that participants not only process the acoustic cues to stress, but that -IA deviants in a +IA standards environment are less expected than +IA deviants in a -IA standards environment.

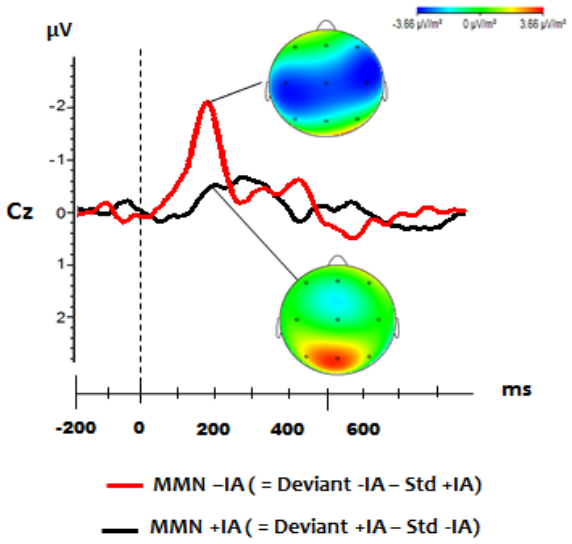


Figure 4: *MMN component for -IA and +IA deviants, for each participants' group. Grand average ERPs recorded at the Cz (central) electrode. Amplitude (µV) is represented on the ordinate, with negative voltage up, and time (msec) on the abscissa. Topographical maps are computed at the MMN peak latency.*

3.2.2. Between participants' MMN comparison

The previous within participants' results indicate that +IA deviants do not elicit an MMN. In order to unravel whether it is merely due to low-level, acoustic processing or whether it reflects higher-level processes (stress pattern rarity processing), we calculated MMN effects between identical deviant and standard stimuli (-IA deviants subtracted from -IA standards, and same procedure for +IA stimuli) across participant groups.

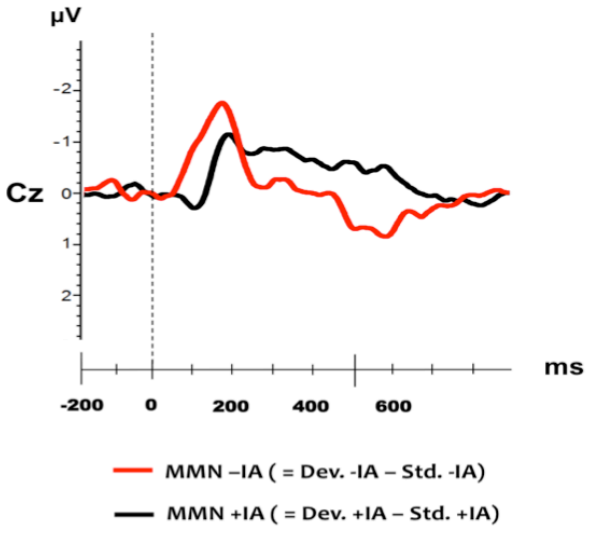


Figure 5: *Difference waves for (deviant – standard) -IA stimuli, and +IA stimuli respectively, across participants. Identical ERPs' graph and statistical settings than for Figure 4.*

Results for deviants and standards ± IA identical stimuli show a significantly larger MMN amplitude for -IA stimuli than for +IA stimuli ($F(1, 23) = 2.94$, $p < .05$).

4. Discussion

The aim of this study was to determine whether French listeners discriminate and process stress patterns in French. Since initial accents (IA) seem to play an important role in prosodic structuring and speech segmentation, the focus of this study was on IA processing at the word level. Despite still inconsistent accounts in prosodic models as to its phonological role, and its description as a 'loose boundary marker', our results indicate that French listeners readily perceive IA and that +IA stress patterns on lexical words are the expected, preferred patterns.

We chose to investigate IA processing through the MisMatch Negativity ERP component because it allows accounting for stress pattern processing at a pre-attentive level and independently from other levels of linguistic processing. In our study, participants were also submitted to an active version of the task in order to enrich the electrophysiological investigation with behavioral data.

Behavioral results show significantly faster Reaction Times to detect +IA deviants in a -IA standard background than to detect -IA deviants in a +IA standard background, which indicates facilitating processing for +IA stimuli. Error rates are also informative insofar as +IA deviants are significantly less numerous than -IA deviants. Altogether, behavioral results revealed that +IA words are discriminated and processed more easily and automatically than -IA words.

ERP data analysis revealed that French listeners process -IA and +IA word stimuli differently. If French listeners did not process stress and were 'deaf' to accentuation, both oddball versions would have yielded similar results, i.e. no MMN component would have emerged when -IA were deviants in +IA stimuli background and vice-versa. The presence or absence of f0 acoustical saliency would not have been discriminated. On the contrary, our results not only indicate that listeners process saliency differences but that they have a preferred stress pattern. Indeed, a large MMN emerges for -IA deviants, indicating that -IA deviants are surprising in a +IA context and showing that the absence of f0 variation on the word was processed (Figure 4). But if listeners were sensitive to low-level acoustic features only, a similar MMN should emerge when +IA deviants occur in a -IA context, indicating a mere processing of acoustic difference. However, Figure 4 shows no such MMN for +IA deviants. Because the MMN is a difference wave between deviant and standard stimuli, this latter result calls for further investigation. The MMN is a measure of the rarity of a stimulus in a context. This short-term memory rarity processing may be emphasized by pattern comparison in long-term memory. It may be that -IA patterns are not expected stress patterns and elicit a negativity wave even when presented as standards. Or it could be that -IA stimuli do not leave a strong enough trace in the short-term memory so as to elicit an ample MMN for +IA deviants, because they do not call upon long-term memory representations. Hence, the rarity processing of +IA deviant stimuli is reduced in the difference wave and no MMN emerges. To explore the rarity processing effect independently from acoustic and prosodic phonological effects, we calculated MMN for identical deviant and standard stimuli *across* participants. Figure 5 reveals a MMN for +IA rare stimuli but this MMN is significantly less ample than for -IA stimuli. Rare -IA stimuli thus remain more surprising than +IA deviants in matched contexts. Thus, a mere acoustic interpretation of ±IA processing has to be ruled out, because a MMN is still present between similar acoustical stimuli. On the contrary, our results show that ±IA stimuli both rely on short-term memory processing (the *absence* of saliency is surprising) and on long-term memory representation (-IA stress patterns are more surprising than +IA stress patterns and mismatch the long-term stress pattern representation).

Altogether, these results point out a preference for +IA stress patterns in French.

5. Conclusions and Perspectives

This study is, to our knowledge, the first investigation of stress pattern processing in French using the oddball paradigm. Results indicate that French listeners not only discriminate stress patterns but show preference for +IA stress patterns. This is a first step towards showing that IA is encoded at the lexical word level and it reinforces its role in speech encoding and decoding processes. These results echo similar findings in other languages [19; 20; 21]. Further MMN investigations will be conducted with more ±IA lexical items in order to generalize our interpretation of a phonological representation of +IA stress patterns. Following Honbolygo [21], we will also extend our investigation to pseudo-words, to emphasize the interpretation of +IA stress templates in long-term memory. Finally, this paradigm will also be extended to FA processing to address the potential similar role of IA and FA in French prosodic phonology.

6. Acknowledgments

This research is funded by a French National Research Agency (ANR) three-year award to the Principal Investigator Corine Astésano (n° ANR-12-BSH2-0001-01; '*PhonIACog*' project: 01/01/13 - 31/12/15).

7. References

[1] Rossi, M. (1980). Le français, langue sans accent ? In I. Fónagy & P. Léon (Eds.). *L'accent* en *français contemporain (Studia Phonetica), 15*. : 13-51.

[2] Vaissière, J. (1991). 'Perceiving rhythm in French?' *ICPhS'91*, 4, Aix-en-Provence: 258-261.

[3] Beckman, M.E. (1992). 'Evidence for Speech Rhythms across Languages'. In *Speech Perception, Production and Linguistic Structure*. Tohkura; Vatikiotis-Bateson; Sagisaka (eds.), Tokyo: 457-463.

[4] Dupoux, E.; Pallier, C.; Sebastian, N.; Mehler, J. (1997). A destressing "deafness" in French? *Journal of Memory & Language, 36,* 406-421.

[5] Jun, S.-A.; Fougeron, C. (2000). A phonological model of French intonation. In Botinis, A. (Ed.). *Intonation: Analysis, Modelling and Technology*. Kluwer, Boston: 209–242.

[6] Post, B. (2000). *Tonal and Phrasal Structures in French Intonation*. Thesus, The Hague.

[7] Welby, P. (2003). *The slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation*. PhD dissertation, Ohio State University.

[8] Di Cristo, A. (1999, 2000). Vers une modélisation de l'accentuation du français. *French Language Studies, 9*, 143-179 & *10*, 27-45.

[9] Astésano, C.; Bard, E.; Turk, A. (2007). Structural influences on Initial Accent placement in French. *Language and Speech, 50 (3)*: 423-446.

[10] Fonagy I. (1980). L'accent en français : accent probabilitaire. In Ivan Fonagy et Pierre Léon (eds.), *l'accent en français contemporain*. Studia Phonetica (pp. 123.233). Paris: Didier.

[11] Rolland, G.; Lœvenbruck, H. (2002). Characteristics of the Accentual Phrase in French: An acoustic, articulatory and perceptual study. In *Speech Prosody 2002, International Conference*.

[12] Bagou, O.; Frauenfelder, U. H. (2006). Stratégie de segmentation prosodique: rôle des proéminences initiales et finales dans l'acquisition d'une langue artificielle. *Proceedings of the XXVIèmes Journées d'Etude sur la Parole*, 571-574.

[13] Welby, P. (2007). The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication*, *49*(1), 28-48.

[14] Astésano, C.; Bertrand, R.; Espesser, R.; Nguyen, N. (2012). Perception des frontières et des proéminences en français. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP,* 353–360.

[15] Cutler, A.; Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113-121.

[16] Böcker, K. B.; Bastiaansen, M.; Vroomen, J.; Brunia, C. H.; Gelder, B. (1999). An ERP correlate of metrical stress in spoken word recognition.*Psychophysiology*, *36*(6), 706-720.

[17] Magne, C.; Astésano, C.; Aramaki, M.; Ystad, S.; Kronland-Martinet, R.; Besson, M. (2007). Influence of syllabic lengthening on semantic processing in spoken French: behavioral and electrophysiological evidence. *Cerebral cortex*,*17*(11), 2659-2668.

[18] Li, X.-q.; Ren, G.-q. (2012). How and when accentuation influences temporally selective attention and subsequent semantic processing during on-line spoken language comprehension: An ERP study. *Neuropsychologia, 50*, 1882–1894.

[19] Honbolygó, F.; Csépe, V.; Ragó, A. (2004). Suprasegmental speech cues are automatically processed by the human brain: a mismatch negativity study. *Neurosci. Lett. 363*, 84–88.

[20] Ylinen, S.; Strelnikov, K.; Huotilainen, M.; Näätänen, R. (2009). Effects of prosodic familiarity on the automatic processing of words in the human brain. *International Journal of Psychophysiology 733*, 362–368.

[21] Honbolygó, F.; Csépe, V. (2012). Saliency or template? ERP evidence for long-term representation of word stress. *International Journal of Psychophysiology 87,* 165-172.

[22] Winkler, I. (2007). Interpreting the mismatch negativity. *Journal of Psychophysiology 21 (3/4)*, 147–163.

[23] Näätänen R.; Paavilainen P.; Rinne T.; Alho K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology, 118:* 2544–2590.

Effects of dynamic pitch and relative scaling on the perception of duration and prosodic grouping in American English

Alejna Brugos and Jonathan Barnes

Boston University, Boston, Massachusetts, USA

abrugos@bu.edu

Abstract

Results of two perception experiments suggest that using timing measures alone to compute prosodic structure misses valuable information from pitch. Previous research showed that pitch can distort perceived duration: tokens with dynamic or higher f0 are perceived as longer than comparable level-f0 or lower-f0 tokens, and silent intervals bounded by tokens of widely differing pitch are heard as longer than those bounded by tokens closer in pitch (the kappa effect). Phrase edges (signalled by increased duration, pause, phrase tones, and f0 reset) set the scene for pitch to modulate perceived duration. Two new experiments used the same duration and f0 manipulations (level vs. varying-slope rises, at varying pitch ranges) of segmentally-identical base files, in two separate tasks: 1) a linguistic grouping task using an ambiguously-structured phrase and 2) a psychoacoustic study on perceived duration. Results show that effects on perceived duration due to dynamic pitch can be either strengthened or nullified depending on relative scaling of compared tokens. These same manipulations push grouping judgments beyond what would be expected from distortions of perceived duration. This suggests that listeners integrate pitch and timing cues when judging linguistic structure, supporting measures of relative boundary size that combine duration and pitch measures.

Index Terms: duration perception, auditory illusions, dynamic pitch, timing, prosodic grouping, boundary tones

1. Introduction

Edges of prosodic groups are known to be marked (at least in many languages, Fon (2012) [1]) by pre-boundary lengthening, silent pauses, phrase tones and reset. Phonetic measures of these features are typically taken independently, without consideration of how they may interact in perception. However, perception of time can be systematically affected by a range of contextual factors (Brown, 2008) [2], including pitch (Hoopen, 2008) [3]. A growing body of work from a diverse range of fields shows that pitch and timing are not entirely perceptually independent.

1.1. Pitch-time interaction

Since Lehiste (1976) [4] showed that subjects perceived vowels with dynamic f0 as longer in duration than static-f0 vowels of the same objective duration, many studies have tried to replicate this finding, with varying results (see Cumming, 2011 [5] for an overview). Cumming (2011) [5] and Yu (2010) [6] both reproduced this effect, with differing methodologies and languages; Yu also found that vowels with higher f0 were perceived as longer than lower-f0 vowels.

Outside of speech research, there is a substantial body of experimental work on pitch-timing interaction in perception. Henry (2011) [7] showed that perception of duration of non-speech tone glides can be modulated by the pitch change

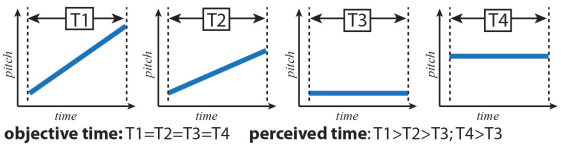


Figure 1: *Schematic of dynamic pitch & scaling effects on perceived duration of filled intervals. The intervals (T1, T2, T3 & T4) are objectively equal in duration; those with dynamic pitch sound longer than level, and more so with steeper pitch. Higher pitch intervals are also perceived as longer than those with lower pitch.*

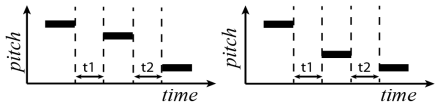


Figure 2*: Schematic example of the auditory kappa effect, whereby relative tone height affects timing perception. Silent intervals (t1, t2) are equal, but t1 sounds shorter at left, longer at right.*

velocity of the target glide and of the standards: glides with greater pitch change velocity are perceived as longer than those with lesser pitch change velocity, or level-pitch tones.

Other work has focused on the effects on perceived duration of silent intervals bounded by filled intervals of varying pitch distance, a phenomenon known as the auditory kappa effect: Silent intervals bounded by tones of closer pitch proximity are perceived as shorter in time than those of equal objective duration bounded by tones of a greater pitch distance (Cohen et al., 1953[8], 1954[9]). While typically demonstrated using non-speech tones (Shigeno, 1993 [10], Crowder & Neath, 1995 [11]; MacKenzie, 2007 [12]; *inter alia*), Brugos & Barnes (2012b) [13] showed that the auditory kappa effect also obtains for spoken language, such that the perceived duration of silent pauses in speech was modulated by the pitch distance across those pauses. In a second study using identical materials, Brugos & Barnes (2012a) [14] found that the effect of these pitch manipulations was even greater on perceived prosodic grouping of these phrases: even effects of objective duration differences on grouping perception were in some cases overridden. These results suggest that relative pitch proximity of neighboring prosodic phrases, described in the literature as phrase-initial reset (Jun, 2003) [15], should be taken into account for estimating boundary size, and support a trading relationship between pitch and timing cues in prosodic grouping (Beach, 1991 [16]; Cumming, 2011b [17]; Jeon & Nolan, 2013 [18]).

Of course, since phrase boundaries in natural speech commonly play host to dynamic pitch (e.g., boundary tones), pitch jumps (e.g., reset), and durational variation (phrase-final lengthening/pauses), we might expect all these pitch and timing cues to enter into cue-trading relationships for the signalling of prosodic grouping. In fact, something of this sort has been shown in a variety of studies. When f0 cues are neutral, grouping can be cued by duration cues alone (Scott,

1982 [19]; Wagner & Crivellaro, 2010 [20]); likewise f0 cues alone can cue grouping when pitch cues are held constant (House, 1990 [21]). When pitch and timing are manipulated together, the picture of cue interactions for prosodic grouping becomes more complex (Beach, 1991 [16]; Jeon & Nolan, 2013 [18], 2010 [22]; Cumming, 2011b [17]).

2. Two experiments

To investigate how dynamic pitch and duration might interact in both duration and grouping perception in American English, a pair of experiments was designed using the same stimuli in two different tasks: 1) a linguistic judgment of grouping in an ambiguous phrase based on cue interpretation and 2) a psychoacoustic judgment of perceived duration.

2.1. Methods

The ambiguous phrase chosen as context for the linguistic grouping task was a string of color terms *blue and green and purple* (following methodology of Beach et al., 1996 [23]). This phrase can be parsed variously: 1) ungrouped (a simple list of 3 colors) or 2) two groups, one pair of colors and a third color on its own, i.e.: *blue and (green and purple)* (B-GP) or *(blue and green) and purple* (BG-P).[1]

The f0 pattern of base recordings of *blue* was manipulated to include both dynamic f0 and plateau contours, and crossed with a continuum of duration manipulations leading to changes in relative duration of the words *blue* and *green*. Assuming that *blue* being longer than *green* cues more B-GP responses, and that dynamic f0 cues longer perceived duration, then we might predict dynamic F0 in *blue* likewise to cue more B-GP responses: results of duration and grouping perception tasks, in other words, should be largely overlapping. However, if, like in Brugos & Barnes 2012a [14], the effects of pitch go beyond their modulation of perceived duration, results from the two tasks are expected to diverge.

2.1.1. Stimuli

Manipulations of pitch and duration were performed to an identical base recording of *blue*, and these same resultant resyntheses were used in both the grouping perception and duration perception tasks as described below. A single version of the complete phrase was used as a base file for additional manipulations of f0 and duration to produce stimuli in the experiments. In order to see whether pitch manipulations modulate grouping perception by way of timing perception, it was necessary first to create a neutral condition in which timing manipulations alone might shift perceived grouping. A female native speaker of American English (the first author) produced multiple versions of the phrase, and the eventual base file token was selected through an extensive process of evaluation, resynthesis, and concatenation of naturally spoken words. Durations and intensity of each of the words were adjusted to produce a natural-sounding concatenation that did not strongly cue either B-GP or BG-P grouping. Base durations and duration continuum points were chosen based on a pre-experiment screening with 12 subjects via web form.

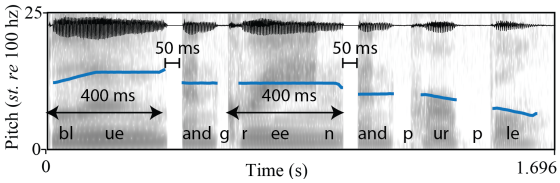


Figure 3: *The base/neutral file. The words* blue *and* green *are ~ 400* ms *long. The f0 contour of* blue *is a plateau with a 2* st *rise, and then level f0,* and green *is level f0, 2* st *below the max f0 of blue,* and purple *starts another 2* st *lower, and ends in a 4* st *fall.*

Pitch and timing manipulations: The base recording of *blue* was resynthesized to create 3 f0 contour shapes at 5 durations, and at 3 steps affecting f0 range. All tokens of blue began with a rise similar to what was seen in natural productions: the 2 *st* rise began at the onset of voicing, through the [l] and into the beginning of the vowel [u]. In order to reduce segmental variation in the onset that might cue differences in perceived prominence and grouping, all manipulations for duration and contour were done only to the /u/ portion of the word following this pivot point (at 158 milliseconds into the word). From this point, the f0 did one of 3 things: 1) stayed level to the end of the word ("plateau"), 2) rose an additional 2 *st* to the end of the word ("2-st-rise") or 3) rose 4 *st* from the pivot ("4-st-rise"). 5 duration manipulations were performed on this same post-pivot interval such that the total duration of the word equalled 300 *ms*, 350 *ms*, 400 *ms*, 450 *ms*, and 500 *ms*, creating 15 time-by-contour manipulations (Figure 4, left).

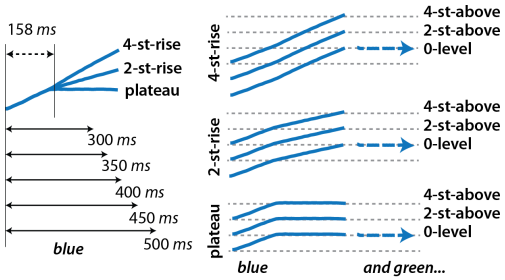


Figure 4*: The base file* blue *was resynthesized to create 3 contour shapes at 5 durations, and duration of the initial 2* st *rise was held constant (left). F0 contours for* blue *were each shifted to 3 f0 steps.*

Unfortunately, introducing a comparison of dynamic vs. static pitch into an experiment such as ours turns out to be far from simple. As Figure 5 shows, any attempt to alter f0 dynamicity during the word *blue* in our sequence necessarily alters the pitch gap across the following boundary also. Given the results of Brugos & Barnes (2012b)[14], this turns out to be a serious potential confound for any investigation of the effects of dynamic pitch on perceived duration.

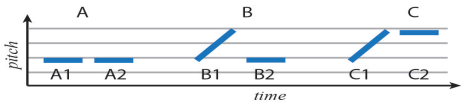


Figure 5: *Schematic showing how dynamic f0 can introduce pitch differences across phrases.*

[1] It should be noted that there is controversy in the literature as to whether it is primarily just relative boundary size, rather than categorical identity of the boundaries involved, that matters most for interpretation (Price et al. 1991 [24]; Clifton et al., 2002 [25]; Jun, 2003 [15]; inter alia). We remain agnostic here as to whether the prosodic phrases used in this study are instantiations of specific levels of the Prosodic Hierarchy (Selkirk, 1986) [26], or instead recursive or gradiently-sized groups (Ladd, 1986 [27]; Wagner, 2005 [28]; Shreuder, 2006 [29]; Kentner & Féry, 2013 [30]). The points we wish to make regarding the implementation of phrasing hold equally well under both scenarios.

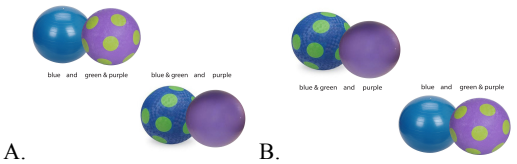
In order to control for this confound, we chose to actively manipulate pitch step orthogonally to the manipulations of dynamic pitch so that we could better separate the effects. Each resulting contour/time combination was resynthesized at 3 different f0 ranges ("pitch steps") based on the pitch relationship between the end point of *blue* and the f0 of the immediately following words in the grouping experiment (*and green),* which were always level at 202 Hz. 3 pitch steps were chosen: 1) 2 *st* above *green* (the level of the neutral condition for the plateau contour shown in Figure 3, above), 2) ending 4 *st* above *green* and 3) ending level to *green*. In each case, the entire contour was shifted up or down in pitch space.

Because manipulations to *blue* alone were not sufficient to cue grouping differences for all listeners in a pre-experiment screening, and as it was obvious to some listeners when the post-*blue* phrase was unchanging, manipulations to the duration of *green* were also added. Base *green* was approximately 400 *ms* in duration, and additional manipulations yielded *green* at 350 and 450 *ms* as well.

2.1.2. Subjects, presentation and task

16 native speakers of American English (age 18 to 22 years) participated in Experiment 1 for a payment of $10. 9 of these subjects additionally participated in Experiment 2 for an additional $10. (Experiment 2, the duration task, was always presented to subjects after completion of Experiment 1.) Subjects faced a laptop and listened to stimuli over headphones. Both experiments were forced-choice tasks, with responses indicated via a button box or designated keys on the laptop. Each experiment took about 25 minutes, including breaks and training. Subjects read a brief introduction to the study, then proceeded to a training section, to ensure that they understood the task. For Experiment 1, subjects were presented with natural examples of prosodic grouping (of repeated digits) produced by a naïve speaker. Subjects proceeded to the experimental phase after answering at least 75% correct of at least 10 training trials. For Experiment 2, training consisted of presentations of plateau tokens of differing durations, using only duration differences of 100 *ms* or greater. Subjects proceeded to the experimental phase upon correctly answering 5 training trials in a row.

Experiment 1: Grouping perception. The screen presented 2 images representing two grouping choices: 1) One solid blue ball, and another purple with green spots (B-GP) and 2) one blue ball with green spots, and another solid purple (BG-P). The text "blue and green & purple" and "blue & green and purple" accompanied each image. For each trial, subjects were played a recording of the complete phrase *blue and green and purple*, and asked to indicate whether they heard the phrase as corresponding to B-GP or BG-P. They were not instructed to attend to any specific aspect of the signal. Trials included phrases with the 45 blue manipulations (3 contours x 5 durations x 3 time steps), paired with the phrase completion,



A. B.

Figure 6: *Images presented to subjects for Experiment 1, with text and images suggesting the 2 groupings. Subjects either saw only A, or only B.*

including 3 durations of green (350, 400, and 450 *ms*): 4 repetitions with 400-*ms green*, and 2 repetitions each for the other 2. These 360 trials were randomized for each subject.

Experiment 2: **Duration perception**. Target versions of the word *blue* identical to those in Experiment 1 were compared to level-f0 standards of the same base file *blue*. Standards were completely level-f0 versions of *blue* at 202 Hz (the level of *and green* in the grouping experiment), and presented in the same 5 durations of the targets (300, 350, 400, 450 and 500 *ms*). Trials consisted of a target followed by a standard, with 200 *ms* of silence interceding. Each pairing was also presented in the opposite order, that is, standard followed by target (45 targets x 5 standards x 2 orders = 450 trials). An additional repetition of each of the 2 rise contours in the target-standard order was included to maximize repetitions of the comparisons of greatest interest, for a total of 600 trials, randomized for each subject. Subjects were asked to indicate which of the two repetitions of the word *blue* sounded longer.

3. Results and analysis

Experiment 1: Results presented are from 5598 experimental trials for 16 subjects. Figure 7 displays proportion of B-GP responses as a function of duration difference between the words *blue* and *green* (durations of *and* & *purple* remain constant for all conditions). Positive time values indicate that *blue* is longer than *green*, and negative that *blue* is shorter. At left, individual lines represent the 3 contours (plateau, 2-st-rise, 4-st-rise), and at right the 3 pitch steps (0-level, 2-st-above, 4-st-above). The upward diagonal trend of the lines in both graphs shows that responses are strongly correlated with duration difference between *blue* and *green*. Bigger time values show more responses that *blue* is grouped separately (B-GP), and smaller time values more responses that *blue* is grouped with *green* (BG-P). The three lines by contour (left) show virtually no separation, but the 3 lines by pitch step (right) show clear separation such that trials where the pitch of *blue* ends level with the following words show proportionately greater responses that *blue* grouped with *green* (BG-P), and higher pitch steps show increasingly more responses that *blue* is grouped separately (B-GP).

Results were analyzed using mixed-effects logistic regression, implemented through the lme4 package (Bates & Maechler, 2009 [31]) in R with response ("B-GP" or "BG-P") as dependent variable, and time step, pitch step and f0 contour as fixed factors. Subject was included as a random effect (Baayen et al., 2008 [32]). The result was a model (N = 5598, log-likelihood = -3269) showing an expected significant main effect of time step (Wald Z = 29.57, $p < .001$), as well as main effects for pitch step, with 2-st and 4-st differing from 0-level (2-st: Wald Z = 5.371, $p < .001$; 4-st, Wald Z = 7.54, $p < .001$). There was a weak effect of contour (between plateau and 2-st-rise (Wald Z= 1.98, $p < .05$), also pushing responses toward B-GP. There was also a somewhat complicated series of interactions between contour and pitch step such that dynamic pitch contours showed a slight tendency at the 2-*st* and 4-*st* steps to produce fewer B-GP responses, which will be addressed in the discussion section.

Experiment 2: Results are presented for 9 subjects for 3360 trials (only target-standard order trials are included, being closest to conditions of Experiment 1.) Figure 8 displays the proportion of "target longer" responses as a function of duration difference between target and standard.

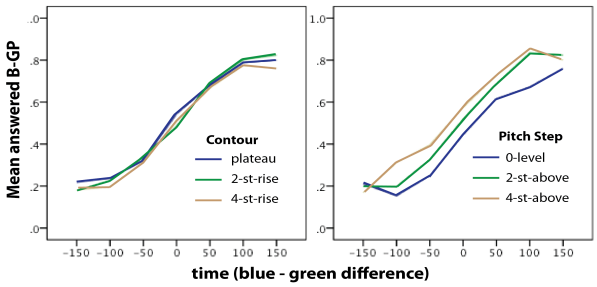


Figure 7: *Experiment 1 results: Mean responses B-GP grouping by time, with lines by contour (left) and pitch step (right). Lines by contour overlap, but lines by pitch step show separation.*

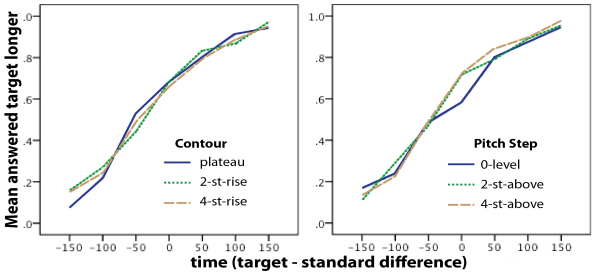


Figure 8: *Experiment 2 results: Mean responses "target longer" by time. Lines by contour (left) overlap. At right, lines by pitch step show some separation at time=0, but otherwise overlap.*

Positive time values indicate a target longer than a standard, and negative that the target is shorter. (For time = 0, target and standard were of identical duration). At left, lines again represent the 3 contours, and at right the 3 pitch steps. The upward diagonal trend of the lines in both graphs shows that subject responses are strongly correlated with the duration, with more responses that the target was longer when the target was indeed objectively longer. The three lines by contour in (figure 8, left) again show virtually no separation, and the 3 lines by pitch step (figure 8, right) only show some suggestion of separation where time = 0.

A mixed model analysis was performed much as in Experiment 1, but with response ("target longer" or "standard longer") as the dependent variable. The result was a model (N = 3360, log-likelihood = -1601) showing significant main effect of time step (target-standard time difference) (Wald Z = 27.47, $p < .0001$). There was no main effect of contour, but also no significant main effect of pitch step. There was, however, a slightly significant interaction between contour and pitch step (Wald Z = 2.43, $p < .05$) only for the 2-st-rise contour with the 4-st-above pitch step.

4. Discussion

The effects of f0 manipulations on grouping perception and duration perception are not identical. The effect of contour (rising pitch vs. plateau) did not strongly influence results in either the grouping perception or the duration perception experiment. Pitch step appears to have influenced responses in both experiments, but potentially in complicated ways.

The complicated interactions between pitch step and contour seen in the results of Experiment 1 are likely a reflection of perceived scaling differences. While pitch step was defined here in terms of pitch distance between the end of *blue* and the following phrase, this likely reflects perceived scaling only for plateau tokens. It is known that for tone glides, perceived pitch is roughly equivalent to a point roughly 2/3 of the way through the glide (Rossi's "2/3 rule," Rossi, 1971 [33]). Estimating perceived pitch thus, tokens where *blue* is either dramatically higher or lower than green will tend to be perceived as B-GP, and where they are similar in pitch there are likely to be more BG-P judgments. Closer pitch across the boundary may influence perceived tonal continuity, and such pitch proximity or similarity of pitch may suggest a weaker boundary. Large pitch changes across a boundary, conversely, create a greater discontinuity, and cue a stronger boundary. Such an observation is in keeping with proposals of prosodic grouping that make reference to gestalt-like principles, such as proposed by Kentner & Féry (2013) [30], and similar to principles proposed for music grouping by Lerdahl & Jackendoff (1983) [34]. It is also worth noting that pitch is recognized as playing a role in connecting phrases into coherent segments in discourses (Wichmann, 2000 [35]; Hansson, 2003 [36]; Hirst, 1993 [37]), but such effects have not been as thoroughly explored in intonational phonology traditions such as AM/ToBI (Beckman & Ayers, 1997 [38]).

In Experiment 2, the lack of effect of pitch contour, seemingly suggests that dynamic pitch does not modulate perceived duration. This may well be due to the orthogonal manipulation of pitch range. When examining the subset of cases where the target and standard durations were equal, more responses that the target was longer can be seen when there is a big pitch jump between target and standard. It is possible that, as shown with the kappa effect, such a pitch jump affects perceived duration of the interceding silent interval. Indeed, it may be the case that subjects are not clearly distinguishing between the durations of the filled intervals vs. the silent intervals. (Note that there is a larger silent interval, 200 *ms* vs. 50 *ms*, in Experiment 2, and this may be reflected in the overall tendency for subjects to hear the first item as longer: 60%, in spite of balanced presentations.) It is possible that at least some of the previously reported effects of dynamic pitch on perceived duration are due to a confound of dynamic pitch introducing pitch step differences across compared tokens. Under certain circumstances dynamic pitch may have an effect, but circumstances may need to be just right, and may be overridden by other effects, such as relative scaling. It is also conceivable that while the psychoacoustic effect of dynamic pitch on perceived duration is real, it may not transfer straightforwardly to speech.

5. Conclusions

These results of these experiments add to our understanding of the effects of pitch on perceived duration, and suggest that pitch factors affect grouping judgments beyond what would be expected from distortions of perceived duration. While pitch factors have been shown to modulate perceived duration, such effects do not account for the degree to which pitch changes affect perceived grouping. Pitch relations across boundaries influence perceived juncture across those boundaries, at times overriding the effects of durational cues. Results support the idea that listeners integrate pitch and timing cues when judging linguistic structure, supporting measures of relative boundary size that combine duration and pitch measures.

6. References

[1] Fon, Y.-J. J., *A Cross-linguistic Study on Syntactic and Discourse Boundary Cues in Spontaneous Speech.* Dissertation. Ohio State University, 2002.

[2] Brown, S. W. "Time and attention: Review of the literature." *Psychology of Time*, 111–138, 2008.

[3] Hoopen, G. T. "Classic Illusions of Auditory Time Perception." *Journal of the Human-Environmental System*, *11*(1), 27–35, 2008.

[4] Lehiste, I.. "Influence of fundamental frequency pattern on the perception of duration." *Journal of Phonetics*, 4, 113–117, 1976.

[5] Cumming, R. "The effect of dynamic fundamental frequency on the perception of duration." *Journal of Phonetics*, 39(3): 375–387, 2011a.

[6] Yu, A., "Tonal effects on perceived vowel duration." In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (Eds.), *Papers in Laboratory Phonology* (Vol. 10). Berlin: Mouton de Gruyter, 2010.

[7] Henry, M. *A Test of an Auditory Motion Hypothesis for Continuous and Discrete Sounds Moving in Pitch Space*. Dissertation. Bowling Green State University, 2011.

[8] Cohen, J., Hansel, C., & Sylvester, J. "A new phenomenon in time judgment", *Nature*, *172*: p. 901, 1953.

[9] Cohen, J., Hansel, C. & Sylvester, J. "Interdependence of temporal and auditory judgments." *Nature*, *174*: 642–644, 1954.

[10] Shigeno, S. "The interdependence of pitch and temporal judgments by absolute pitch possessors." *Perception & Psychophysics*, 54(5): 682–692, 1993.

[11] Crowder, R. & Neath, I. "The influence of pitch on time perception in short melodies", *Music Perception*, 12(4): 379–386, 1995.

[12] MacKenzie, N. *The kappa effect in pitch/time context*. Dissertation. Ohio State University, 2007.

[13] Brugos, A. & Barnes, J. "The auditory kappa effect in a speech context." Speech Prosody, Shanghai, China. 2012b.

[14] Brugos, A. & Barnes, J. "Pitch trumps duration in a grouping perception task," 25th Annual CUNY Conference on Human Sentence Processing, New York, NY. 2012a.

[15] Jun, S. "Prosodic phrasing and attachment preferences." *Journal of Psycholinguistic Research*, *32*(2), 219–249, 2003.

[16] Beach, C. M. "The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations." *Journal of Memory and Language*, *30*(6), 644–663, 1991.

[17] Cumming, R. E. "The interdependence of tonal and durational cues in the perception of rhythmic groups," *Phonetica* 67, 219–242, 2011b.

[18] Jeon, H.-S., & Nolan, F.. "The role of pitch and timing cues in the perception of phrasal grouping in Seoul Korean." *The Journal of the Acoustical Society of America*, *133*(5), 3039-3049, 2013.

[19] Scott, D. R. "Duration as a cue to the perception of a phrase boundary." *Journal of the Acoustical Society of America*, *71*(4), 996–1007, 1982.

[20] Wagner, M., & Crivellaro, S. "Relative Prosodic Boundary Strength and Prior Bias in Disambiguation." in *Proceedings of Speech Prosody 2010*. Chicago, 2010.

[21] House, D., *Tonal Perception in Speech.* Lund, Sweden: Lund University Press, 1990.

[22] Jeon, H.-S., and Nolan, F. "Segmentation of the Accentual Phrase in Seoul Korean," in *Proceedings of Speech Prosody 2010*, Vol. 100023, pp. 1–4, 2010.

[23] Beach, C. M., Katz, W. F., & Skowronski, A. "Children's processing of prosodic cues for phrasal interpretation." *Journal of the Acoustical Society of America*, *99*(2), 1148–1160, 1996.

[24] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. "The use of prosody in syntactic disambiguation." *Journal of the Acoustical Society of America*, 90, 2956-2970, 1991.

[25] Clifton, C. J., Carlson, K. and Frazier, L., "Informative prosodic boundaries," *Language and Speech*, vol. 45, pp. 87–114, 2002.

[26] Selkirk, E. "On derived domains in sentence phonology." *Phonology Yearbook*, 3: 371-405, 2986

[27] Ladd, D. R. "Intonational phrasing: The case for recursive prosodic structure." *Phonology* 3. 311–340, 1986.

[28] Wagner, M. *Prosody and recursion*. Dissertation, MIT, 2005.

[29] Schreuder, M. J. *Prosodic processes in language and music*, Dissertation, Groningen University, 2006.

[30] Kentner, G., & Féry, C. "A new approach to prosodic grouping." *The Linguistic Review*, *30* (2), 2013.

[31] Bates, D. & Maechler, M. "lme4: Linear mixed-effects models using S4 classes." R package version 0.999375-32, 2009.

[32] Baayen, R., Davidson, D. & Bates, D. "Mixed-effects modeling with crossed random effects for subjects and items", *Journal of Memory and Language*, 59: 390-412, 2008.

[33] Rossi, M. "Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole," *Phonetica*, **23**, 1-33, 1971.

[34] Lerdahl, F., & Jackendoff, R. *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press, 1983.

[35] Wichmann, A. *Intonation in Text and Discourse*: *Beginnings, Middles and Ends*. Harlow, UK: Longman, 2000.

[36] Hansson, P., *Prosodic Phrasing in Spontaneous Swedish*. Dissertation, Lund University, 2003.

[37] Hirst, D. "Peak, boundary and cohesion characteristics of prosodic grouping." *ESCA Workshop on Prosody*, 1993.

[38] Beckman, M., & Ayers Elam, G. *Guidelines for ToBI Labelling* (version 3, March 1997).

Distinguishing Phrase-Final and Phrase-Medial High Tone on Finally Stressed Words in Turkish

Canan Ipek[1], Sun-Ah Jun[2]

[1]Department of Linguistics, University of Southern California, Los Angeles
[2]Department of Linguistics, University of California Los Angeles, Los Angeles
ipek@usc.edu, jun@humnet.ucla.edu

Abstract

The goal of this paper is to investigate the nature of the high tones realized on finally stressed words in Turkish. Following Ipek & Jun's [1] AM model of intonational phonology of Turkish, it was hypothesized that the high tone realized on the last syllable of a phrase (i.e., Intermediate Phrase (ip)) is realized differently from that of a phrase-medial prosodic word (PW), reflecting the prosodic hierarchy. Acoustic data show that an ip-final High tone shows larger f0 rise than a PW-final High tone, and the ip-final syllable is longer than the PW-final syllable. Furthermore, the degree of coarticulation is weaker across an ip boundary than a PW boundary. These findings support the prosodic structure and tonal categories proposed in Ipek & Jun's [1] model of Turkish intonation.

Index Terms: Turkish intonation, pitch accent, Intermediate Phrase, boundary tone

1. Introduction

The intonational contour of a Turkish declarative sentence produced in a neutral context typically consists of a sequence of rising tones [L H], followed by a nuclear accented word in a narrower pitch range, and a low boundary tone at the end of the sentence. The domain of a rising tone is a Prosodic Word (PW) with the L tone on the first syllable of a word and the H tone on its stressed syllable, often the last syllable of the word. The primary goal of this paper is to investigate the nature of high tones on finally-stressed words in Turkish that are either (a) phrase medial, or (b) phrase final. Based on the findings, we aim to support Ipek & Jun's [1] model of the intonational phonology of Turkish being developed in the Autosegmental-metrical (AM) framework [2, 3, 4, 5].

Word-final high tones have been analyzed differently in two previous phonological models of Turkish intonation within the AM framework.. In [6], a high tone on a word-final stressed syllable was analyzed as a pitch accent regardless of whether the word is in the middle or at the end of a Phonological Phrase. On the other hand, [7], following [8],assumes that words with final stress under the traditional analysis [9, 10, 11] are lexically unstressed, and instead the High tone is analyzed as a boundary tone (H-) of a Major Phrase (or Phonological Phrase or Accentual Phrase). Since the Major Phrase included only one prosodic word in [7], it is not clear whether H- is a boundary tone of a prosodic word or of a Major Phrase. It is also not clear how a phrase-medial, word-final High tone would be analyzed in this model.

In Ipek and Jun's [3] intonation model, a High tone on a stressed syllable is analyzed as a H* pitch accent regardless of the location of stress, but when the finally-stressed word is the last word in an Intermdiate Phrase (ip), the prosodic unit above the Prosodic Word, the High tone on the ip-final stressed syllable is assumed to have a dual function. First, it marks prominence on the word (the function of a pitch accent); second, it marks the right edge of an ip (the function of a boundary tone). This tone is therefore given the label H*-.

In the present study, we will examine acoustic data from word-final syllables in the two prosodic conditions, PW-final vs. ip-final. The aim is to examine (a) whether the word-final High tones are realized differently depending on their prosodic condition and (b) whether the degrees of final lengthening and of coarticulation with a following word reflect the word-final syllable's location in the prosodic hierarchy proposed in [1]. Previous studies have shown that, compared with syllables at the edge of lower constituents, syllables at the edge of higher constituents will show higher f0 peaks [2, 12, 13], greater degrees of final lengthening [14, 15, 16, 17], and lesser degrees of trans-boundary segmental coarticulation [18]. We will therefore test the following two hypotheses: (i) There will be a higher f0 peak and more phrase-final lengthening for a word bearing a final H*- than one bearing the H*; (ii) Vowels will undergo less coarticulation across a boundary following the H*- than across a boundary following H*.

2. Experiment

2.1. Method

2.1.1. Stimuli

Five pairs (i.e., ten total) of five-word sentences were designed to examine the acoustic properties of word-final H* and H*-, and the prosodic juncture following them (see Table 1). Every sentence started with a subject noun phrase (NP), and the number of words within the subject NP differed between the two sentences in each pair. One sentence of each pair contained a three-word subject NP (Group 1), and the other contained a two-word subject NP (Group 2), but the target word was always the second word (i.e., word2).

According to [6], word2 in each Group receives a pitch accent (L+H*), but according to [7], word2 in Group 2 receives a H- boundary tone marking a Major Phrase. (The H tone of word2 in Group 1 is not analyzed in [7]). In Ipek and Jun's [1] model, the high tone realized on the final syllable of word2 in Gorup1 is in the middle of an ip (therefore, it would be marked with H*), but the final syllable of the same word in Group2 is the last syllable of an ip (corresponding to a right edge of NP), and so would be marked with H*-.

In addition, in order to measure the degree of vowel coarticulation across the boundary between the second and the

third word in each sentence, the last vowel of the second word was fixed as [a] and the first vowel of the third word was fixed as [i], maximizing the difference in the first (F1) and second formant (F2) values between the vowels.

Table 1. *A list of ten target sentences in the two groups. The second word (in bold) in Group1 is ip-medial, marked with H*, but the same second word in Gorup2 is ip-final, marked with H*-. Below each sentence, an English gloss is given word-by-word, followed by the meaning of the sentence.*

Group 1: Intermediate Phrase-medial High tone (H*) {word1 **word2**^{H*} word3}ip word4 word5
1-1. Ülkedeki **yasak** içkiler dışardan geliyor. in the country banned drinks from outside coming "Drinks banned in the country are coming from outside the country."
1-2. Yoldaki **tuzak** işaretler önümüzü kapatıyor. on the road trap signs our front blocking "Trap signs on the road are blocking our sight."
1-3. İnşaattaki **tutsak** işçiler çevreyi temizliyor. in the construction captive workers the neighborhood cleaning "Captive workers in the construction are cleaning the neighborhood."
1-4. Dergideki **korkak** içerikler tamamen kaldırılmış. in the magazine coward contents completely removed "Coward contents in the magazine were completely removed."
1-5. Camdaki **çatlak** izler önümüzü kapatıyor. on the mirror cracked traces our front blocking "Cracked traces in the mirror are blocking our sight."
Group 2: Intermediate Phrase-final High tone (H*-) {word1 **word2**$^{H*-}$}ip word3 word4 word5
2-1. Ülkedeki **yasak** içkiyi çekici kılıyor. in the country ban the alcohol attractive make "The ban in the country makes alcohol attractive."
2-2. Yoldaki **tuzak** işareti tamamıyla kapatıyor. on the road trap the sign completely blocking "The trap on the road is blocking the sign completely."
2-3. İnşaattaki **tutsak** işleri tamamen aksatıyor. in the construction captive the works completely hindering "The captive in the construction is hindering the works completely."
2-4. Dergideki **korkak** içeriği tamamen değiştirmiş. in the magazine the coward the content completely changed. "The coward in the magazine changed the content completely."
2-5. Camdaki **çatlak** izleri tamamen kapatıyor. on the mirror crack the traces completely blocking "The crack in the mirror is blocking the traces completely."

2.1.2. Participants

Data is collected from five native speakers of Turkish (2 males, 3 females) from Istanbul, Turkey, who have been living in America (Los Angeles) for less than five years. Mean age of participants was 33.8 years. None of the participants reported any speech or hearing disorders.

2.1.3. Recording Procedure

The recordings were done in a quiet room, with participants seated in front of a computer screen. A unidirectional USB microphone was placed to the left of the computer, approximately six inches from the speaker's lips. Speakers repeated twenty-five sentences (randomized between ten target and fifteen filler sentences) five times.

2.1.4. Measurements

- Peak f0, Minimum f0, and Magnitude of f0 rise: The f0-peak was measured ten ms before the end of the vowel in the final syllable of the target word (to avoid any effect of microprosody) and the minimum f0 value was measured on the preceding syllable. Each f0 value was converted into semitones and the magnitude of f0 rise was calculated by the difference between the f0 peak and the minimum f0 values.
- Rhyme Duration: Measured the duration between the beginning of the final vowel in the second word and the beginning of the first vowel in the third word. The beginning of the vowel was defined as the point where the second formant energy begins.
- Degree of V-to-V Coarticulation: Measured by the Euclidean distance in the F1 by F2 space between the final vowel of second word (V1) and the first vowel of the third word (V2), i.e., in $V_1C\#V_2$ context, where # represents a prosodic word boundary in Group 1 and an Intermediate Phrase boundary in Group 2.

2.1.5. Statistical Analysis

Statistical analyses were done in R using Linear Mixed-Effects Regression (LMER) package 'lme4' written by [19], with PROSODIC BOUNDARY (prosodic word/intermediate phrase) as the fixed effect, and SUBJECT and SENTENCE as random effects. The simplest and best fitting model for each LMER analysis was derived via model comparison.

In order to obtain *p*-values for the fixed effects, we used *pvals.func* in the 'languageR' package, which computes the relevant *p*-values using Markov chain Monte Carlo sampling (default number of samples=10,000). When the best fitting model involved random slopes, significance for fixed-effect factors was computed using likelihood-ratio test.

2. 2. Results

2.2.1. Peak f0, Minimum f0, and Magnitude of f0-rise

As predicted, the peak f0 values were higher at the Intermediate Phrase boundary (H*-) than at the Prosodic Word boundary (H*), although the difference was not significant (Intercept=13.25, β=0.73, t=1.84, p=0.07). The minimum f0 value was significantly lower before the H*- syllable than before the H* syllable (Intercept= 12.4, β= -0.71, t= -2.95,

p<0.05). Finally, the magnitude of f0 rise at an Intermediate Phrase boundary was significantly larger than that of a Prosodic Word boundary (β=1.435, t=5.925, p<0.05). Figure 1 shows the mean f0 rise in semitone at the end of the second word, i.e., the last syllable of a Prosodic Word (H*) and an Intermediate Phrase (H*-).

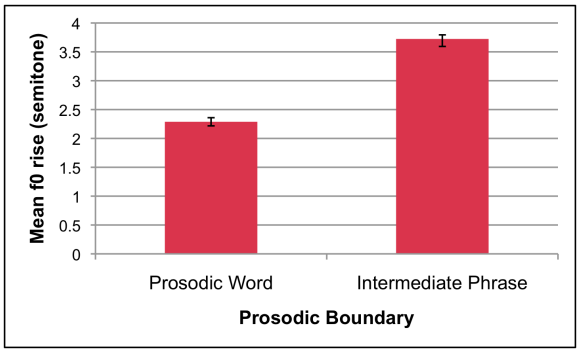


Figure 1: *Magnitude of mean f0 rise at Prosodic Word and Intermediate Phrase boundaries.*

Example pitch tracks of the sentences 1-2 and 2-2 in Table 1 are shown in Figure 2 and 3, respectively. In Figure 2, the second word (*tuzak*) is phrase-medial, thus the f0 peak at the end of the second word receives a pitch accent H*, while the f0 peak of the same word receives H*- in Figure 3 because it marks the end of an Intermediate Phrase.

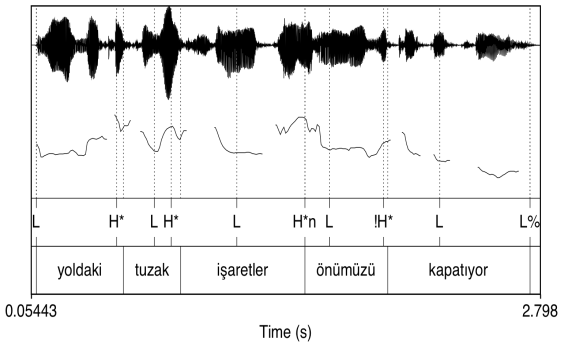


Figure 2: *Sample pitch track of the sentence 1-2 in Table 1, illustrating H* at the end of 2nd word (tuzak), i.e., Prosodic Word boundary.*

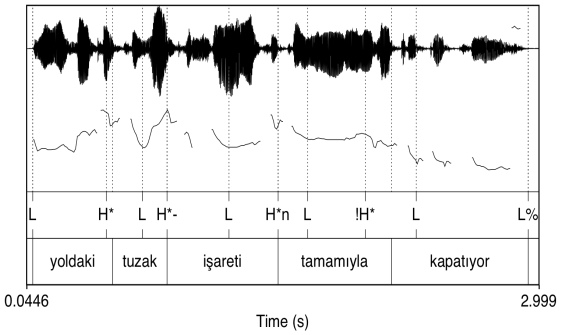


Figure 3: *Sample pitch track of the sentence 2-2 in Table 1, illustrating H*- at the end of 2nd word (tuzak), i.e., Intermediate Phrase boundary.*

2.2.2 Rhyme Duration

Figure 4 shows the mean rhyme duration of the last syllable of the second word at the Prosodic Word and Intermediate Phrase boundaries. As hypothesized, the last syllable of an Intermediate Phrase is longer than that of a Prosodic Word (β=0.067, t=8.38, p < 0.05).

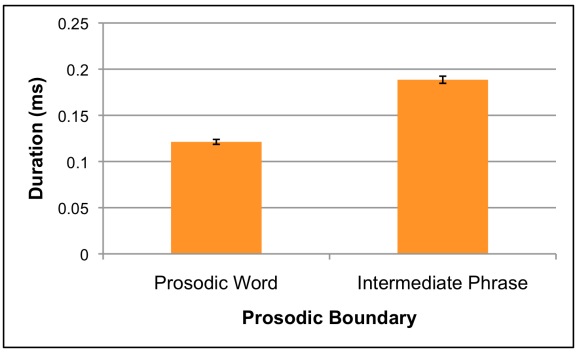


Figure 4: *Rhyme duration of the last syllable at Prosodic Word and Intermediate Phrase boundaries.*

2.3. Degree of Coarticulation

Figure 5 displays the mean Euclidean distance in the F1 by F2 space between each vowel in $V_1C\#V_2$ context. As hypothesized, there is less coarticulation (i.e., larger Euclidean distance) across the Intermediate Phrase boundary than across the Prosodic Word boundary (β=231.1, t=2.145, p<0.05).

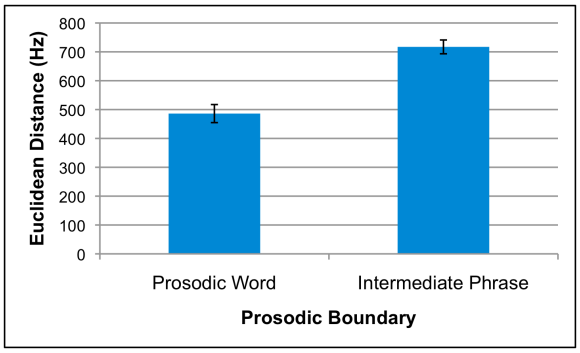


Figure 5: *Euclidean distance across Prosodic Word and Intermediate Phrase boundaries.*

3. Discussion and Conclusion

In this paper, we have shown that the phonetic realization of the high tone on finally-stressed words in Turkish differs depending on its prosodic location. The magnitude of f0 rise is larger at the end of a phrase (Intermediate Phrase/Phonological Phrase/Major Phrase) than in the middle of a phrase. The syllable carrying the high tone is longer when it is the last syllable of an Intermediate Phrase than of a Prosodic Word. Finally, the degree of V-to-V coarticulation is weaker when the two vowels are separated by an Intermediate Phrase boundary than by a Prosodic Word boundary.

The quantitative data found in the experiment therefore support the prosodic structure and tonal categories proposed in Ipek & Jun's model of Turkish intonation. In Ipek & Jun's model, the High tone on the stressed syllable of a prosodic word is labeled as H*, a pitch accent; the High tone realized on the stressed syllable at the end of an ip, however, is labeled as H*-, indicating its function, as both a pitch accent and a boundary tone.

In the experiment reported in the current paper, we only examined finally-stressed words, so an ip-final syllable is always stressed, justifying the H*- symbol. However, as shown in Ipek & Jun [1], the ip-final syllable (or Phonological Phrase/Major Phrase-final syllable in [6, 7]) was still marked by a high tone when the syllable was not stressed. For example, consider Figure 6, where the first two words are non-finally stressed and form a subject NP. Here, each word in the subject NP has a High tone on its stressed syllable (H*) and the second word has an additional High tone on its final syllable, which is also the last syllable of an Intermediate Phrase. Since there is a low f0 target between the two f0 peaks (one over the stressed syllable of the word and the other at the end of the word), and the L tone is closer to the following H tone than the preceding H tone (H*), Ipek & Jun labeled the ip-final f0 rise over an unstressed syllable as LH-. (Note also the shallow falling slope from H* in the first word, interpolating to the initial L boundary tone of the second word in Figure 6). Thus, it is clear that the end of an ip is marked by a high boundary tone regardless of whether the ip-final syllable is stressed or not. However, since the function of the ip-final High tone can differ, the High tone on the ip-final stressed syllable is labeled as H*-, indicating both the prominence marking and the boundary marking functions; the High rising tone on the ip-final unstressed syllable, however, is labeled as LH-, indicating its sole, boundary-marking function.. Further research is needed to investigate whether these tonal categories are perceptually distinct.

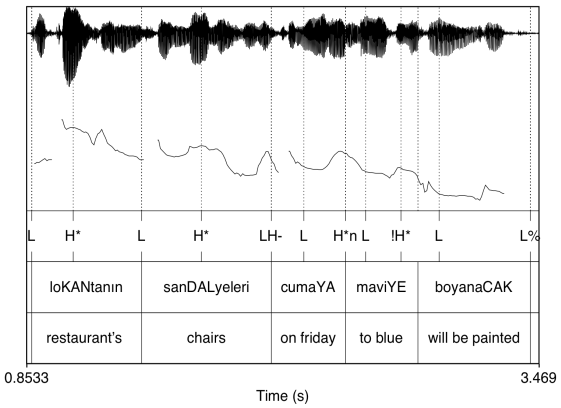


Figure 6: *Sample pitch track of the sentence "Restaurant's chairs will be painted in blue on Friday." The second word (sandalyeleri) is ip-final and is stressed on its second syllable. The second f0 rise on this word, labeled LH-, marks the right edge of ip.*

4. References

[1] Ipek, C. and Jun, S.-A., "Towards a Model of Intonational Phonology of Turkish: Neutral Intonation", in the Proceedings of Meeting on Acoustics(POMA), 9:060230-069238, 2013.

[2] Beckman, M., and Pierrehumbert, J., "Intonational structure in Japanese and English", Phonology Yearbook 3: 255–309, 1986.

[3] Ladd, D. R., "Intonational Phonology", Cambridge University Press, 1996.

[4] Ladd, D. R., "Intonational Phonology", 2nd edition, Cambridge: Cambridge University Press, 2008.

[5] Pierrehumbert, J., "The Phonology and Phonetics of English Intonation", Unpublished Ph.D. dissertation. MIT, 1980.

[6] Kan, S., "Prosodic Domains and the syntax-prosody mapping in Turkish", MA diss. Boğaziçi University, 2009.

[7] Kamalı, B., "Topics at the PF interface of Turkish", Doctoral thesis, Harvard University, 2011.

[8] Levi, S., "Acoustic correlates of lexical accent in Turkish", Journal of the International Phonetic Association 35, 73–97, 2005.

[9] Lees, R., "The phonology of Modern Standard Turkish", Bloomington, IN: Indiana University, 1961.

[10] Lewis, G., "Turkish grammar", Oxford: Oxford University Press, 1967.

[11] Sezer, E., "On non-final stress in Turkish", Journal of Turkish Studies, 5, 61-69, 1981.

[12] Jun, S.-A., "Prosodic Markings of Complex NP Focus, Syntax, and the Pre-/Post-Focus String", in the Proceedings of the 28th WCCFL, pp. 214-230, 2011.

[13] Khan, S. D., "Intonational Phonology and Focus Prosody of Bengali", Unpublished Ph.D. dissertation, University of California, Los Angeles, 2008.

[14] Jun, S.-A and Fougeron, C., "A Phonological Model of French Intonation" in A. Botinis [Ed], Intonation: Analysis, Modeling and Technology, 209-242, Kluwer Academic Publishers, 2000.

[15] Tabain, M., “Effects of prosodic boundary on /aC/ sequences: articulatory results”, Journal of the Acoustical Society of America, 113:516-531, 2003a.
[16] Tabain, M. and Perrier, P., “Articulation and acoustics of /i/ in pre-boundary position in French”, Journal of Phonetics, 33:77-100, 2005.
[17] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J., “Segmental durations in the vicinity of prosodic phrase boundaries”, Journal of the Acoustical Society of America, 91:1707-1717, 1992.
[18] Cho, T., “Prosodic strengthening and featural enhancement: evidence from acoustic and articulatory realization of /a,i/ in English”, Journal of the Acoustical Society of America, 117:3867-3878, 2005.
[19] Baayen, R.H., Davidson, D. J., and Bates, D. M., “Mixed-effects modelling with crossed random effects for subjects and items”, Journal of Memory and Language, 59:390-412, 2008.

The interaction of accent and boundary tone in perception of whispered speech

Willemijn Heeren[1] and Vincent J. van Heuven[1,2]

[1]Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition
Leiden University, The Netherlands
[2]Department of Applied Linguistics, University of Pannonia, Veszprém, Hungary
{w.f.l.heeren,v.j.j.p.van.heuven}@hum.leidenuniv.nl

Abstract

We investigated how the perception of Dutch whispered boundary tones depends on the presence of an accent in the utterance-final word, i.e. the boundary tone landing site. Listeners performed near ceiling in normal speech, whereas the same listeners' performance dropped about 30% in whisper, while processing speed decreased in whisper compared to normal speech. Accent position furthermore influenced boundary tone perception. Initial-stress words showed a question bias that affected recognition of that speech act when accent and boundary tone did not coincide. On final-stress words, in which boundary tone and accent coincided, statements and questions were identified equally well.

Index Terms: boundary tone, nuclear accent, perception, whispered speech

1. Introduction

In whispered speech – where voicing and therefore a fundamental frequency (f0) are absent – listeners still perceive, albeit less reliably than in normal speech, prosodic differences that normally heavily depend on f0 presence. For instance, in whisper listeners recognize questions and statements expressed by different boundary tones (H% versus L%) in cases where prosody, rather than lexico-syntax, codes the crucial information [1-3]. Listeners also discriminate intended pitch height [4], differentiate emotional from neutral speech [5], and identify lexical tones [e.g. 6-8]. Many of these studies, however, assessed perception in single syllables, rather than in multi-syllabic or multi-word phrases, whereas the latter would be more ecologically valid. Multi-syllabic utterances will display some form of ranking as to the relative prominence of those syllables (e.g. imposed by lexical stress).

Though earlier work may indicate that intonation in whisper is perceptible, it does not provide much evidence on the perception of whispered intonation in more complicated linguistic structures. In the present study, we investigated how the perception of Dutch whispered boundary tones depends on characteristics of the tone-bearing word, by using disyllabic minimal stress pairs as boundary tone landing sites. In the case that lexical stress, realized as a nuclear accent, lands in final position, the two tonal events fall on the same syllable. In the case that lexical stress lands in initial position, the tonal events fall on adjacent syllables.

For Dutch, as found in studies on normal speech, the most reliable acoustic correlate of lexical stress in sentence context is relative syllable duration [9, 10]. Perceptually, duration also is a reliable cues to stress [11], but for the perception of prominence, f0 is taken to be the primary cue in Dutch [12], as well as in English [13]. In the absence of f0, expressing intonational contrasts in whispered speech seems to be necessarily more intertwined with segmental characteristics than in the case of normal speech. For instance, in whisper formants are not only used to code vowel identity, but also seem to contribute to expressing differences in height [e.g. 14, 15]. Moreover, if different intonational events land on the same syllable, the restricted resources in whisper may be burdened even further.

To our knowledge, one earlier study has addressed the interaction of accent and boundary tones in whispered speech [1], but in a descriptive manner only. In that investigation, Hungarian listeners classified disyllabic minimal stress pairs that were produced either as question or as statement into one of four categories: two lexical stress positions by two boundary tones. A confusion matrix of classification responses showed that boundary tones were identified above chance level, that declaratives were identified correctly more often than interrogatives, and that accent positions were confused less than boundary tones. In addition, there was confusion across accent positions and boundary tones. For instance, ten percent of final-stress declaratives were identified as initial-stress interrogatives, and such across-tonal event responses seem to support the claim that, in whisper, accents and boundary tones may interact in perception. The same type of utterances was classified without errors in normal speech.

To better understand prosody perception in whispered speech communication, the interaction of accent position and boundary tone perception in Dutch was studied in whispered compared to normal speech. A within-subjects design was used that also included reaction time measurements. We predict that listeners perform better when tonal events do not coincide on the same syllable.

2. Method

Perception of the speech act, i.e. interrogative versus declarative, as expressed through the boundary tone (H% versus L%, respectively) was determined in a classification task with reaction time measurements. Boundary tones were produced on disyllabic nouns with lexical stress, realized as a nuclear accent, in either initial or final position. In the latter case, boundary tone and lexical stress coincide on the same syllable (prosodic clash); in the former case lexical stress falls on the syllable preceding the one carrying the boundary tone (no clash). Minimal stress pairs were used, so that segmental structure would be comparable. To verify that the boundary tone does not alter its bearer's interpretation, perception of the items' stress positions was measured using the same task.

2.1. Materials

Four Dutch minimal stress pairs were used: (1) *'ca·non/ka'non*, /kanɔn/, 'canon/cannon', (2) *'Ser·visch/ser'vies*, /sɛrvɪs/, 'Serbian/crockery set', (3) *'Pla·to/pla'teau*, /plato/, 'Plato; plateau', and (4) *'voor·naam/voor'naam*, /vornam/, 'first name/dignified'. Target words were recorded in a neutral

carrier sentence *Hij zei…* 'He said…', which orthographically ended in either a full stop (to elicit L%) or a question mark (to elicit H%), and which forced the nuclear accent onto the target word, thus establishing the prosodic crowding contrast.

Twelve (self-reported) normal-hearing, Dutch native speakers (6 female) participated in 20-minute recording sessions (informed consent was obtained), and were paid a small amount for their efforts. For each speaker, a different listener was present to judge the recordings. This speaker-listener set-up was intended to prompt the speaker to use listener-directed rather than read speech.

Speakers received written instructions, and completed a short practice session, using different minimal pairs than during the actual recording, for both normal and whispered speech. The order of the speech modes was counterbalanced across speakers. Recordings were made using an Edirol R-44 portable recorder and Røde NTG-2 condenser microphone with 'dead cat' windscreen at 44.1 kHz, 24 bits in a sound-treated booth in the phonetics laboratory of Leiden University. Affirmative and interrogative targets were presented to the speaker one by one and in written form on a computer screen, in a pseudo-random order. The listener was seated outside the booth in a silent classroom, wearing Sennheiser HD 414 SL headphones, and used a keyboard to classify each of the speaker's utterances as affirmative or interrogative. Before the next target was presented, the speaker got feedback about the listener's understanding of the previous one. By keeping the listener outside the booth, and invisible to the speaker, the only cues the speaker could provide were auditory. Two repetitions per utterance were recorded and saved as separate wave files, resulting in 32 files per speaker.

The listeners who were present during the recordings labeled the boundary tones in normal speech correctly in 94% of the cases and in whisper, in 68% of the cases. All speaker-listener pairs were different. Using a 6-point Likert scale (1 = very difficult, 6 = very easy) speakers rated the difficulty of their task for both speech modes. According to a Wilcoxon signed ranks test for paired samples, the task was judged more difficult in whisper (median=3.0) than in normal speech (median = 4.5), $Z = -2.5$, $p = .013$. In neither speech mode was the task judged as particularly easy.

Per lexical item, one instance was annotated manually, and that annotation was used to automatically annotate all other instances of the same item using a dynamic time warping procedure in PRAAT [16]. These annotations were manually checked, and corrected if necessary. Target words were cut from the carrier sentences, and intensity was normalized by setting recordings within a speaker and speech mode to 60 dB (rms = 0.020), which corresponded to the minimum intensity of whispered items after scaling peaks to the maximum intensity range (using PRAAT's 'Scale peaks…'). There were 192 stimuli per speech mode: 8 items (4 initial, 4 final stress) × 2 speech acts (question, statement) × 12 speakers.

2.2. Participants and procedure

Twenty-four, right-handed Dutch native listeners (17 females), aged 19-57 (mean = 22 years), were hearing-screened to have normal hearing at octave frequencies between 0.125 and 8 kHz (informed consent obtained). Each of the 192 items per speech mode was presented once to each listener in a blocked design over tasks. Half of the subjects heard the first half of the materials in the Speech Act classification task, and the second half in the Lexical Stress Position classification task. The other half of the subjects listened to the complementary stimulus sets in each task. The set of materials was halved by including only one boundary tone realization, either H% or L%, per speaker and per target word in each half. Subjects received a small fee for participation in the 45-minute session.

Subjects were seated in a sound-treated booth wearing Sennheiser HD 414 SL headphones. After general instructions in written form, more detailed instructions were presented on a computer screen. Response options were shown on screen, while listeners were asked to press one of two response buttons on a keyboard using their index fingers. During speech act classification listeners indicated whether the target sounded like a question or a statement. During lexical stress position classification listeners indicated whether the initial or the final syllable was more prominent. Both tasks were presented once with normal speech materials, and once with whispered speech, resulting in four subsequent tests. Speech modes, response keys and task orders were counterbalanced across subjects. To allow for within-subjects analyses including the factor speech mode (normal vs. whisper), corresponding whispered and normal speech items from the same speaker were presented to the same subject in the same task.

3. Analysis and results

Percent correct responses was computed for both subtasks, i.e. boundary tone (BT) and lexical stress (LS) classification, and transformed to rationalized arcsine units (RAU) [17]. Reaction times (RTs) were measured from target word onset. RTs under 500 ms and over two standard deviations beyond the mean, computed per listener-per speech mode, were excluded (BT: 2.4% of the data; LS: 4.2% of the data). RTs were transformed to their inverse (1/RT) for analysis. Both RAU scores and inverse RTs were subjected to repeated measures ANOVAs with within-subjects factors Speech Mode (normal, whisper), Speech Act (interrogative, declarative), Lexical Stress Position (initial, final) and Minimal Pair (4). If sphericity was violated, Huynh-Feldt correction was applied.

3.1. Boundary tone classification

Significant effects are presented in Table 1. Figure 1a shows that in whisper, boundary tone classification was significantly poorer than in normal speech (61 vs. 94%, respectively), but above chance level [binomial test: $N = 2304$, $p = ½$, $Z = 11.1$, $p < .001$]. Across speech modes, declaratives were classified correctly more often than interrogatives. The interaction of speech mode by speech act showed that in whisper, the difference in correct responses to declaratives versus interrogatives was larger than in normal speech (whisper: 71 vs. 52%; normal speech: 96 vs. 93%, respectively).

Across speech modes, stimuli with final stress yielded similar scores for the two speech acts, but stimuli with initial stress received more correct responses on declaratives than interrogatives. This interaction was found within both speech modes [normal speech: $F(1,23) = 53.2$, $p < .001$; whisper: $F(1,23) = 31.7$, $p < .001$]. Absolute differences were largest in whisper (see Fig. 1a), where for words with initial stress, declaratives were classified correctly well above chance level at 80%, whereas interrogatives were classified below chance level at 41% [$N = 576$, $p = ½$, $Z = -4.2$, $p < .001$]. On final-stress words, scores were 62.5 and 62.3%, respectively. Trends were comparable between minimal stress pairs, and followed the two main effects. Only for the *Plato/plateau* pair, did

speech mode and speech act interact, showing a larger performance difference between the interrogatives and declaratives across speech modes.

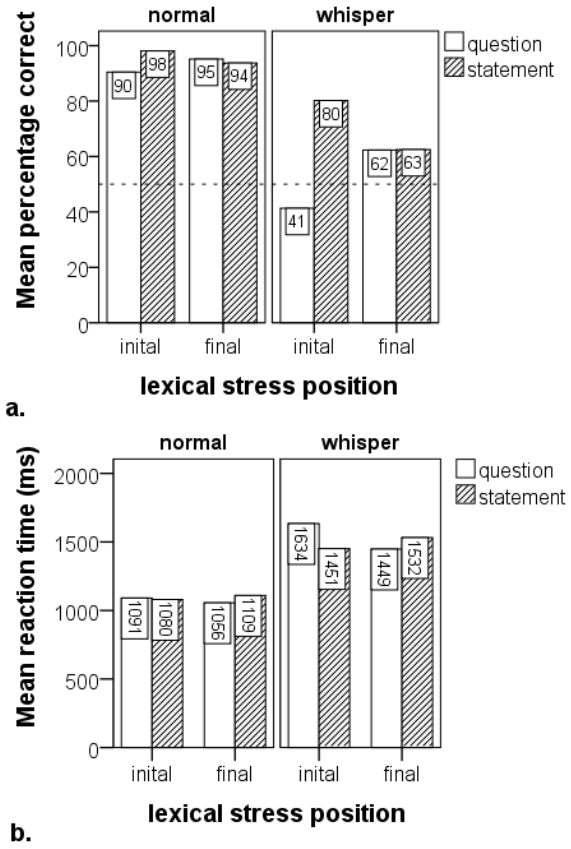


Figure 1: *BT task results, per speech mode, speech act and stress position. a) Mean percentage correct (chance level = 50%). b) Mean reaction time (ms).*

Reaction time results (see Fig. 1b) showed that listeners were faster in normal speech than in whispered speech (1084 and 1502 ms, respectively), faster on words with final stress than with initial stress (1244 and 1255 ms, respectively), and RTs also varied with minimal pair. The effect of stress position, following the main effect, was only significant within whispered speech [$F(1,18) = 7.9$, $p = .012$], as reflected by a marginally significant speech mode by stress position interaction [$F(1,18) = 4.4$, $p = .051$].

Across speech modes, declaratives were responded to equally fast in words with initial and final stress, but interrogatives were responded to faster in final-stress words. This speech act by stress position interaction was found in similar ways in both speech modes [whisper: $F(1,18) = 17.6$, $p = .001$, normal speech: $F(1,23) = 16.5$, $p < .001$]. Across minimal pairs, RTs were longer in whispered than in normal speech, but not exactly to the same extent. Across speech modes, responses to the different minimal pairs followed the main effect of faster responses to words with final stress, but for the *voornaam* pair the trend was in the opposite direction with faster responses to initial-stressed words. The four-way interaction showed that the differences in response times by lexical stress position were mainly caused by differences measured in whisper.

Table 1. *Significant effects and interactions of the RM ANOVAs on BT classification and reaction time data.*

Classification		
Speech mode	$F(1,23) = 403.5$	$p < .001$
Speech act	$F(1,23) = 26.3$	$p < .001$
Minimal pair	$F(3,69) = 3.7$	$p = .016$
Speech mode × speech act	$F(1,23) = 20.2$	$p < .001$
Speech act × stress position	$F(1,23) = 47.8$	$p < .001$
Sp. mode × sp. act × stress pos.	$F(1,23) = 14.8$	$p = .001$
Sp. mode × sp. act × min. pair	$F(3,69) = 5.4$	$p = .002$
Reaction times		
Speech mode	$F(1,18) = 129.1$	$p < .001$
Stress position	$F(1,18) = 6.9$	$p = .017$
Minimal pair	$F(3,54) = 8.9$	$p < .001$
Speech act × stress position	$F(1,18) = 29.6$	$p < .001$
Speech mode × minimal pair	$F(3,54) = 3.6$	$p = .020$
Stress position × minimal pair	$F(3,54) = 3.1$	$p = .033$
Sp. mode × sp. act × min. pair	$F(3,54) = 3.0$	$p = .039$
Sp. mode × sp. act × stress pos.× min. pair	$F(3,54) = 3.0$	$p = .041$

3.2. Lexical stress position classification

Table 2 lists the significant effects for both lexical stress position classification and reaction times. The absence of a speech mode main effect shows that identification of lexical stress position in whisper went as well as in normal speech (89 and 91%, respectively). Across speech modes, there was some variation in mean classification scores per minimal pair, but this difference remained under 4% between the lowest and highest mean scores per pair. Words pronounced as declaratives received more correct responses for initial stress, whereas the correctness of responses to words pronounced as interrogatives was comparable for both stress positions. This trend was observed in both speech modes, but the speech act by stress position interaction was only significant in whisper [$F(1,23) = 18.0$, $p < .001$], not normal speech ($p = .069$).

Table 2. *Significant effects and interactions of the RM ANOVAs on LS classification and reaction time data.*

Classification		
Minimal pair	$F(3,69) = 3.6$	$p = .018$
Speech act × stress position	$F(1,23) = 12.7$	$p = .002$
Reaction times		
Speech act	$F(1,23) = 31.1$	$p < .001$
Stress position	$F(1,23) = 14.4$	$p = .001$
Minimal pair	$F(3,69) = 4.7$	$p = .005$
Speech mode × stress position	$F(1,23) = 19.0$	$p < .001$
Speech act × stress position	$F(1,23) = 11.1$	$p = .003$
Speech act × minimal pair	$F(3,69) = 4.8$	$p = .004$
Sp. act × stress pos.× min. pair	$F(3,69) = 4.0$	$p = .011$

The absence of a speech mode main effect in the RTs indicated that listeners were as fast at identifying lexical stress position in whispered as in normal speech (1304 and 1336 ms, respectively). Across speech modes, responses were faster to declaratives (1279 ms) than to interrogatives (1344 ms), especially for words with initial stress. Responses were faster to final-stress (1289 ms) than to initial-stress (1352 ms) words, but the latter effect did not assume significance for normal speech ($p = .318$), only for whisper [$F(1,23) = 31.2$, $p < .001$].

There was variation in the response times to different words, with fastest responses to *ka'non* (1279 ms) and slowest responses to *'servisch* (1399 ms). The trend for responses to declaratives to be faster than to interrogatives was present in all minimal pairs, but the size of the difference varied between them. Finally, the speech act by stress position interaction was observed in three out of four minimal pairs; for *Plato/plateau*, however, responses to either speech act were equally fast.

4. Discussion

It was expected that in whisper – given the more restricted means of conveying intonation – listeners would have more difficulty correctly identifying boundary tones, and especially when coinciding with lexical stress position. Though the general performance decrease was obtained as expected, the main effect of stress position was not found. Reaction times reflected that processing of final-stress words was in fact somewhat faster, which may hint at easier processing; this effect became significant for whispered stimuli only.

Listener performance showed comparable means around 60% correct on whispered words with initial and final stress; across stress positions, performance was better on declaratives than on interrogatives. This in general suggests that cues to interrogativity were less clear in whisper, which was also found in [1]. But as Fig. 1 shows, performance varied with the stimulus' stress position, especially in whisper. On whispered words with initial stress, where boundary tone and accent do not coincide, performance was much better on declaratives than on interrogatives. Effectively, at 40% correct, questions were not recognized as such on words with initial stress. Moreover, RTs were generally longer for this type of stimulus. For whispered words with final stress, performance was comparable between the speech acts. These results suggest that only on whispered stimuli in which accent and boundary tone coincided on the same (i.e. final) syllable (clash condition), were listeners able to reliably identify the boundary tone.

As performance on interrogatives pronounced on initial-stress words was, in fact, below chance level, we looked for potential response biases in the data. Chi square analyses per speech mode per lexical stress position revealed that in both speech modes, listeners gave a majority of 'interrogative' responses to words with initial stress, whereas equal numbers of either response category were expected (normal speech: 543 out of 1008 'interrogative' responses, $\chi^2(1) = 6.0$, $p = .014$; whisper: 711 out of 1008 'interrogative' responses, $\chi^2(1) = 170$, $p < .001$). In normal conversational speech, statements occur (much) more often than questions [e.g. 18]. As speech perception generally reflects differences in the token frequencies of categories, we expect listeners to respond with the statement category unless there is clear evidence to the contrary. This was not what listeners did. Possibly, listeners interpreted the accent in initial position as prominence in a more general sense that was then associated with an interrogative reading of the utterance as a whole. Alternatively, first-syllable prominence may have been interpreted as a direct cue to a potentially upcoming question, which was not overruled by evidence provided in the relatively weak final syllable. [19] showed that the size of an object accent influenced listener expectations about whether an utterance was a statement or a question: larger object accents triggered stronger question expectation. In the present study, the accent on the first syllable may have similarly signaled a potentially upcoming question, especially when f0 was absent.

For words with lexical stress on the same (i.e. final) syllable as the boundary tone, whispering speakers were able to convey the speech act. In comparison with other studies on boundary tone identification in whisper [2, 3], the task seems to have been relatively more difficult in the present study. The difference may be due to higher demands placed on processing by the two intonational events in close proximity. On the one hand, this is taken to reflect that more complicated linguistic structures, as in the present study, may moderate earlier results on the processing of prosody in whisper (see also [1]). On the other hand, the exclusive use of minimal pairs in the present study may have made listeners aware of the lexical contrast in addition to the speech act difference, also during the boundary tone task, which in turn may have influenced performance.

There was a large difference in the mean reaction times to whispered versus phonated boundary tones. This cannot be explained by the difference in stimulus duration between the speech modes, as this difference was only on the order of 100 ms (693 vs. 583 ms means for whispered and normal speech items, respectively), whereas the reaction time difference was on the order of 400 ms. The slower responses in whisper therefore seem to mainly reflect an increase in processing time due to a difference in cues to boundary tones between the speech modes, including the absence/presence of f0.

Over the same set of stimuli, listeners classified lexical stress position with high accuracy and with similar reaction times in the two speech modes. These results are consistent with the finding that lexical stress position is most reliably realized by durational differences [10], which also form a main cue for listeners [11]. A planned acoustic analysis of the data is expected to reflect the presence of durational information. Moreover, we take the high listener scores to indicate that lexical meaning was generally not influenced by boundary tone realization, in either speech mode.

Words with initial stress pronounced as declaratives were more often identified correctly with respect to stress position than their counterparts with final stress, whereas no difference was found when the same words had been pronounced as interrogatives. Mainly in whisper, responses to interrogatives were around 60 ms faster on words with final stress than on words with initial stress, whereas the average durational difference between the words types was very small (~10 ms). This hints at a small processing benefit for the former type of words, which may be explained by a smaller demand on short term memory for items with final stress.

In sum, depending on the listening task, the same stimuli were responded to very differently. Whispered speech was as clear as normal speech with respect to lexical stress position. For boundary tone perception, however, listeners performed near ceiling in normal speech, whereas the same listeners' performance dropped about 30% in whisper, while processing speed decreased significantly. Accent position furthermore influenced boundary tone perception. Initial-stress words showed a question bias that affected recognition of that speech act. On final-stress words, in which boundary tone and accent coincided, the speech acts were identified comparably.

5. Acknowledgements

This work was supported by a VENI grant made available to the first author by the Netherlands Organisation for Scientific Research (NWO).

6. References

[1] Fónagy, J. (1969). “Accent et intonation dans la parole chuchotée,” Phonetica, 20:177-192, 1969.

[2] Heeren, W. F. L. and Van Heuven, V. J., “Perception and production of boundary tones in whispered Dutch”, in Proc. Interspeech 2009, Brighton, 2411-2414, 2009.

[3] Heeren, W. F. L. and Lorenzi, C., “Perception of prosody in whispered French”, J. Acoust. Soc. Am., to appear.

[4] Higashikawa, M., Nakai, K., Sakakura, A. and Takahashi, H., “Perceived pitch of whispered vowels–relationship with formant frequencies: a preliminary study”, J. Voice, 2:155-158, 1996.

[5] Tartter, V. C. and Braun, D. “Hearing smiles and frowns in normal and whisper registers”, J. Acoust. Soc. Am., 96:2101-2107, 1994.

[6] Abramson, A. S., “Tonal experiments with whispered Thai”, in A. Valdman [ed.], Papers on linguistics and phonetics to the memory of Pierre Delattre, 29–44, The Hague: Mouton, 1972.

[7] Miller, J. D., “Word tone recognition in Vietnamese whispered speech”, Word, 17:11-15, 1961.

[8] Liu, S. and Samuel, A. G., “Perception of Mandarin lexical tones when F0 is neutralized”, Lang. Speech, 47:109-138, 2004

[9] Nooteboom, S.G., “Production and Perception of Vowel Duration. A Study of durational Properties of Vowels in Dutch”, Unpublished Doctor’s Thesis, Utrecht: University of Utrecht, 1972.

[10] Sluijter, A. and Van Heuven, V. J., “Spectral balance as an acoustic correlate of linguistic stress”, J. Acoust. Soc. Am., 100:2471-2485, 1996.

[11] Sluijter, A. M., van Heuven, V. J. and Pacilly, J. J., “Spectral balance as a cue in the perception of linguistic stress”, J. Acoust. Soc. Am., 101:503-513, 1997.

[12] Van Katwijk, A., “Accentuation in Dutch”, Amsterdam/Assen: Van Gorcum, 1974.

[13] Fry, D. B., “Experiments in the perception of stress”, Lang. Speech, 1:126-152, 1958.

[14] Higashikawa, M. and Minifie, F. D., “Acoustic–perceptual correlates of “whisper pitch” in synthetically generated vowels”, J. Speech, Lang. Hear. Res. 42:583-591, 1999.

[15] Meyer–Eppler, W., “Realization of prosodic features in whispered speech,” J. Acoust. Soc. Am., 19:104-106, 1957.

[16] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer [Computer program],” retrieved from http://www.praat.org/, 2013.

[17] Studebaker, G. A., “A "Rationalized" Arcsine Transform”, J. Speech Hear. Res. 28:455-462, 1985.

[18] Van Heuven, V. J., Haan, J. and Pacilly, J. J., “Global and local characteristics of Dutch questions in play-acted and spontaneous speech”, in Proc. ESCA workshop on sound patterns of spontaneous speech, La Baume-les-Aix, 139-142, 1998.

[19] Van Heuven, V. J. and Haan, J., “Temporal development of interrogativity cues in Dutch”, in C. Gussenhoven and N. Warner [Eds.], Laboratory Phonology 7, 61-86, Berlin: Mouton de Gruyter, 2002.

Links between Manual Punctuation Marks and Automatically Detected Prosodic Structures

Katarina Bartkova[1], *Denis Jouvet*[2]

[1] ATILF - Analyse et Traitement Informatique de la Langue Française
Université de Lorraine, ATILF, UMR 7118, Nancy, F-54063, France
[2] Speech Group, LORIA
Inria, Villers-lès-Nancy, F-54600, France
Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
katarina.bartkova@atilf.fr, denis.jouvet@loria.fr

Abstract

This paper presents a study of the links between punctuation and automatically detected prosodic structures, as observed on large speech corpora that were manually annotated during French speech transcription evaluation campaigns. These corpora contain more than 3 million words and almost 350 thousands punctuation marks. The detection of the prosodic boundaries and of the prosodic structures is based on an automatic approach that integrates little linguistic knowledge and mainly uses the amplitude and the direction of the F0 slopes as described in [1], as well as phone durations. The paper first analyzes the occurrences of the punctuation marks with respect to various sub-corpora, which also highlights the variability among annotators. Then, we focus on analyzing prosodic parameters with respect to the punctuation marks, followed or not by a pause, and on analyzing the links between the automatically detected prosodic structures and the manually annotated punctuation marks.

Index Terms: prosodic structure, speech, punctuation

1. Introduction

Speech is structured by prosodic means to allow the listener to access to lexical units and therefore to the meaning conveyed by the speech signal. For optimal results, most of the automatic speech processing techniques (automatic translation, information retrieval…) need to recover the speech prosodic structuring. This corresponds to adding punctuation marks to the raw streams of words supplied by automatic speech transcription systems. Though orthographic conventions used to capture the speech prosody cannot reflect all the various linguistic and extra-linguistic meanings of the speech prosody, however, they allow to mark its most elementary linguistic functions such as the modality and the finality of a sentence through full stops, exclamation marks, question marks… and through commas the intention of the speaker after a deep prosodic boundary (prosodic group inserted closer to the root than to the leaves in the prosodic tree, cf. Figure 1) to continue to develop the same sentence.

The pattern of the prosodic parameters (mainly F0 movement and syllable duration lengthening) depends on the type of the prosodic boundary. A continuation (major – deeper boundary; or minor – shallower boundary) is indicated by the slope of the F0 movement whose direction can be rising or falling. The end of a declarative sentence is indicated by a falling F0 movement [2],[1] although other (flat or rising) movements are also observed in French spontaneous speech [3]. The modality of the sentence is expressed by the direction and the steepness of the F0 movement: polar (yes-no) questions in French are marked only by a sharply rising F0 slope while an order is marked by a sharply falling F0 slope.

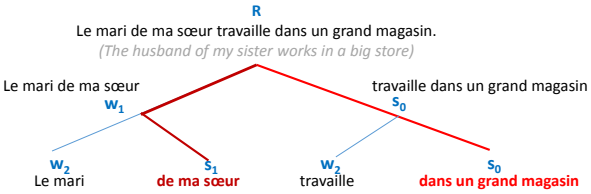


Figure 1: *Metric tree with the deepest prosodic boundaries at nodes s_1 and s_0 (little cohesion to the right)*

In French the syllable duration is closely related to the prosodic boundaries: the duration of a syllable on a continuation prosodic boundary is significantly lengthened compared to the duration of syllables on non-prosodic boundaries. Moreover, the syllable duration, although lengthened on sentence final position, is however shorter than on clause final (major or minor continuation) position [4].

The automatic detection of the prosodic structure used in this study, is based on a theoretical description of prosodic trees; the framework was first developed for prepared speech [1] and later adapted for semi-spontaneous speech [5]. The approach was recently revisited and applied on various types of speech material, including spontaneous speech [6].

In this study, the location of the manually set punctuation marks is compared with prosodic units detected by the automatic approach described in [6]. This comparison allows us observing whether human annotators were or were not influenced by the prosodic parameter values considered as pertinent for the automatic segmentation and also if there is coherence between the location of the punctuation marks and the depth of the prosodic boundaries estimated from their prosodic parameters.

The paper is organized as follows. Section 2 presents the speech corpora used as well as some global statistics about the punctuation marks observed in the manually transcribed data. Section 3 summarizes the process used to obtain automatically the prosodic structures of the speech data. Then, Section 4 analyzes the relations between the punctuation marks and the prosodic parameters, the prosodic groups and the prosodic structures.

2. Speech corpora and punctuation

Several French speech corpora have been used in this study. They come from the recent ESTER2 [7] and ETAPE [8] speech transcription evaluation campaigns, and from the EPAC project [9],[10]. The ESTER2 data are French broadcast news collected from various radio channels; thus mainly prepared speech, plus some interviews. The EPAC data are mainly spontaneous speech, recorded during the ESTER1 campaign, and correspond to shows from three French radios [9]. The ETAPE data corresponds to debates collected from radio and TV channels, and is mainly spontaneous speech. The training part of the ESTER2 and ETAPE data, plus the transcribed part of the EPAC corpus correspond to a total of almost 300 hours of signal and 4 million running words. The development and test parts of the corpora have been left aside for further experiments.

The values of F0 in semitones and of the energy are computed every 10 ms from the speech signal using the ETSI/AURORA [11] acoustic analysis. The phonetic transcription of the text, with pronunciation variants, is obtained using the BDLEX [12] lexicon and an automatic grapheme-to-phoneme transcription system [13] which is applied for words absent from the lexicon. The forced text-speech alignment is carried out with the Sphinx tools [14]. This provides the speech segmentation into phonemes and words, which is then used to compute the sound durations, as well as to obtain the location and the duration of the pauses. As the speech signal quality is rather good, it can be assumed that the segmentation is obtained out without major problems. After the forced alignment step, short pauses of less than 100 ms occurring before a plosive or a fricative are removed. Finally, end of speech events, as well as last word before a speaker change, falls in the columns "plus pause" in the following tables.

Punctuation was present in most of the manual transcription files, which correspond to more than 3 million words (lexical and grammatical words). Table 1 exhibits some global statistics for each of the three corpora. There is on average one punctuation symbol every 8 to 10 words (every 8.3 words for EPAC, every 9.8 words for ETAPE).

Table 1. *Size of the speech corpora used, and number of words and punctuation symbols with respect to position (followed or not by a pause).*

	Files	Words			Punctuation symbols		
		count	plus pause	no pause	count	plus pause	no pause
ESTER	292	2.01 M	18.2%	81.8%	225 k	66.7%	33.3%
EPAC	106	0.92 M	18.5%	81.5%	94 k	50.9%	49.1%
ETAPE	44	0.24 M	20.8%	79.2%	29 k	61.9%	38.1%

With respect to pauses, the three corpora have a similar behavior, on average a pause is observed after every 5 words (about 20% of the words are followed by a pause). However there are more differences for what concern manually set punctuation symbols. About one third of the punctuation symbols are not followed by a pause for the ESTER and ETAPE corpora, whereas this is almost one punctuation symbol out of two which is not followed by a pause in the EPAC corpus. This exhibits a rather large variation in punctuation annotation which is due to annotators as the high frequency of punctuation symbols not followed by a pause in the EPAC corpus does not match neither with the annotations on ESTER (similar radios) nor with the annotations on ETAPE (also mainly spontaneous speech).

Table 2. *Analysis of the various punctuation marks in the three speech corpora.*

	dot	excl.	inter.	3 dots	semi col.	two dots	comma
ESTER	32.9%	1.8%	2.3%	0.8%	1.3%	2.6%	58.3%
EPAC	23.0%	1.6%	4.2%	0.0%	5.9%	5.5%	59.8%
ETAPE	32.0%	2.0%	5.1%	0.5%	0.2%	2.3%	57.9%

Table 2 presents the distribution of the various punctuation marks in the 3 corpora. The frequency of the comma is rather similar between the 3 corpora. However, other punctuation symbols on EPAC exhibit a different behavior, especially with respect to semi-column and two dots marks which are much more used by the annotators of the EPAC corpus than by the annotators of the ESTER and ETAPE corpus.

3. Automatic detection of prosodic structures

As mentioned before, the approach used for the automatic detection of prosodic structures is based mainly on prosodic parameter values and little linguistic knowledge. The process starts by an initial segmentation of the text (i.e., the sequence of words corresponding to the speech signal) into potentially stressed prosodic units. This is carried out by grouping grammatical words with lexical words. Prosodic parameters are then considered only on the vowels of the last syllables of the potentially stressed groups. Two main parameters, the F0 slope and the normalized duration of the vowels in final positions (other than the schwa vowel in final position when the word is plurisyllabic) are used to detect prosodic boundaries. The duration threshold that separates stressed vowels from unstressed vowels was determined from the analysis of the distribution of the duration of vowels in unstressed positions (only syllables other than last syllables of the lexical units were considered for this estimation) and in stressed positions (only syllables followed by a pause were considered for this estimation). A similar approach was applied for determining the threshold that separates the values of the slopes of F0 on prosodic and non-prosodic boundaries. A third parameter, the variation of the F0 value (obtained as the difference in F0 between the current vowel and the previous one not separated from the current vowel by a pause) is also calculated for the last vowel in the prosodic groups. If this variation of the F0 value is higher than 5 semi-tones then a prosodic boundary is set on this syllable.

The automatic approach also evaluates the depth of the prosodic boundaries. A prosodic boundary which is marked by a steep F0 slope (higher than the glissando threshold for speech) and a long vowel duration, or a very long vowel duration and a more moderate F0 slope (higher than the glissando threshold for vowels), receives the symbolic annotation C0. To capture smaller variations of the prosodic parameters, symbolic annotation ranging from C1 (deeper prosodic boundary) to C5 (shallower prosodic boundary) are used. To avoid a too fine-grained prosodic segmentation, prosodic boundaries whose symbolic annotation is C3 and whose length is less than 2 syllables, are neutralized and attached to the following prosodic group. Also, when the prosodic group exceeds 10 syllables, an intermediate prosodic boundary is searched around the middle part of the group

using this time lower threshold values for vowel duration and F0 slope detection. When an appropriate split is found, the prosodic group is cut into 2 groups; otherwise the group regardless of its length is maintained as one single prosodic group. The symbolic annotations (C0, C1, ...) are used to construct prosodic trees for each breath group (speech signal preceeded and followed by a long pause). In the prosodic tree construction, a prosodic group is attached to the next prosodic group if the next group has a lower symbolic depth (i.e. if the next group is closer to the root of the prosodic tree).

4. Prosodic structure and punctuation

This section is dedicated to analyzing in details, on the ESTER training data, the punctuation marks with respect to the prosodic parameters, the prosodic groups and the prosodic structures that were automatically obtained. Besides Table 3, most of the analyzed items are presented as normalized frequency histograms; that means that the figures show the distribution of some parameters (e.g., pause duration, F0 slope …) with respect to the presence or absence of punctuation symbols at the end of prosodic groups. To keep the figures simple, only the dot and the comma punctuation symbols are considered, plus the "*no punctuation*" case, corresponding to end of prosodic groups that are not associated to any punctuation symbol. Moreover, two histograms are drawn in each case, whether a pause follows or not.

4.1. Punctuation and prosodic groups

The first analysis concerns the position of the punctuation marks with respect to the automatically obtained prosodic groups. Table 3 shows that most of the punctuation marks that are followed by a pause (column *plus pause*) match with the end of automatically detected prosodic groups. Moreover, almost two thirds of the punctuation marks that are not followed by a pause also match with the end of prosodic groups. Table 3 shows that, overall, less than 14% of the punctuation marks fall inside the automatically detected prosodic groups. For these cases, if we left aside the few cases were the detection of the prosodic groups is not correct, the annotator's decision was probably influenced more by semantic or syntactic information not marked by prosodic parameters.

Table 3. *Analysis of the various punctuation marks with respect to position (any place or end of automatically detected prosodic groups).*

ESTER		Any place		End prosodic group	
		plus pause	no pause	plus pause	no pause
dot	74154	89.0%	11.0%	89.0%	7.3%
excl.	4026	89.1%	10.9%	88.1%	5.9%
inter.	5163	86.6%	13.4%	86.6%	8.6%
3 dots	1742	84.3%	15.7%	83.5%	9.3%
semi-col.	3018	81.3%	18.7%	80.1%	12.4%
two dots	5969	65.8%	34.2%	63.3%	22.6%
comma	131280	52.1%	47.9%	50.3%	29.7%
Total	*225352*	*66.7%*	*33.3%*	*65.5%*	*20.8%*

For what concerns the length of the prosodic groups, Figure 2 shows that prosodic groups reduced to a single word are almost never followed by a punctuation mark. Prosodic groups followed by a punctuation mark are longer than those not associated to any punctuation mark. Also, the distributions of the length of the prosodic groups followed by a dot or by a comma are very similar. Moreover, the distribution of the length of the prosodic groups is not significantly different whether the prosodic groups are followed by a pause or not.

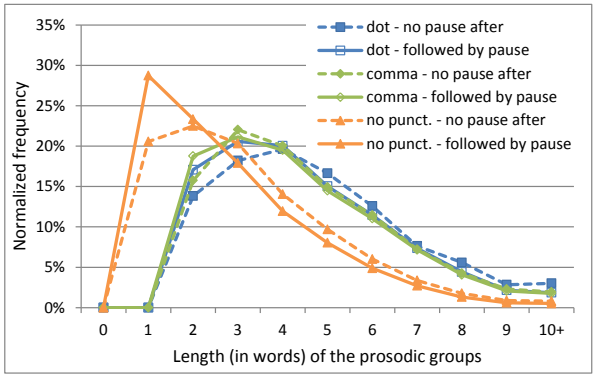


Figure 2: *Normalized frequency histograms of the length of the prosodic groups*

Figure 3 presents the distributions of the pause durations when observed after a prosodic group. When there is no punctuation associated to the prosodic group, the duration of the following pause is most of the time smaller than 200 ms. When a punctuation mark is associated to the prosodic group, the duration of the following pause if typically between 300 ms and 500 ms. Overall, the duration of the pause following a prosodic group tends to be longer for prosodic groups associated to dots than for prosodic groups associated to commas, which are themselves longer than for prosodic groups that are not associated to any punctuation mark.

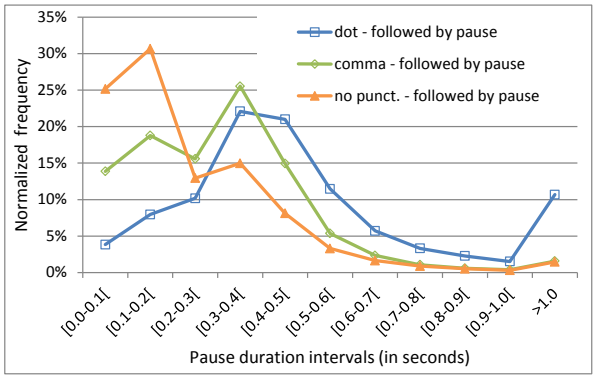


Figure 3: *Normalized frequency histograms of the duration of the pauses after prosodic groups.*

4.2. Punctuation and prosodic parameters

The first prosodic parameter considered here is associated to the global F0 slope, that is the longest F0 slope that ends in the last syllable of the prosodic group. Figure 4 presents the normalized histograms of the absolute variation of F0 (delta F0) between the beginning and the end of the global slope. The delta F0 (Figure 4) and the F0 slope (not represented here) are slightly higher for prosodic groups associated to the dot and comma punctuation marks than for prosodic groups not associated to any punctuation mark. That means that this

parameter is not perceived as pertinent by human annotators for the decision of the punctuation marks.

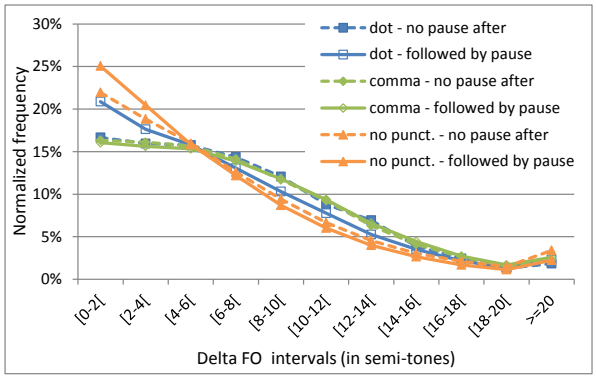


Figure 4: *Normalized frequency histograms of the delta F0 absolute values of the global slope ending in each prosodic group.*

The analysis of the last F0 value on the prosodic groups with respect to the associated punctuation mark is reported in Figure 5. The F0 value was first normalized according to the speaker F0 range, and thus expressed as a percentage of the F0 speaker range (0.0 meaning the lowest F0 value, 1.0 meaning the highest F0 values for the given speaker). It appears from the figure that the highest last F0 values are frequently associated to the comma punctuation mark, intermediate last F0 values are generally not associated to any punctuation mark, whereas dots are associated either to low last F0 values when the prosodic group is followed by a pause or to high last F0 values when there is no following pause.

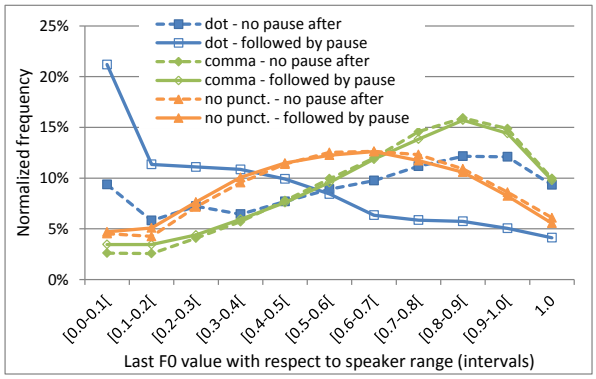


Figure 5: *Normalized frequency histograms of the F0 values at the end of each prosodic group (F0 value expressed as a ratio of the F0 speaker range).*

4.3. Punctuation and prosodic structure

Figure 6 illustrates the relation between the punctuation marks and the level of the prosodic group in the prosodic structures detected automatically. The figure shows that when there is no following pause, the dot and comma are almost always associated to prosodic groups of level 0, i.e., the prosodic group with the deepest boundary in the considered speech segment. However, many prosodic groups of level 0 are also associated to the no punctuation case. When there is no following pause, dot and comma punctuation marks are frequently associated to prosodic groups of level 1. Higher level prosodic groups (level 2 or more) are more frequently observed for the no punctuation case.

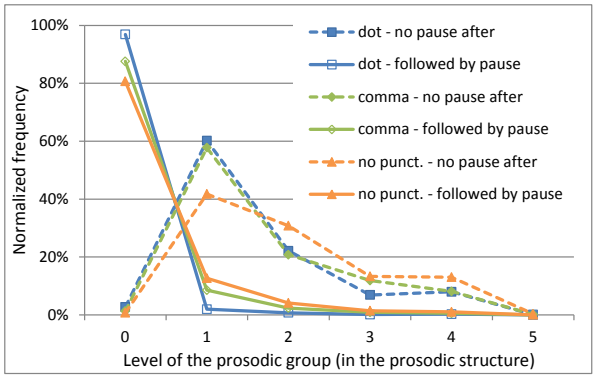


Figure 6: *Normalized frequency histograms of the level of the prosodic group (level in the associated prosodic structure).*

A more detailed analysis shows that for the no punctuation case and for the prosodic groups of level 1, there is a tendency to relate the punctuation symbol to the level of the following prosodic group: a dot if the following group is of level 2 or more, a comma if the following group is also of level 1, and no punctuation if the following group is of level 0.

5. Conclusions

This paper has analyzed the links between the punctuation marks and the prosodic parameters and prosodic structures of the speech data. The analysis was conducted on French speech corpora that were manually transcribed and annotated for speech transcription evaluation campaigns. Several hundred hours of signal were considered. A first statistical analysis of the punctuation marks on several sub corpora showed that the annotation of the punctuation bears a rather large variability due to human annotators.

The prosodic structures of the speech data were obtained automatically through an approach that integrates little linguistic knowledge; the approach mainly relies on the amplitude of the F0 slopes, as well as on phone durations. Most of the manually set punctuation marks match with the end of automatically detected prosodic groups (few punctuation marks fall inside automatically detected prosodic groups).

Prosodic groups, prosodic parameters and prosodic structures were also analyzed with respect to the presence or absence of punctuation marks, whether they are followed or not by a pause. Two punctuation marks were particularly studied: dot and comma. Parameters were analyzed through normalized frequency histograms revealing different behaviors for dot, comma and no punctuation case occurrences.

However, the distribution of the parameters still largely overlap, and each of the prosodic parameters cannot be used alone to decide on the punctuation (if any) that should be associated to the end of a prosodic group. Further studies will investigate the application of automatic classifiers that could handle simultaneously all the parameters for deciding on the presence and on the type of punctuation mark at the end of a prosodic groups.

6. References

[1] Martin, P.: “Prosodic and rhythmic structures in French”. *Linguistics* 25, pp. 925–949, 1987.

[2] Delattre, P.: “Les dix intonations de base du français”. *The French Review* 40 (1), pp. 1-14, 1966.

[3] Avanzi, M., Martin, P.: "L'intonème conclusif : une fin (de phrase) en soi ?". *Nouveaux cahiers de linguistique française*, 28. pp. 247-258, 2007.

[4] Bartkova K., Sorin, C. “A model of segmental duration for speech synthesis in French”. *Speech Communication* 6 (3), pp. 245-260, 1987.

[5] Segal, N., Bartkova, K.: “Prosodic structure representation for boundary detection in spontaneous French”. In *Proc. ICPhS 2007*, Saarbrücken, Germany, pp. 1197–1200, 2007.

[6] Bartkova, K., Jouvet, D.: “Automatic Detection of the Prosodic Structures of Speech Utterances”. In *Proc. SPECOM 2013*, Pilsen, Czech Republic, pp. 1-8, 2013.

[7] Galliano, S., Gravier, G., Chaubard, L.: “The Ester 2 evaluation campaign for rich transcription of French broadcasts”. In *Proc. INTERSPEECH 2009*, Brighton, UK, pp. 2583–2586, 2009.

[8] Gravier, G., Adda, G., Paulsson, N., Carr, M., Giraudel, A., Galibert, O.: “The ETAPE corpus for the evaluation of speech-based TV content processing in the French language”. In *Proc. LREC 2012*, Istanbul, Turkey, 2012.

[9] Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., Farinas, J.: “The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news”. In *Proc. LREC 2010, European Conf. on Language Resources and Evaluation*, Valetta, Malta, 2010.

[10] Corpus EPAC: Transcriptions orthographiques. Catalogue ELRA, reference ELRA-S0305, http://catalog.elra.info.

[11] Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression Algorithms, ETSI ES 202 212, 2005.

[12] de Calmès, M., Pérennou, G.: “BDLEX: a Lexicon for Spoken and Written French”. In *Proc. LREC 1998*, Grenade, pp. 1129–1136, 1998.

[13] Jouvet, D., Fohr, D., Illina, I.: “Evaluating grapheme-to-phoneme converters in automatic speech recognition context”. In *Proc. ICASSP 2012*, Kyoto, Japan, pp. 4821–4824, 2012.

[14] Sphinx (2011), http://cmusphinx.sourceforge.net/

Perception of Peak Placement in Tashlhiyt Berber

Timo B. Roettger[1], Rachid Ridouane[2] & Martine Grice[1]

[1]IfL Phonetik, University of Cologne;
[2]Laboratoire de Phonétique et Phonologie (UMR 7018) CNRS/Sorbonne Nouvelle
timo.roettger@uni-koeln.de; martine.grice@uni-koeln.de; rachid.ridouane@univ-paris3.fr

Abstract

Previous production studies on Tashlhiyt Berber have demonstrated that questions and statements have similar intonation contours, i.e., a final rise to a F0 peak and subsequent fall. The contours tended to differ in overall pitch register and peak location: questions (a) revealed a stronger tendency to be realized with the F0 peak on the final syllable than statements and (b) even within the same syllable, peaks were often aligned later in questions than in statements. The peak location, however, was reported to vary strongly both within and across speakers, interpreted as free alternation of tonal association. Given this high degree of variation, the question arises as to how relevant this variation is for communication. The present perception study shows that both pitch register (low vs. high) and tonal placement (peak on penultimate vs. final syllable) affect listeners' judgments on sentence modality as well as reaction times. Whereas peak alignment within the syllable (early vs. late) did not affect judgments, it did have a marginal effect on reaction times. By demonstrating their perceptual impact, this study confirms that the patterns found in production are communicatively relevant.

Index Terms: Tashlhiyt Berber, intonation, tonal association, tonal alignment, pitch register

1. Introduction

Berber is an Afro-Asiatic language spoken across large parts of North Africa. Tashlhiyt Berber, the variety investigated here, is one of three major Berber varieties in Morocco.

Recent studies on Tashlhiyt have shown that the intonation of this language involves a high degree of variation both within and across speakers [1,2,3,4]. In particular, it was shown that in polar questions and contrastive statements pitch peaks consistently co-occur with available sonorant nuclei [3,4]. The authors accounted for this placement in terms of a phonological association. More specifically, the tone is analysed as an edge tone, H, of a prosodically defined phrase with a secondary association to a tone bearing unit, a syllable with a sonorant nucleus. However, there was considerable variation with regard to which sonorant nucleus the peak co-occurs with. In fact, speakers showed a certain degree of free alternation: Consider the examples in Figure 1 which demonstrate this apparent free alternation. The same speaker produced the sentence /inːa tugl/ (*'He said 'she held''*), embedded in the same context, with an intonational peak on the final syllable (Top) or on the penultimate syllable (Bottom). Nonetheless, this alternation was not found to be entirely free. Even though there was a strong trend to realize the H on the final syllable in both types of utterances, there were more occurrences of H on the final syllable in questions than in statements. Impressionistically the peak placement goes hand in hand with some degree of segmental prominence, i.e., the syllable the peak co-occurs with is longer and louder than it would be without a peak (compare /u/ and /l/ in /tugl/ of Figure 1 with and without an intonational peak)

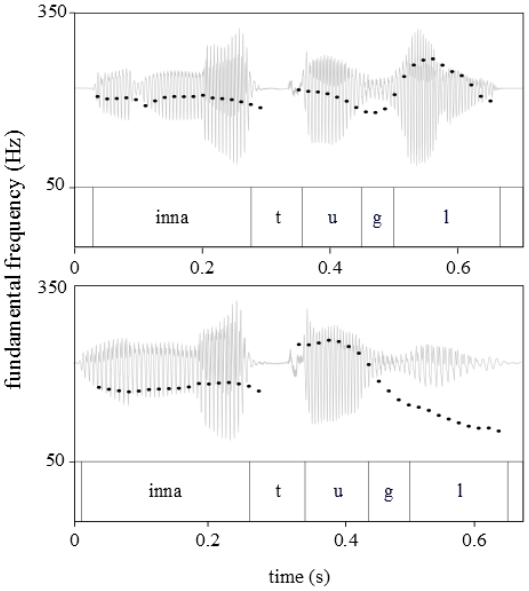


Figure 1: *Waveform and F0 contours of statements /inːa tugl/ (3ms-say 3fs-held) 'He said 'she held'' produced by the same speaker in the same context. Top: Peak placed on final syllable; Bottom: peak placed on penultimate syllable.*

Moreover, this discrete asymmetry, interpreted as phonological association of the H tone, was reflected in more gradual phonetic detail, i.e., the phonetic alignment within the syllable [cf. 4]. Peaks in questions were aligned later in the syllable than in statements. This tendency appeared to be highly speaker dependent. In addition to these differences in peak placement, questions differed from statements in terms of pitch register. Questions were produced with a higher baseline and a concomitant steeper rise to the pitch peak than statements.

The discussed different preferences in questions and statements have to be considered as functionally motivated to distinguish sentence modality. The remaining variability, however, might reflect the redundant nature of peak location in Tashlhiyt. Polar questions are marked by an initial preverb /is/. This morphosyntactic marker could already be sufficient for indicating sentence modality. Tonal cues might thus be redundant to some degree. However, there is also an echo question in this language (inːa tugl? '*He said* '*she held*'*?*'). This type of question is morphosyntactically identical to statements, and differs from statements by intonation only. The present paper reports on a perception study which investigates the contribution of tonal cues to the discriminability of the contrast between statements and echo questions.

On an alternative account, the observed variability could be due to the sample tested in previous studies. The production data of [3,4] was based on Tashlhiyt speakers that have lived in

Paris for a significant amount of time, which might have had an influence on their productions. Thus, the present study attempts to validate the obtained patterns in production (based on speakers living in France) by testing these contours perceptually with Tashlhiyt speakers living in Morocco.

2. The present study

To summarize, polar questions and statements in Tashlhiyt have been mainly shown to differ according to following dimensions:

(a) Questions are more likely to have the peak on the final syllable than statements.
(b) Peaks in questions are aligned later within the syllable than peaks in statements.
(c) Questions reveal a greater pitch register than statements.

The present study investigates the impact of those three factors on the perception of sentence modality.

2.1 Methodology

2.1.1 Speech material

We used stimuli with a resynthesized F0 contour in order to control for pitch register and peak placement. As base stimuli we used four short phrases /inːa baba/, /inːa bibi/, /inːa dimaː/, and /inːa ʁila/ '*3ms-say* ('*father*, *turkey*, *always*, *now*')' produced by a trained native speaker of Tashlhiyt. For each phrase, the speaker produced two contours corresponding to two different tonal associations of the high tone as displayed in Figure 2, resulting in two sets of stimuli. One set (PU) contained a rise to an F0 peak on the prefinal syllable (Figure 2 bottom) and the other set (F) contained the rise and peak on the final syllable (Figure 2 top). As mentioned above, tonal association was accompanied by segmental prominence, i.e., the syllable the peak was co-occurring with was longer and louder (cf. Figure 1-2).

Both sets were resynthesized using PSOLA in Praat [5]. F0 was manipulated resulting in two different pitch register conditions: The *low* register condition started with a baseline of 130 Hz the *high* register condition started 4 semitones higher (~164 Hz).

Generally, F0 was manipulated to start rising at the offset of /inːa/ towards two different F0 maximum locations for each set: In the *early* peak condition F0 reached its maximum at 1/3 of the vowel (penult in set PU and final in set F), in the late peak condition F0 reached its maximum at 2/3 of the vowel. The maximum F0 value was 4 semitones higher than the baseline (~164 Hz and ~206 Hz respectively). After reaching its maximum, F0 fell towards the baseline located at the end of the target word. These manipulations resulted in 32 stimuli (4 target words * 2 tonal associations (penultima vs. final) * 2 peak alignments (early vs. late) * 2 pitch registers (low vs. high)) (cf. Figure 3).

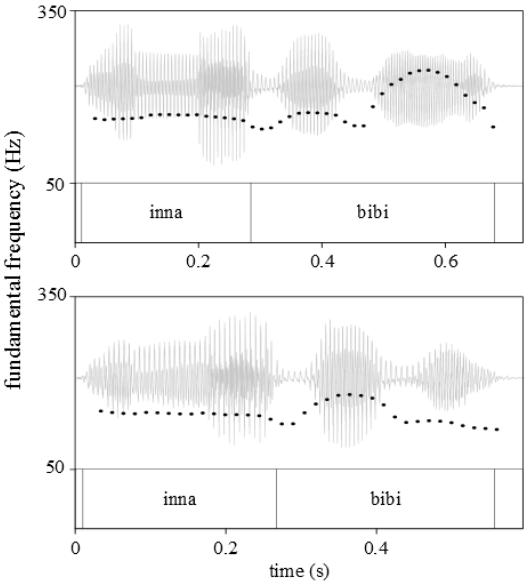


Figure 2: *Waveform and F0 contours of base stimulus productions /inːa bibi/ (3ms-say turkey) 'He said 'turkey'' as an echo question with a peak on the final syllable (Top) and as a statement with a peak on the penultimate syllable (Bottom).*

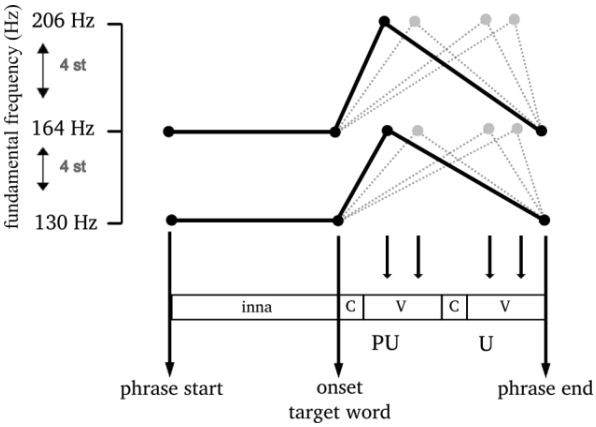


Figure 3: *Schematized manipulation condition displaying the differences of pitch register, tonal association and peak alignment.*

2.1.2 Participants and procedure

Eight native speakers of Tashlhiyt Berber (average age = 21; 4 men; 4 women) participated in our experiment. All live in Agadir, Morocco, and speak at least Moroccan Arabic and to a lesser extent French. Participants were seated in front of a computer screen in a quiet room at the Ibn Zohr University in Agadir. Participants were told that they were going to listen to a robot which does well in speaking Tashlhiyt, however, struggles with producing the difference between statements and questions.

The experiment was controlled using Superlab [6]. At the beginning of each trial, a fixation stimulus consisting of a '+'

was presented in the centre of the screen for 1500 ms. Following this, the stimulus was presented auditorily. Simultaneously two sentences appeared on the right and left side of the screen. Participants had to press a left or right button on the computer keyboard. On one side the statement was displayed in Latin script in blue (e.g. inna baba !), on the other side the question was displayed in red (e.g. inna baba ?). The position of question vs. statement was kept constant within participants, but was counterbalanced across participants. After response delivery, a blank screen appeared for 500 ms.

2.1.3 Analyses

All data were analyzed with generalized linear mixed models, using R [7] and the package *lme4* [8]. To analyze responses categorically, we used mixed logistic regression models with "RATING" (question or statement) as the dependent measure. As fixed effects we included TONAL ASSOCIATION (PU vs. F), PEAK ALIGNMENT (early vs. late), PITCH REGISTER (low vs. high), in addition to TARGET WORD and mean-centered REPETITION. To analyze reaction times (RTs, measured from the offset of the audio stimulus), we used models with Gaussian error distribution with RTs as dependent variable. As fixed effects we included the two-way interactions of RATING and TONAL ASSOCIATION, RATING and PEAK ALIGNMENT, and RATING and PITCH REGISTER, in addition to TARGET WORD and mean-centered REPETITION. For both analyses, we included a term for random intercepts for speakers as well as correlated random slopes for the fixed effects TONAL ASSOCIATION, PEAK ALIGNMENT and PITCH REGISTER (and RATING). For both dependent variables, we tested whether the inclusion of any of the fixed effects did improve the models prediction significantly via likelihood ratio tests.

3. Results and Discussion

Rating: Figure 4 depicts the rating results according to the factors TONAL ASSOCIATION, PEAK ALIGNMENT and PITCH REGISTER. Overall, participants rated the stimuli to correspond to questions in 43% of the cases. There was a significant effect of TONAL ASSOCIATION ($\chi^2(1)=8.66$, $p=0.003$) such that items with the peak on the final syllable were significantly more often rated as a question than statements (61% vs. 25%). There was a significant effect of PITCH REGISTER ($\chi^2(1)=9.14$, $p=0.003$), as well, such that items with a high pitch register were significantly more often rated as a question (58% vs. 28%). PEAK ALIGNMENT did not have a significant effect on responses. Early peaks were rated to correspond to questions comparably as often as late peaks (44% vs. 41%) ($\chi^2(1)= 1.18$, $p=0.28$). As can be seen in Figure 4, there was no apparent interaction of TONAL ASSOCIATION and PITCH REGISTER. Those effects rather add up, with a final peak in a high register being the most preferred question type, and a prefinal peak in a low register being the least preferred question type. However, there appears to be no clear cut. Even the least preferred intonational pattern for questions (low register and peak on PU) shows a considerable amount of question ratings (14%).

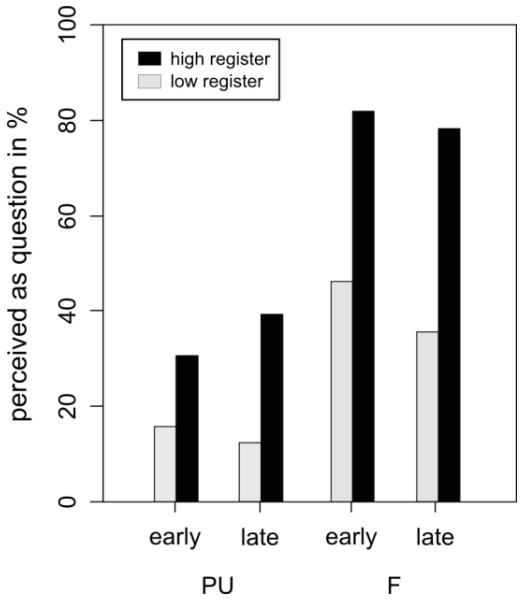


Figure 4: *Ratings as a function of tonal association (PU vs. F), peak alignment (early vs. late) and pitch register (low vs. high).*

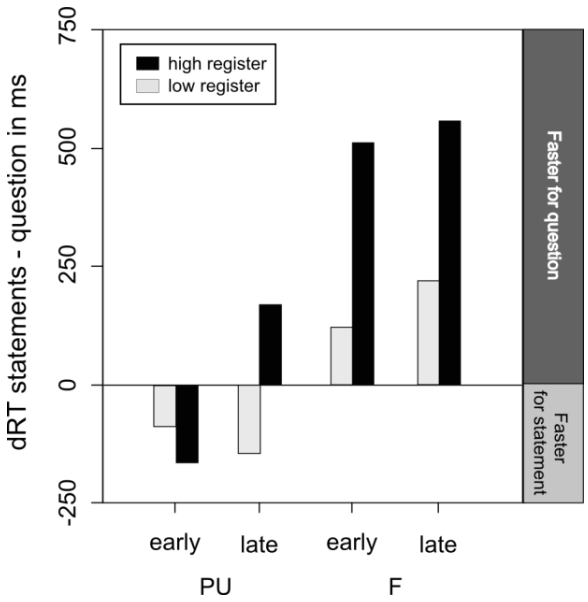


Figure 5: *Differences in response latency between statement and question as a function of tonal association (PU vs. F), peak alignment (early vs. late) and pitch register (low vs. high). Positive values indicate faster responses for questions; negative values indicate faster responses for statements. dRT = 0 indicates no RT difference between statement and question rating.*

RTs: Figure 5 depicts the RT results according to the factors TONAL ASSOCIATION, PEAK ALIGNMENT and PITCH REGISTER. There was a significant interaction of RATING and TONAL ASSOCIATION ($\chi^2(1)=13.49$, $p=0.002$) such that items with the peak on the final syllable were responded to faster when rated as a question (307 ms), while items with the peak on the

penultimate syllable were responded to faster when rated as a statement (123 ms). There was a significant interaction of rating and pitch register ($\chi^2(1)=19.72$, $p<0.0001$), as well, such that there was a greater response latency advantage for questions when stimuli were in a high register (305 ms) as opposed to low register (47 ms). Thus the RT analyses reflect the general finding that both a higher pitch register and a peak on the final syllable are preferred properties of questions. Despite the absence of an effect of PEAK ALIGNMENT on RATING, we find a marginally significant interaction of RATING and PEAK ALIGNMENT for RTs ($\chi^2(1)=2.86$, $p=0.09$), such that there was a greater response latency advantage for questions when stimuli had late peaks (191 ms) as opposed to early peaks (91 ms) (cf. Figure 5).

To sum up, we were able to identify two main factors influencing the perception of the distinction between statements and questions. First, contours with H tones associated to the final syllable are perceived more frequently and faster as questions than H tones associated to the penultimate syllable. However, even contours with a peak on the penultimate syllable appear to be acceptable for questions. Second, contours in a high pitch register are perceived more frequently and faster as questions than contours in a low register. The evidence for peak alignment within each syllable is somewhat weaker. While there was no significant rating asymmetry between early and late peaks, response latencies indicate that late peaks are processed faster when rated as a question (marginally significant). These findings are comparable to results from production, in which [3,4] found some degree of free alternation of tonal association of H tones in polar questions and statements, although questions had a stronger tendency to be realized with a high tone on the final syllable than statements. They further report on later peaks in questions in terms of where the peak is located within the syllable it co-occurs with. Finally, based on impressionistic observations, they report on higher pitch register in questions.

4. General Discussion

The most important contribution of the present study is that Tashlhiyt speakers not only exhibit free alternation of H tone association in production, but are also tolerant with regard to peak placement in perception. For example in questions, there is a probabilistic trend for the peak to be placed on the final syllable, nonetheless, speakers not only produce peaks on the penultimate syllable but also accept those as questions in perception.

We conclude that this variability in locating the tone is free alternation of tonal association. While there have been reports on free alternation of word prominence for Indonesian [9], free alternation in higher prosodic domains are so far unattested. Even though such an alternation of tonal association has not been reported yet, there is some degree of alternation in a number of languages, depending on the properties of syllables at or near the phrase edge. For instance, in Standard Greek, there is secondary association to a lexical stressed syllable if there is one available, otherwise the edge tone is associated with the final syllable [10,11].

We have also observed a preference to place the peak as far to the right as possible. The general preference for late peaks in questions (in terms of both association and alignment), i.e., a rise in pitch, is common across languages [12]. Moreover, several studies showed that speakers use pitch peak alignment to disambiguate sentence types. For example, [13] showed that peak alignment plays a role in disambiguating Hungarian statements and polar questions, whereby an early pitch peak (in the accented vowel) is associated with declaratives and a late peak with interrogatives (see also [14], for Neapolitan Italian; [15], for Swedish; and [16], for Russian). As mentioned above, pitch peaks in Tashlhiyt questions have been observed to exhibit a concomitant steeper rise than in statements. Thus, in production, peak alignment differences could be an artifact of reaching a higher peak. It is well known that a higher peak can perceptually have the same effect as peak delay [14,15,17]. The present study indicate that Tashlhiyt listeners might be sensitive to peak delay irrespective of peak height (although only for RTs).

The remaining variability found in both production and perception, might reflect the redundant nature of peak location in Tashlhiyt. Global pitch cues such as pitch register could already be sufficient for marking sentence modality. The actual peak location phrase finally is thus one of a number of cues and redundant to some degree.

It is important to stress, that while earlier reports on Tashlhiyt intonation [1,3,4] were based on Tashlhiyt speakers living in Paris. The present study recruited subjects that have lived in Morocco for all their lives. We were thus able to confirm the observed patterns in production making it unlikely that they are merely due to interference. Thus, this study not only reveals information about the perception of tonal cues in Tashlhiyt for the first time, it is also a first validation of obtained patterns in production.

5. Acknowledgements

We would like to thank all members of the faculty of Amazigh studies and the dean of the humanities Ahmed Sabir from the Ibn Zohr University in Agadir for their support. We are grateful for all subjects that patiently participated in this experiment. We also thank three anonymous reviewers for their valuable comments. This research was funded by the Volkswagen Stiftung (Project: *Tonal Placement - the Interaction of Qualitative and Quantitative Factors*: *ToPIQQ*).

6. References

[1] Grice, M, Roettger, T. B., Ridouane, R. and Fougeron, C., "Tonal association in Tashlhiyt Berber", Proc. 17th ICPhS, 775–778, 2011.

[2] Gordon, M. and Nafi, L., "The acoustic correlates of stress and pitch accent in Tashlhiyt Berber", Journal of Phonetics, 40: 706–724, 2012.

[3] Roettger, T. B., Ridouane, R. and Grice, M., "Phonetic alignment and phonological association in Tashlhiyt Berber", Proc. 21st ICA, 2013.

[4] Roettger, T. B., Ridouane, R. and Grice, M., "Sonority and syllable weight determine tonal association in Tashlhiyt", Proc. 6th Speech Prosody, 2012.

[5] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer", [Computer program], http://www.praat.org/, 2013.

[6] Abboud, H., "SuperLab", Wheaton, MD: Cedrus, 1991

[7] R Core Team, "R: A language and Environment for Statistical Computing", R foundation for statistical computing, Vienna, URL: http//www.R-project.org.

[8] Bates, D., Maechler, M. and Bolker, B., "lme4: Linear mixed–effects models using S4 classes", R package version 0.999999–0, 2012.

[9] Goedemans, R.W.N. and van Zanten, E.A., "Stress and accent in Indonesian", in V.J. van Heuven and E.A. van Zanten [Eds.] Prosody in Indonesian Languages, LOT: Occasional Series 9: 35-62, 2007.

[10] Arvaniti, A., D. R. Ladd and Mennen, I., "Tonal association and tonal alignment: evidence from Greek polar questions and contrastive statements", Language and Speech, 49: 421–450, 2006
[11] Grice, M., Ladd, D. R. and Arvaniti, A., "On the Place of Phrase Accents in Intonational Phonology", Phonology, 17: 143–185, 2000.
[12] Ultan, R., "Some general characteristics of interrogative systems", in J. Greenberg, [Ed.], Universals of human language, Vol. 4: Syntax. Stanford: Stanford University Press, 211–248, 1978.
[13] Gosy, M. and Terken, J., "Question marking in Hungarian: Timing and height of pitch peaks", Journal of Phonetics, 22: 269-281, 1994.
[14] D'Imperio, M., and House, D., "Perception of questions and statements in Neapolitan Italian", In Proc. Eurospeech, 97: 251-254, 1997.
[15] House, D., "Perceiving question intonation: The role of pre-focal pause and delayed focal peak", Proc. of 15th ICPhS, 755–758, 2003.
[16] Makorova, V., "The Effect of Pitch Peak alignment on sentence type identification in Russian", Language and Speech, 50: 385-422, 2007.
[17] Gussenhoven, C., "The phonology of tone and intonation", CUP: Cambridge, 2002.

The meaning of French "implication" contour in conversation

Cristel Portes [1], *Uwe Reyle* [1, 2]

[1] Aix-Marseille Université, CNRS, LPL, UMR 7309, Aix-en-Provence, France
[2] IMS, University of Stuttgart
cristel.portes@lpl-aix.fr, uwe.reyle@ims.uni-stuttgart.de

Abstract

French intonational contours inventory have a rising-falling tune which presents very interesting semantic properties. It has been called "intonation d'implication" by Delattre [1] suggesting that the contour triggers an implicit meaning, i.e. an implicature in Gricean terms. Besides, the "implication" contour have been claimed to convey various attitudinal meanings from obviousness to exasperation, and also to mark contrastive focus. The aim of the present paper is to give a unified account of these seemingly differing semantic descriptions of the "implication" contour in French, using a dynamic semantic framework, namely Discourse Representation Theory (DRT). We claim that the main semantic component of the "implication" contour is to convey a contradiction (or a contrast). We first present our DRT-theoretical approach, and then apply it to occurrences of the "implication" contour in a corpus of conversational dialogue.

Index Terms: intonation, intonational meaning, "implication" contour, semantics, Discourse Representation Theory, dialogue, conversation, French

1. Introduction

The rising-falling contour called "intonation d'implication" (implication contour) by Delattre [1] is one of the tunes in French that has been attributed a whole range of different kinds of meaning. Delattre himself proposed that the meaning of the contour is to link the meaning of the actual utterance to an implicit content which must be recovered from the context: it may convey various meanings such as obviousness, exasperation or, on the contrary, politeness. Another role attributed to French "implication" is related to a high degree of expressivity or emphasis. For instance Rossi called it "expressème" ([2], [3]) and Di Cristo & Hirst [4] spoke about "emphase contrastive" (contrastive emphasis). This latter proposal refers to another meaning that has been attributed to the contour, i.e. contrast. A related idea expressed in more semantic terms is found in Mertens' and in Ladd's more recent proposals ([5], [6]) where the contour is said to convey speaker commitment. These approaches give very detailed and rich accounts of the phonetics and phonological aspects of the "implication contour" in French. However, its semantic aspects are mostly presented in broad attitudinal terms, which do not account for the dialogical dimension of its meaning.

On the other hand, the semantic literature gives more and more attention to both dialogue and intonation. For instance, working in a semantic framework developed by Ginzburg [7], Beyssade & Marandin [8] proposed that intonational meaning relies crucially on the attribution of attitudes to the addressee. For Gunlogson [9], rising intonation in declarative questions expresses the speaker commitment to a proposition but, at the same time, marks it as contingent on ratification by the addressee. Very recently, Groenendijk & Roelofsen [10] have proposed an "inquisitive" semantic framework where assertions bear inquisitive contents that are inviting responses from other participants. Using this framework, Westera [11] claimed that final rises in English signal that a conversational maxim is violated. And Portes & Reyle [12] followed Krifka's proposal [13] to interpret speech acts by development of spaces of commitments assigned to the discourse participants, in order to explain the meaning of four contours of French inventory.

Convoking both phonological and semantic literature, the present paper aims at showing that French "implication" contour conveys a complex meaning whose different dimensions can be accounted for in a dynamic semantic approach modeling dialogue. After a brief exposal on the phonology of the contour in section 2, section 3 develops a semantic account of its meaning using Discourse Representation Theory (DRT). Then, section 4 verifies the reliability of the proposed unified meaning in a corpus of conversational data, before section 5 concludes.

2. The phonology and phonetics of French "implication" contour

The "implication" contour is not the only rising-falling movement of French intonational inventory. It is sometimes (but must not be) mixed up with a rise-fall the high f0 target of which is localized on the *penultimate* syllable of the accentual phrase (AP: the basic constituent of French phrasing), while it occurs on the *final* (full) syllable for the "implication" contour. The present section gives a brief account of what must be known about the phonology and phonetics of the contour under investigation here.

The clearest account of the phonological contrast and the phonetic implementation of the "implication" rise-fall has been given by Post ([14], [15]). Phonologically, she distinguished this contour from the fall from penultimate contour by attributing two different pitch accents to them. Hence, the former is coded LH*L% with a monotonal H* pitch accent while the latter is coded LH+H*L% with a bitonal H+H* pitch accent.

Phonetically, the "implication" contour LH*L% is said to be implemented with a global difference in temporal alignment compared to the rise LH*H%, the H target of which also occurs on the last syllable of the AP. The alignment of LH*L% is earlier both for its initial L target and for its H* target as shown in Figure 1 and 2 below. This regular phonetic difference has been confirmed by quantitative measures in a large corpus study on naturally occurring data carried out by Portes [16].

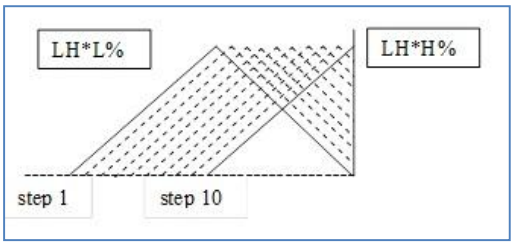


Figure 1: *Design of the ten step alignment continuum between the "implication" contour LH*L% and the rising contour LH*H% used by Post in a categorical perception experiment. The temporal alignment of both the first L and the H* targets are delayed from step 1 to 10.*

Another important aspect of the "implication" contour is that it can occur at the end of an intonational phrase (IP) but also at the end of an intermediate phrase (ip). In this case the final low boundary tone L% may be preceded by a low phrasal tone L- triggered by a narrow focus occurring on a non final constituent, as proposed by Jun & Fougeron [17]. This L- phrasal tone spreads until the end of the IP up to the L%, triggering deaccentuation on the material following the narrow focus constituent, at least when it is marked as background information by the speaker. Figure 2 depicts a corpus example of such a use of the "implication" contour borrowed from Bigi and colleagues [18].

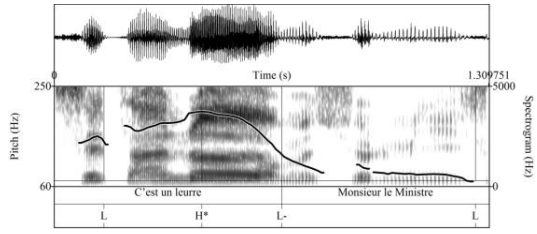


Figure 2: *A corpus example of an ip-final "implication" contour resulting in a LH*L-L%. The background part of the utterance, which is also syntactically right detached, is deaccented through the spreading of the phrasal tone L-. The text is "C'est un leurre, Monsieur le Ministre" (This is an illusion, Minister).*

3. A DRT account of the meaning of the French "implication" contour

We now come back to the meaning of the "implication" contour. In the semantics of French intonation proposed by Beyssade and Marandin [8], this tune belongs to a group of "non falling contours" which are appropriate in what they call a "defective" context (following Stalnaker [19]), namely a context where the assumed beliefs of speaker and addressee are not compatible. More recently, Portes and Beyssade [20] proposed a compositional account of the meaning of French intonation where the "implication" contour is said to convey disagreement. The data in (1) confirm this claim. The contours are depicted with the following codes: F for the simple fall and IRF for the "implication" rise-fall.

(1) a. L1: *Dans cette ville, il n'y a de restaurants que pour les carnivores.* F

In this town, there are restaurants only for carnivores.

b. L2: *Non, il y a un restaurant végétarien* F

No, there is a vegetarian restaurant.

c. L2: *Il y a un restaurant végétarien* IRF

There is a vegetarian restaurant.

Consider the declarative in (1b). Its content q contradicts the assertion p made by L1 in (1a) (with a falling contour). It has been stressed by several authors (cf. Lascarides and Asher [21]) that disagreement should be made explicit, whereas undenied commitments persist in dialogue. For that reason, L2 explicitly rejects p by the preceding "no" in case she wants to use the simple fall F, as in (1b). But (1c) shows that L2 has other means to mark the rejection, and hence to react negatively on the expectation to add p to the common ground (CG), namely by her use of the "implication" contour IRF. This shows that the use of the "implication" contour in (1c) conveys the rejection of p just as the explicit "no" does in (1b).

Conversely, the "implication" contour IRF is not appropriate in (2b) below contrary to the simple fall F. This is because, in this case, there is no contradiction between p (the content of (2a) which is equivalent of that of (1a)) and r, the content of (2b).

(2) a. L1: *Dans cette ville, il n'y a de restaurants que pour les carnivores.* F

In this town, there are restaurants only for carnivores.

b. L2: *Il n'y a pas de restaurant végétarien* F but #IRF

There is no vegetarian restaurant.

Examples (1) and (2) suggest that the "implication" contour expresses a *contradiction*, or more generally a *contrast*. We claim that the implication contour *presupposes*[1] such a contrast and that the different meanings of the contour can be explained on the basis of how this presupposition is resolved in the context of the dialogue. The main ingredients of the explanation are:

(i) Following [22] we assume that the defining criterion for "contrast" is the awareness of a manageable set of alternatives; the set of alternatives is given by the context.

(ii) Implicit Questions under Discussion (QUDs) are present at any stage of the discourse ([22], [23], [7]).

(iii) The interpretation of the implication contour presupposes a contrast, i.e. a set of alternatives to be identified with a contextually given QUD. The particular meaning of the contour then follows from the information that has brought about this QUD at the first place.

To make this more precise, consider (1). L1's statement involves the focus-sensitive operator "il n'y a que" (only). In the framework of an alternative semantics for focus ([24]) such an operator requires a set of alternatives, e.g. "What kinds of restaurants are there in Aix", which can be considered as the QUD to which L1 tries to give an answer by his asserting (1a). The meaning of "il n'y a que" implies that the QUD is fully answered, i.e. all alternatives except p are

[1] A presupposition is a condition associated with a sentence or utterance which must be fulfilled in the context in which the utterance occurs (or the sentence is used) in order that this sentence or utterance succeeds in determining a well-defined proposition.

excluded. If we assume that there are three types of restaurants, non-vegetarians (p), vegetarians (q) and vegans (r), that are relevant in the discussion of L1 and L2, we have before L1's uttering (1a) the QUD = {p,q,r}. With his utterance (1a), L1 proposes a complete answer to L2, i.e. a fully resolved QUD = {p}, as shown in the final line of the following diagram.

utterance	presupposition	QUD
		{p, q, r}
p (= (1.a))		
		{p}

According to our assumption the implication contour on (1c) triggers the presupposition that there must be a contrast, given by a contextually relevant set of alternatives to q. In the context of (1) we may assume that the alternatives are {p,q,r} again. But if we look at the utterance (1c) in isolation, this contrast is underspecified. The only thing we know is, that there are alternatives to q, i.e. the contrast has the form {q}∪C, where C is a non-empty, contextually determined set of alternatives to q. In our example (1), C will be identified with {p,r}. As each assertion requires acknowledgment of the hearer, L2 has to signal disagreement with (1a) if she wants to object. Suppose she agreed and uttered (1c) with the "implication" contour nevertheless. Then the presupposition that there is a QUD = {q}∪C cannot be resolved, because the current QUD consists of the singleton set {p} only, but the presupposition requires a non-singleton set containing q. This explains why uttering (1c) with a fall F (L*L%) is not appropriate. In case L2 doesn't agree with (1a), she may signal her objection with a pure "no", as in (1b). The effect of the rejection is that the original QUD remains active. We take the felicitousness of (1c) with the implication contour as another possibility to express disagreement with (1a). So, once (1a) is rejected by the implication contour, the original QUD is still accessible and the presupposition triggered by the contour can be resolved to {p,q,r}. At the same time, L2 claims to resolve the QUD by his assertion of ¬p[1].

...	...	...
q (= (1.c))	{q}∪C	{p, q, r}
	C = {p,r}	{q}

The dialogue (3), taken from the CID-corpus, has the same structure as (1). The situation and the interpretation of the implication contour is, however, different, because YM's assertion only partially settles the original QUD, i.e. the implicit question of "what there is, that they have at their windows".

(3) YM : *il y a il y a pas de volets quoi*

There are there are no shutters

AG : *ah oui ils y ont des rideaux* IRF *hein*

Ah yes they have curtains havn't they

Let us assume, that there are shutters (p), curtains (q), or nothing at all (r) at the windows, i.e. the implicit question that YM answers with his assertion is QUD = {p,q,r}. Then his utterance only partially resolves this issue, YM only excludes possibility p and leaves the remaining options for further discussion.

utterance	presupposition	QUD
		{p, q, r}
¬p (= YM)		
		{q, r}
q (= AG)	{q}∪C	
	C = {r}	{q}

AG accepts YM's proposal by his *ah oui*. Nevertheless he uses the "implication" contour for his utterance of *ils ont des rideaux*. But this time the presupposition of the contour can be resolved, because the original QUD is not restricted to a singleton, but to {q, r}. AG's assertion is not contradictory to YM's assertion. It tries to settle the original QUD raised but not completely resolved by YM. And this is why – to the extend of AG accepting YM as authority wrt. the original QUD – AG's assertion is understood as a confirmation request.

Our claim is that even cases of confirmation request, information retrieval, exclamation or politeness can be argued to be derived from the general unified meaning we proposed. The politeness effect in (4) is easy to explain on the basis of our contrast-based analysis.

(4) Context: the speaker, while opening a door, says to the hearer:

Après-vous IRF, *cher Monsieur* (deaccented)

After you, Sir

The speaker asserts that he will go after the hearer (p) and, by mean of the contour, contrasts his assertion with the proposition q to go first. Politeness follows as conventional implicature (suggesting that the speaker would never q, i.e. go first). But note, that q need not be uttered (or even thought) by the hearer.

4. A corpus based evaluation

In order to evaluate our semantic proposal of the meaning of the "implication" contour on conversational data, we carried out an analysis of all the occurrences of the contour in a one hour dialogue extracted from the CID corpus [26]. In this corpus, two male speakers, AG and YM, well-knowing each other, were gathered in an anechoic room and fitted out with individual microphones in order to be recorded on separated

[1] In [25], we gave a formal analysis of the meaning of the implication contour that corresponds to the rejection and contradiction case of (1). We assumed that the use of the contour in an utterance of p triggers a presupposition that is more specified than the one we assume in this paper, because the contradiction that explains this particular meaning of the contour is already built in the presupposition itself. This implies that the contour is considered ambiguous between the particular meaning we investigated and the other meanings described in the literature. This paper starts from the assumption that the implication contour is not ambiguous, but underspecified and receives its final meaning by specification in context.

tracks. They were requested to talk freely about unusual events that have happened to them.

Thanks to the hearing and visual inspection of the sound tracks using Praat [27] carried out by the first author of the present paper, 167 occurrences of the "implication" contour were found. 62 were produced by speaker AG and 105 by speaker YM.

We classified the occurrences into 12 different types of situations depending on different uses of the contour. Table 1 below shows the number of occurrences for each type of situation by speaker. The lines in italic correspond to uses for which the meaning is clearly *defective* (i.e. involves a *contradiction* or a *contrast*), what we have described as the crucial dimension of the contour's meaning.

Type of situation	**AG**	**YM**
Contradiction	*3*	*12*
Correction	*7*	
Auto-correction	*5*	*6*
Disagreement, protest	*2*	*9*
contrast	*16*	*33*
paradox	*2*	*3*
"Je sais pas" (I don't know)	*1*	*7*
Adressee's incredulity	*10*	
Confirmation request	4	3
Information retrieval	1	3
Exclamation (emphasis)	3+1	8+9
obviousness	7	12
TOTAL	**62**	**105**

Table 1. *Number of occurrences of the "implication" rise-fall for each type of situation by speaker.*

These contradictory/contrastive items can be said to confirm the meaning proposed in section 2 above, especially by involving a contradiction between a proposition p and its negation ¬p, even if it is sometimes indirectly, i.e. p is not the actual content of the utterance, or by involving a contrast between two incompatible referents or situations. Table 2 gives the proportions of the contradictory/contrastive occurrences versus the other occurrences by speaker. It shows that for both speakers, 70% of the occurrences of the "implication" contour are used in situations and with meanings that confirm our semantic proposal.

Type of situation	**AG**	**YM**
defective	74%	67%
others	26%	33%

Table 2. *Proportions of the defective occurrences of the "implication" rise-fall versus the other occurrences by speaker.*

Here are some examples extracted from the dialogue under study that will make the claim more explicit.

(5) AG : *non* IRF *ça se voyait peut-être je me rappelle plus trop mais je crois pas que ça se voyait* IRF

No perhaps it was visible I don't remember well but I don't think it was visible

Example (5) illustrates a direct contradiction where the speaker AG explicitly negates his addressee's proposition.

(6) AG: *c'est des châtaignes* IRF *ben bien sûr* IRF *ouais il y a que ça* IRF *qui est comestible*

That's chestnut of course yeah only this is good to eat

In (6), the first contour contrasts with anybody's (except the speaker's) potential assumption that it's not des *chataîgnes*, but des *marrons*, the second expresses obviousness (as shown by the words), and the third reinforces the contrast already expressed by the first. Here, the disagreement is not with the addressee, but with a general opinion. The contour on *bien sûr* (of course), shows that even the obviousness use indeed refers to a potential or actual disagreement.

The case of the expression *Je sais pas* (I don't know) is more difficult to explain but also very regular in the corpus. It is used idiomatically with the "implication" contour, once by AG but 7 times by YM, in order to refute in advance the implicit request by the addressee of certified information due to Grice's cooperation principle [28]. Hence, these systematically implicate "I should know" (i.e. ¬p). Example (7) illustrates this case.

(7) *et c'était des glaces y avait je sais pas* IRF quinze litres de glace

And that was ice cream there were I don't' know fifteen liters of ice cream

Even cases of confirmation request, information retrieval and exclamation can be argued to be derived from the general unified meaning we proposed. Confirmation requests involve an issue {p, ¬p} and a commitment of the speaker towards p. Information retrieval examples refer to the presence versus absence of the relevant information. Finally, exclamative or emphatic items rhetorically refer to the incredibility of the information. In all three cases, a potential alternative ¬p is implicated.

5. Conclusion

In this paper, a unified and detailed meaning is proposed for French "implication" contour. This meaning centrally relies on an underspecified presupposition of contrast. The interpretative task of justifying this presupposition in context explains why the contour appears disguised in different kinds of meanings, labeled by other authors as obviousness, exasperation, politeness, emphasis, confirmation request, information retrieval, etc. This unified meaning reliably explains most of the uses of the contour in a spontaneous dialogue extracted from the CID corpus. These results give important support to two important theoretical issues: i) meaning (and especially intonational meaning) contains "inquisitive" components ([8], [10]), and ii) Gussenhoven's "linguistic normalcy" view, that intonational contours have meaning on their own [29].

6. Acknowledgements

This research is funded by a grant from the French National Research Agency (ANR-12-BSH2-0001-01; "*PhonIACog*: *The role of the Initial Accent in prosodic structuring in French – From phonology to speech processing.*

7. References

[1] Delattre, P., “Les dix intonations de base du français”, The French Review : 40, 1966.

[2] Rossi, M., “Vers une théorie de l’intonation”, in Rossi et al. [Eds] L’intonation : de l’acoustique à la sémantique, Paris : Klincksieck, 1981.

[3] Rossi, M., “L’intonation, le système du français : description et modélisation”, Paris : Ophrys, 1999.

[4] Di Cristo, A.and Hirst, D., “Vers une typologie des unites intonatives du français”, 16émes Journées d’Etude sur la parole, Société française d’acoustique (ed.), Avignon, 219-222, 1996.

[5] Mertens, P., “Syntaxe, prosodie et structure informationnelle : une approche prédictive pour l'analyse de l'intonation dans le discours”. Travaux de Linguistique 56(1), Duculot, 87-124, 2008.

[6] Ladd, R. D., “Intonational Phonology”. Second edition: Cambridge University Press, 2008.

[7] Ginzburg, J., “The Interactive Stance: Meaning for Conversation”, Oxford, 2012.

[8] Beyssade, C. and Marandin, J.-M., “French intonation and attitude attribution”, In P. Denis, E. McCready, A. Palmer, and B. Reese (Eds) Proceedings of the 2004 Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface, 2007.

[9] Gunlogson, C.” A question of commitment”. In De Brabanter, P. and Dendale, P. (Eds) Commitment, 101–136, 2008.

[10] Groenendijk, J. and Roelofsen, F., “Inquisitive semantics and pragmatics”. Presented at the Workshop on Language, Communication, and Rational Agency at Stanford, May 2009.

[11] Westera, M., “ ‘Attention, I’m violating a maxim!’ A unifying account of the final rise”, in Fernández, R. and Izard, A. (Eds) Proceedings of the 17thWorkshop on the Semantics and Pragmatics of Dialogue, Amsterdam, decembre 2013.

[12] Portes, C. and Reyle, U., “Intonational meaning triggers expectations”, poster at Discourse Expectations: Theoretical, Experimental and Computational Perspectives, Tübingen,

[13] Krifka, M., “Negated polarity questions as denegations of assertions”, in F. Kiefer and C. Lee (Eds) Contrastiveness and scalar implications. Springer. 2013.

[14] Post, B., “Solving a controversy in the analysis of French rising pitch movements”, Proceedings of ICPhS1999. San Francisco, 965-968, 1999.

[15] Post, B., “Tonal and phrasal structures in French intonation”, published PhD dissertation,The Hague: Holland Academic Graphics, 2000.

[16] Portes, C., “Prosodie et économie du discours : Spécificité phonétique, écologie discursive et portée pragmatique de l'intonation d'implication”. PhD thesis, Aix-Marseille Université, 2004.

[17] Jun, S.-A. and Fougeron, C., “A phonological model of French intonation”, in A. Botinis (Ed.), Intonation: Analysis, modelling and technology. Kluwer, Boston, 2000.

[18] Bigi, B., Portes, C.; Steuckardt, A. and Tellier, M., “Multimodal Annotations and Categorization for Political Debates”, Proceedings of ICMI Workshop on Multimodal Corpora for Machine learning, Alicante, Spain, 2011.

[19] Stalnaker, R., “Assertion”, Pragmatics, Syntax and Semantics, 9, 1978.

[20] Portes, C. and Beyssade, C., “Is intonational meaning compositional”, Verbum, forthcoming.

[21] Lascarides, A. and Asher, N., “Agreement and disputes in dialogue”, Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, Stroudsburg, PA, USA, 29-36, 2008.

[22] Roberts, C., “Information structure in discourse: Towards an integrated formal theory of pragmatics”. Ohio State University Working Papers in Linguistics, 49, 1996.

[23] Büring, D., “On d-trees, beans, and b-accents”, Linguistics & Philosophy, 26:5:511–545, 2003.

[24] Rooth, M., “A theory of focus interpretation”, Natural Language Semantics 1,75-116, 1992.

[25] Reyle, U., Portes, C., “The meaning of French H*L%-contour”. In Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue (DialDam), Amsterdam: December 2013.

[26] [26] Bertrand, R., Blache, P., Espesser, R, Ferré, G., Meunier, C., Priego- Valverde, B., Rauzy, S. “Le CID - corpus of interactional data - annotation et exploitation multimodale de parole conversationnelle », Traitement automatique des langues (TAL), 105-134, 2008.

[27] Boersma, P. and Weenink, D. “Praat: doing phonetics by computer [Computer program]”. Version 5.3.60, retrieved 8 December 2013 from http://www.praat.org/

[28] Grice, H. P., “Logic and conversation”, in Cole P. and Morgan, J., L. (Eds) Syntax and Semantics III: Speech Acts. Academic Press, New York, NY, 41–58, 1975.

[29] Gussenhoven, C. “A semantic analysis of the nuclear tones in English”, in C. Gussenhoven, On the Grammar and Semantics of Sentence Accents, 193-267. Foris, Dordrecht, The Netherlands, 1983.

Combination of variations of pairwise classifiers applied to multiclass ToBI pitch accent recognition

César González-Ferreras, Carlos Vivaracho-Pascual, David Escudero-Mancebo, Valentín Cardeñoso-Payo

Departamento de Informática, Universidad de Valladolid, Spain

{cesargf,cevp,descuder,valen}@infor.uva.es

Abstract

In this paper we present some experiments on multiclass ToBI pitch accent classification. The system is based on the fusion of pairwise classifiers, which are specialized in the distinction of pairs of prosodic labels. Several machine learning techniques, including neural networks, decision trees and support vector machines, are combined in different ways in order to find the best overall combination. Variations of pairwise classifiers are introduced in order to take into account the influence of the samples of the remaining classes during the training of the binary classifiers. The use of these techniques allowed us to improve the results, both the overall classification accuracy and the balance across the different ToBI pitch accent classes.

Index Terms: automatic prosodic labeling, ToBI, classifier combination, pairwise classifiers

1. Introduction

Automatic multiclass pitch accent classification remains a challenging problem in computational prosody. There is a high perceptual similarity between some ToBI labels and some classes are more difficult to identify than others. On the other hand, some prosodic events are more frequent than others, which causes the corpora used in experiments to be clearly imbalanced, and, therefore, the classification performance is negatively affected.

In our previous work we reported a classification strategy based on pairwise classifiers which provided good performance [1]. Pairwise classifiers are specialized in the distinction of the prosodic labels in pairs. Basically, the multiclass classification problem is divided into a set of binary classification subproblems. The distinction of classes in pairs is an easier problem than the distinction between multiple classes and the combination of binary decisions provides improved classification results [2, 3].

In this paper we evaluate two variations of the pairwise strategy: *training with remaining classes* and *correcting classifiers* (to be described in sections 3.3.1 and 3.3.2 respectively). These variations try to avoid the problem that a binary classifier trained to distinguish between two particular classes l and m, might provide unreliable estimations for instances which belong neither to class l nor to class m. We experimented with the fusion of different configurations of the pairwise classifiers based on these variations and on different types of classifiers: neural networks, decision trees and support vector machines. Different types of classifiers appear to behave differently when they attempt to discriminate different classes and their outputs can be complementary.

The use of these machine learning techniques for prosody recognition allowed us to improve the results in multiclass pitch accent classification. As a conclusion, it is difficult to improve at the same time the total classification accuracy and the accuracy rate of each individual class. Thus, we selected two different configurations of the final system: one which improves the total classification accuracy and one which provides more balanced rates among all the prosodic classes.

The structure of the paper is as follows. First, we review the state of the art on automatic prosodic labeling. Then, the classification procedure and the experimental setup are described. Finally, we analyze the results and present some conclusions.

2. State of the art

Automatic detection and classification of ToBI events have been performed using different machine learning techniques: decision trees [1, 4, 5, 6, 7, 8, 9], Markov models [4, 10, 11], maximum entropy models [12], neural networks [1, 7, 8, 13, 14], GMM [13, 15, 16, 17], n-grams [10, 13, 18], Bayesian networks [19], conditional random fields [7, 8] and support vector machines [7, 8, 9, 14]. In most of those works, a combination of these techniques was used.

A common finding of previous work is that accuracy rates are highly dependent on the task: the identification of boundary tones and breaks is easier than the identification of pitch accents. Besides, the results were significantly better in prosodic event detection than in classification. The most efficient classifiers use morpho-syntactic features in conjunction with prosodic acoustic features (F0, intensity and duration) and their temporal evolution. Accuracy rates over 90% are reported in the detection of pitch accents [8]. Nevertheless, accuracy rates in classification are lower, 70.8% in [1], showing a high dependence on the number of classes and speakers, as shown in table 1. Although results can be improved by reducing the number of classes, we decided to keep the original set of classes in this work, since they convey linguistic meaning as defined in the standard [20], which should be preserved.

3. Classification method

In this section we describe the classification procedure used in the experiments, which is an evolution of the system presented in [1]. First we describe the strategy of multiple classifier combination and the base classifiers used in the experiments. Then, two variations of pairwise classification are explained. Finally, we present the experimental setup.

Table 1: Accuracy of pitch accent tone classification for different mappings of the ToBI labels, as reported in the state of the art. All the experiments used the Boston University Radio News Corpus.

Mapping	H*	H*	H*	H*	high	high	high
	L+H*	L+H*	L+H*	L+H*	high	high	high
	!H*	!H*	H*	!H*	downstepped	downstepped	downstepped
	H+!H*	H+!H*	H+!H*	ignored	high	high	high
	L+!H*	L+!H*	L+H*	ignored	downstepped	downstepped	downstepped
	L*	L*	L*	L*	low	low	low
	L*+H	L*+H	L*+H	ignored	low	low	low
	no label	none	ignored	ignored	unaccented	unaccented	unaccented
	#Classes	8	5	4	4	4	4
	Reference	[1]	[21]	[18]	[10]	[22]	[6]
	Level	word	word	word	syllable	syllable	syllable
	#Words/Syllables	27,767	29,578	28,300	14,599	14,599	14,377
	#Speakers	6	6	6	1	1	1
	Accuracy	**70.8%**	**63.99%**	**56.4%**	**80.17%**	**81.3%**	**87.17%**

3.1. Multiple classifier combination

The pairwise coupled approach basically divides a given multiclass classification problem into a number of binary classification subproblems, whose results must be combined to obtain the final classification result [2, 3]. According to this approach, let us refer by $\hat{P}(l|x,\lambda^k_{l,m})$ to an estimation of the probability $P(y = l|x, \quad y = l \vee m)$, where l and m are two different prosodic labels; x is the input of the classifier (in our case, the prosodic features); y is the class label; and $\lambda^k_{l,m}$ is a pairwise classifier of type k that is trained to separate classes l and m (neural network, $k = 1$; decision tree, $k = 2$; support vector machine, $k = 3$).

From these estimators, we build $\hat{P}(l|x,\lambda^k)$, which is obtained with classifiers of type k by:

$$\hat{P}(l|x,\lambda^k) = \prod_{\substack{m=1..C \\ l\neq m}} \hat{P}(l|x,\lambda^k_{l,m}) \quad (1)$$

where C is the number of classes, or prosodic labels.

Then, the results of K different types of classifiers are combined, so that the final estimation of $P(l|x)$, $\hat{P}(l|x)$, is computed as follows:

$$\hat{P}(l|x) = \prod_{k=1..K} \hat{P}(l|x,\lambda^k) \quad (2)$$

For each classifier type, there are as many classifiers as there are combinations of pairs of C classes: $\frac{C\cdot(C-1)}{2}$. Each classifier, $\lambda^k_{l,m}$, provides the posterior probability estimates $\hat{P}(l|x,\lambda^k_{l,m})$ and $\hat{P}(m|x,\lambda^k_{l,m})$.

Since the labeling of a given word depends on the context in which the word has been uttered, we introduce language model dependence. Experiments reported in [4, 13, 23] showed an improvement in results when a model of the sequence of labels was used. A detailed description of the process can be found in [1, 4, 13]. To search for the most likely prosodic label sequence, we applied the Viterbi algorithm [24]. The SRILM toolkit was used to build trigram prosodic language models [25], with Katz backoff for smoothing. The training data was used to build these models.

3.2. Base classifiers

We used three different types of classifiers in this work: decision trees (DT), neural networks (NN) and support vector machines (SVM). The reason for using different types of classifiers is that different classifiers behave differently on the discrimination of prosodic labels [1, 26].

A multilayer perceptron (MLP) was used, trained by means of the standard Error Backpropagation learning algorithm. Nonlinear sigmoid units were used in the hidden and output layers. A single hidden layer was used and a total of 100 training epochs. In the output layer we used as many units as classes, one per each class to classify. The POS feature was transformed into quantitative values by using a binary coding of the 33 values, using 6 bits. Normalization techniques were applied, using Z-Norm normalization across the same speaker.

The Weka toolkit [27] was used to build C4.5 decision trees (J48 in Weka). Different values for the confidence threshold for pruning have been tested, although the best results were obtained with the default value (0.25). The minimum number of instances per leaf was also set to the default value (2). This classifier was trained with qualitative POS features and unnormalized data. To obtain better class probability estimates, we turned off pruning, turned off *collapsing* and calculated class probabilities with the Laplace correction, as described in [28].

We used the Weka machine learning toolkit [27] implementation of the support vector machines. We tested different kernels and selected the polynomial kernel. To obtain probability estimates, logistic regression models were used at the output of the support vector machine. This classifier was trained with qualitative POS features and unnormalized data.

3.3. Variations of pairwise classification

In the canonical pairwise classification scheme, each pairwise classifier is trained to distinguish between two particular classes l and m. Then, only samples of this two classes are used in the learning stage. In the classification stage, each individual pairwise classifier, $\lambda^k_{l,m}$, is coupled with the others in order to get the final output for each test sample x. Given that x can belong to any class, the input of a particular classifier can belong to its target classes (l or m) or not. In this last case, the problem, observed in our work and in the literature [2, 29], is that the

Table 2: *Accuracy of the base classifiers (DT: Decision Tree; NN: Neural Network; SVM: Support Vector Machine; RC: training with Remaining Classes; CC: Correcting Classifiers).*

	DT	DT-RC	DT-CC	NN	NN-RC	NN-CC	SVM	SVM-RC	SVM-CC
H*	61.6%	74.0%	76.0%	64.0%	61.0%	72.2%	44.6%	61.2%	63.5%
L+H*	30.7%	21.2%	19.1%	31.8%	40.4%	34.1%	48.6%	41.9%	36.1%
!H*	35.1%	32.4%	32.7%	36.3%	45.4%	36.3%	44.2%	54.6%	52.0%
H+!H*	17.1%	13.1%	13.8%	18.1%	23.0%	10.1%	36.2%	11.8%	17.7%
L+!H*	7.4%	4.9%	3.9%	14.1%	14.3%	3.0%	29.6%	0.2%	1.3%
L*	18.6%	13.5%	10.3%	16.1%	29.6%	7.2%	45.5%	24.6%	30.2%
L*+H	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
none	86.4%	91.1%	91.6%	88.1%	85.9%	90.8%	82.9%	85.3%	86.5%
Total	66.2%	70.7%	71.2%	68.1%	67.9%	71.2%	63.5%	67.9%	68.6%

classifier might provide an unreliable output. Moreover, this unreliable output could be a high value, which might cause the instance to be incorrectly assigned to class l or to class m.

In order to cope with the described problem, we propose the use of two techniques: training with remaining classes (RC) and correcting classifiers (CC). Therefore, for each classifier type, we have three different classifier configurations: original pairwise classifiers, training with remaining classes and correcting classifiers.

3.3.1. Training with remaining classes

During the training phase, each classifier $\lambda^k_{l,m}$, is trained with examples of three classes: l, m and $\neg lm$ (this last is composed by the training examples of the rest of classes).

In the case of NN, the output layer is composed by two cells, $\{O_1, O_2\}$, assigning each cell at a certain target class, e.g., O_1 to l and O_2 to m. In the standard training method, the desired outputs are fixed to $\{1.0,0.0\}$ for l class samples and $\{0.0,1.0\}$ for m class samples. In the test stage, the input, x, is assigned to the class with the corresponding higher output, i.e, if the higher is O_1 x is assigned to l and if the higher is O_2 x is assigned to m. In the *training with remaining classes* method, the desired outputs in the learning stage are fixed at: $\{1.0,0.0\}$ for the l class training examples, $\{0.0,1.0\}$ for the m class training examples and $\{0.5,0.5\}$ for the $\neg lm$ class training examples. That is, the MLP is trained to provide high outputs when the input belongs only to the l or m classes.

In the case of DT and SVM, a similar method is applied. We extended the binary pairwise classifiers and built classifiers that can distinguish between three classes: l, m and $\neg lm$. Thereby, the probability estimates $\hat{P}(l|x, \lambda^k_{l,m})$ and $\hat{P}(m|x, \lambda^k_{l,m})$ provide high values only when the input belongs to classes l or m.

3.3.2. Correcting classifiers

For each pairwise classifier $\lambda^k_{l,m}$, separating class l from class m, an additional classifier is trained, $\phi^k_{l,m}$, separating classes l and m from all the other classes [29]. This additional classifier generates $\hat{Q}(lm|x, \phi^k_{l,m})$, an estimation that sample x belongs to either class l or class m, and can be included in equation (1), which becomes:

$$\hat{P}(l|x, \lambda^k) = \prod_{\substack{m=1..C \\ l \neq m}} \hat{P}(l|x, \lambda^k_{l,m}) \hat{Q}(lm|x, \phi^k_{l,m}) \quad (3)$$

The drawback of this technique is the cost of training $\frac{C \cdot (C-1)}{2}$ additional classifiers for each classifier type.

3.4. Experimental setup

We used the Boston University Radio News Corpus (BURNC) [30]. The experiments were performed using the word as the reference unit. All utterances in the corpus with ToBI labels from all the speakers were used. Pitch accents considered in this paper (and the number of samples of each) were: *H** (7,587), *L+H** (2,383), *!H** (2,144), *H+!H** (586), *L+!H** (638), *L** (517), *L*+H* (44) and *none* (13,868). We used oversampling in order to reduce the negative impact of imbalanced data on the final result [1, 9, 26]. Ten-fold cross-validation was applied in all the experiments.

We used similar features to the ones used in other experiments [13]. *Frequency features*: within-word F0 range, difference between maximum and average within-word F0, difference between average and minimum within-word F0, difference between within-word F0 average and utterance average F0. *Energy features*: within-word energy range, difference between maximum and average within-word energy, difference between average and minimum within-word energy. *Vowel nucleus duration*: we used the maximum normalized vowel nucleus duration from all of the vowels of the word. *Part of speech*: we used the POS tags that come with the BURNC corpus, which were automatically obtained and were hand-corrected [31].

In order to model the temporal evolution of the pitch contour along the unit of reference, we included additional features: Tilt and Bézier parameters. *Tilt* is probably the most widely applied technique for parameterizing the pitch contours [32]. Tilt has been explicitly used in the state of the art of prosodic event detection [9, 18]. *Bézier stylization* is based on the approximation of the pitch contours with Bézier functions [33]. The minimum square fitting approximation technique is used to represent the shape of the F0 contour along a given reference unit. In this work, we use 4 control points of the spline as parameters.

The use of context features can improve the classification results [1, 9, 21, 22, 34]. We decided to select the features to model the context using the Correlation-based Feature Selection (CFS) algorithm [35]. Without the use of context, for each word, we use 18 features. The CFS algorithm selected 8 features to be used as context features. We used 2 previous words and 2 following words as context [1].

Table 3: *Accuracy of the fusion of classifiers, with and without applying the Viterbi algorithm.*

	without Viterbi	with Viterbi
DT + NN + SVM	70.85%	71.29%
DT + NN + SVM-RC	71.26%	71.49%
DT + NN + SVM-CC	71.46%	71.70%
DT + NN-RC + SVM	70.65%	71.47%
DT + NN-RC + SVM-RC	71.07%	71.61%
DT + NN-RC + SVM-CC	71.12%	71.74%
DT + NN-CC + SVM	71.91%	72.07%
DT + NN-CC + SVM-RC	72.02%	72.15%
DT + NN-CC + SVM-CC	72.08%	72.19%
DT-RC + NN + SVM	72.01%	72.23%
DT-RC + NN + SVM-RC	72.05%	72.22%
DT-RC + NN + SVM-CC	72.24%	72.40%
DT-RC + NN-RC + SVM	72.10%	72.46%
DT-RC + NN-RC + SVM-RC	72.17%	72.37%
DT-RC + NN-RC + SVM-CC	72.20%	72.55%
DT-RC + NN-CC + SVM	72.38%	72.51%
DT-RC + NN-CC + SVM-RC	72.46%	72.58%
DT-RC + NN-CC + SVM-CC	72.56%	72.61%
DT-CC + NN + SVM	72.28%	72.45%
DT-CC + NN + SVM-RC	72.33%	72.57%
DT-CC + NN + SVM-CC	72.43%	72.59%
DT-CC + NN-RC + SVM	72.43%	72.51%
DT-CC + NN-RC + SVM-RC	72.25%	72.41%
DT-CC + NN-RC + SVM-CC	72.37%	72.54%
DT-CC + NN-CC + SVM	72.60%	72.64%
DT-CC + NN-CC + SVM-RC	72.62%	72.62%
DT-CC + NN-CC + SVM-CC	72.62%	72.54%

4. Experimental results

Table 2 shows the classification results of the base classifiers, before the fusion. The total accuracy of the different classifiers ranges from 63.5% for the SVM classifier to 71.2% for the DT-CC and NN-CC classifiers. The strategies RC and CC improve the results of their baseline counterparts: for instance, DT improves from 66.2% to 70.7% and 71.2% respectively.

Another important result in table 2 is that some classifiers are more effective in the identification of a given class than others. This justifies the improvements achieved with the classifier fusion strategy. For example, the SVM classifier is the most efficient in identifying class *L**, with a rate of 45.5%. For this class, DT classifiers only obtain 18.6% at most.

Table 3 shows the results of the fusion of classifiers, with and without applying the Viterbi algorithm. A first conclusion from these results is that the fusion improves the results achieved with the base classifiers. The best global results are achieved when the Viterbi algorithm is used, because it allows to search for the most likely prosodic label sequence, instead of considering the accents in isolation. However, this global improvement is mainly due to the improvement of classes *H** and *none* (the most frequent ones), as shown in table 4.

As we are interested in multiclass classification, higher classification rates in each of the classes are also important. Table 4 compares two alternative combinations with the baseline of our previous work. In the third column, the classifier DT-CC+NN-CC+SVM+Vit provides higher total accuracy rate, but is clearly specialized in the *H** and *none* classes, with accuracies of 78.0% and 91.8% respectively. In the second column, the classifier DT+NN-RC+SVM provides a better balance in accuracy across the different pitch accent classes. This classifier obtains the highest rates of all configurations for classes *L+H**, *!H**, *H+!H**, *L+!H** and *L**. These classes proved to be very difficult to recognize.

Table 4: *Rate of ToBI labels for different combinations of classifiers. We show the combination which provides more balanced results among classes and the combination which provides higher accuracy rate (to select the most balanced configuration we calculated the geometric mean of the classification rate of all classes except class L*+H. DT: Decision Tree; NN: Neural Network; SVM: Support Vector Machine; RC: training with Remaining Classes; CC: Correcting Classifiers; Vit: Viterbi).*

	Previous work [1]		More Balanced		Higher rate	
H*	72.5%	DT+NN+Vit	66.8%	DT+NN-RC+SVM	78.0%	DT-CC+NN-CC+SVM+Vit
L+H*	25.3%		37.3%		25.9%	
!H*	35.2%		46.9%		36.5%	
H+!H*	12.1%		25.3%		10.4%	
L+!H*	6.0%		11.4%		2.2%	
L*	11.4%		32.1%		9.1%	
L*+H	0.0%		0.0%		0.0%	
none	91.0%		88.4%		91.8%	
Total	**70.8%**		**70.7%**		**72.6%**	

Table 4 also shows that with the experiments reported in this paper we have outperformed the results of our previous work [1].

5. Conclusions

We have presented a system for the multiclass classification of ToBI pitch accents, which is based on classification by pairwise coupling and is an extension of our previous work [1]. A classifier for each pair of classes is built and the final label is assigned combining all the pairwise predictions. Several machine learning techniques are used to build the base classifiers: neural networks, decision trees and support vector machines.

We have described two different techniques in order to incorporate the samples of the other classes during the training of the pairwise classifiers: training with remaining classes and correcting classifiers. The use of both techniques provided us with various different configurations of the base classifiers, which seemed to be complementary. The combination of these configurations allowed us to improve our previous results [1]. Some combinations improve the overall classification accuracy: the classifier DT-CC+NN-CC+SVM+Vit improves the total rate from 70.8% to 72.6%. Other combinations provide more balanced accuracies among the different pitch accent classes: the classifier DT+NN-RC+SVM doubles the identification rate (or close to double the rate) of the classes *L**, *L+!H** and *H+!H**.

6. Acknowledgements

This work has been partially supported by Ministerio de Ciencia e Innovacion, Spanish Government (Glissando project FFI2011-29559-C02-01).

7. References

[1] C. González-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso-Payo, "Improving Automatic Classification of Prosodic Events by Pairwise Coupling," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 20, no. 7, pp. 2045–2058, September 2012.

[2] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *The annals of Statistics*, vol. 26, no. 2, pp. 451–471, April 1998.

[3] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, December 2004.

[4] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 469–481, October 1994.

[5] J.-S. Lee, B. Kim, and G. G. Lee, "Automatic corpus-based tone and break-index prediction using k-tobi representation," *ACM Transactions on Asian Language Information Processing*, vol. 1, pp. 207–224, September 2002.

[6] X. Sun, "Pitch accent prediction using ensemble machine learning," in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 16–20.

[7] C. Ni, W. Liu, and B. Xu, "From English pitch accent detection to Mandarin stress detection, where is the difference?" *Computer Speech and Language*, vol. 26, no. 3, pp. 127 – 148, 2012.

[8] C. Ni, W. Liu, and B. Xu, "Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features," in *Proceedings Interspeech*, 2011, pp. 2017 – 2020.

[9] A. Rosenberg, "Automatic Detection and Classification of Prosodic Events," Ph.D. dissertation, University of Columbia, USA, 2009.

[10] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech & Language*, vol. 10, no. 3, pp. 155–185, 1996.

[11] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2005, pp. 269–272.

[12] V. Rangarajan Sridhar, S. Bangalore, and S. Narayanan, "Exploiting Acoustic and Syntactic Features for Automatic Prosody Labeling in a Maximum Entropy Framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 797–811, May 2008.

[13] S. Ananthakrishnan and S. Narayanan, "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, January 2008.

[14] J. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2009, pp. 4565–4568.

[15] Y. Ren, S. Kim, M. Hasegawa-Johnson, and J. Cole, "Speaker-independent automatic detection of pitch accent," in *Proceedings of Speech Prosody*, 2004, pp. 521–524.

[16] K. Chen, M. Hasegawa-Johson, A. Cohen, and J. Cole, "A Maximum likelihood Prosody Recognizer," in *Proceedings of Speech Prosody*, 2004, pp. 509–512.

[17] K. Chen, M. Hasegawa-Johson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2004, pp. 509–512.

[18] S. Ananthakrishnan and S. Narayanan, "Fine-grained pitch accent and boundary tone labeling with parametric F0 features," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2008, pp. 4545–4549.

[19] M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S.-S. Kim, A. Cohen, T. Zhang, J.-Y. Choi, H. Kim, T. Yoon, and S. Chavaria, "Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus," *Speech Communication*, vol. 46, no. 3–4, pp. 418–439, 2005.

[20] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labelling English prosody," in *International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 867–870.

[21] A. Rosenberg, "Classification of Prosodic Events using Quantized Contour Modeling," in *HLT/NAACL*, 2010, pp. 721–724.

[22] G. Levow, "Context in Multi-lingual Tone and Pitch Accent Recognition," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2005, pp. 1809–1812.

[23] A. Rosenberg, "Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2011.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[25] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 901–904.

[26] C. González-Ferreras, C. Vivaracho-Pascual, D. Escudero-Mancebo, and V. Cardeñoso-Payo, "On the automatic ToBI accent type identification from data," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2010, pp. 142–145.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[28] F. Provost and P. Domingos, "Tree induction for probability-based ranking," *Machine Learning*, vol. 52, no. 3, pp. 199–215, 2003.

[29] M. Moreira and E. Mayoraz, "Improved pairwise coupling classification with correcting classifiers," in *European Conference on Machine Learning (ECML)*, ser. Lecture Notes in Computer Science, vol. 1398. Springer, 1998, pp. 160–171.

[30] M. Ostendorf, P. Price, and S. Shattuck, "The Boston University Radio News Corpus," Boston University, Tech. Rep., 1995.

[31] M. Marcus, M. Marcinkiewicz, and B. Santorini, "Building a large annotated corpus of English: The Penn Treebank," *Computational linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[32] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *Journal of Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[33] D. Escudero and V. Cardeñoso Payo, "Applying data mining techniques to corpus based prosodic modeling," *Speech Communication*, vol. 49, no. 3, pp. 213–229, 2007.

[34] A. Rosenberg and J. Hirschberg, "Detecting Pitch Accent at the Word, Syllable and Vowel Level," in *HLT/NAACL*, 2009.

[35] M. A. Hall, "Correlation-based feature subset selection for machine learning," Ph.D. dissertation, University of Waikato, Hamilton, New Zealand, 1998.

Production-Comprehension (A)symmetry: Individual Differences in the Acquisition of Prosodic Focus-marking

Aoju Chen

Utrecht University & Max Planck Institute for Psycholinguistics

aoju.chen@uu.nl

Abstract

Previous work based on different groups of children has shown that four- to five-year-old children are similar to adults in both producing and comprehending the focus-to-accentuation mapping in Dutch, contra the alleged production-precedes-comprehension asymmetry in earlier studies. In the current study, we addressed the question of whether there are individual differences in the production-comprehension (a)symmetricity. To this end, we examined the use of prosody in focus marking in production and the processing of focus-related prosody in online language comprehension in the same group of 4- to 5-year-olds. We have found that the relationship between comprehension and production can be rather diverse at an individual level. This result suggests some degree of independence in learning to use prosody to mark focus in production and learning to process focus-related prosodic information in online language comprehension, and implies influences of other linguistic and non-linguistic factors on the production-comprehension (a)symmetricity.

Index Terms: prosody, focus, comprehension-production asymmetry, individual differences, first language acquisition

1. Introduction

A common pattern in language acquisition is that comprehension precedes production [1, 2]. There are however aspects of language which are characterised by production-precedes-comprehension asymmetries [2]. The acquisition of sentence-level prosody (or: intonation) is a case in point. A most widely discussed phenomenon is the mapping between accentuation and focus. Focus refers to the predication on a topic and typically contains new information to the hearer [3, 4]. In many languages, there is a strong association between focus and accentuation. Specifically, speakers tend to accent the focused constituent and deaccent the unfocused constituents, especially if they are post-focal [5]. Listeners take this into account in online language comprehension such that appropriate focus-to-accentuation mapping speeds up comprehension and inappropriate focus-to-accentuation mapping slows down comprehension [6]. Further, listeners use this mapping to anticipate the upcoming referent in online reference resolution [7]. Adult-like competence in prosodic focus-marking thus entails that children not only can place accentuation to encode focus but also efficiently exploit the focus-to-accentuation mapping in language comprehension. The literature on the acquisition of the focus-to-accentuation mapping over the past two decades has been dominated by the claim that children can use accentuation to mark focus before they can interpret or efficiently use the focus-to-accentuation in comprehension [8].

Recently [9] has pointed out that the alleged asymmetry in the acquisition of the focus-to-accentuation mapping could arise from asymmetries in the test materials used in the production and comprehension studies. More specifically, the test materials in the comprehension studies were usually syntactically more complex and semantically more demanding than the materials used in the production studies. Consequently, what children were supposed to comprehend went beyond the simple focus-to-accentuation mapping in SVO or SV sentences, e.g. the use of accentuation to disambiguate pronouns, accentuation as a cue to contrastive focus in sentences with the focus particle 'only'. [9] examined 4- to 5-year-olds' and adults' production and online comprehension of the focus-to-accentuation mapping in Dutch SVO sentences and found that the 4- to 5-year-olds were similar to the adults in both production and comprehension and the differences between them were of a gradient nature (i.e. more frequent use of accentuation in post-focus constituents and slower reaction times in the children). This is the first evidence suggesting that there is no asymmetry in the acquisition of prosodic focus-marking when the syntactic complexity is controlled for in production and comprehension.

However, prosodic focus-marking is more than the use of accentuation. Accent type also plays a role in focus-marking [10]. For example, in Dutch the preferred accent type to mark focus is H*L in both contrastive and non-contrastive focus [11]. When the same accent type occurs, speakers vary the phonetic realisation of the accent for the purpose of focus marking [12, 13]. Four- to five-year-old Dutch-speaking children are adult-like in choice of accent type in sentence-initial position but not in sentence-final position in SVO sentences [11]. Further, they cannot use phonetic realisation for focus-marking purposes until the age of 8 [12]. Together with the differences in the frequency of accentuation and deaccentuation between children and adults, these differences may have perceptual consequences. It can thus be very useful to examine children's use of prosody in focus-marking through adults' evaluation of children's production.

More importantly, the children in [9]'s production experiment were not the same children as the ones in her comprehension experiment, although they were similar in age and recruited from the same schools. Individual differences in children's intonational skills have been reported for both production and comprehension tasks in earlier work [9, 14]. Can we reduplicate [9]'s results if we examine both production and comprehension in the same group of children? Further, if at the group level, children's production is similar to their comprehension, relative to adults' production and comprehension, does it then mean that production and comprehension go in tandem in every child? In other words, are there individual differences in the production-comprehension symmetricity?

To address the aforementioned questions, we investigated the use of prosody to mark focus in production and the processing of focus-related prosodic information in a single group of Dutch 4- to 5-year-olds. Production data were obtained from the children in a semi-spontaneous setting. Sentence produced by the children were subsequently

evaluated for the appropriateness of the prosody in the corresponding context by trained raters. Comprehension was examined using the same method as in [9]. The focus conditions at issue were narrow focus on the subject NP and narrow focus on the object NP.

2. The production study

The production study consisted of a production experiment and an evaluation experiment. In the production experiment, SVO sentences were elicited in different focus conditions. In the evaluation experiment, trained raters evaluated the appropriateness of the prosody in the participants' sentences.

2.1. The production experiment

2.1.1. Data elicitation

The picture-matching game used in [9] was adapted for the current purpose. The game was played in experimenter-participant dyads. The participant's task was to help the experimenter to find the matching picture for each of her pictures by answering her questions. The conversation between the experimenter and the participant was primarily composed of short question-answer dialogues. The participant had direct access to the information that the experimenter needed and could respond directly to the experimenter's queries. This was achieved by providing the participant with his own set of pictures, each of which depicted a complete event including an agent, a patient and an action. Prior to the game, each participant completed a picture-naming task, in which he named each animal, personage, object and action present in the game and got corrected if he misnamed an entity or used a non-target form. The entities in the pictures were thus referentially given to the participant at the start of the game, rendering the use of a definite article in reference in the game appropriate.

In the game, the experimenter showed the participant one picture a time, drew the participant's attention to the picture, briefly described it (e.g. Look! The girl. There is also the pan. It seems that the girl cooks something.), and then asked the participant a question about the picture (e.g. What does the girl cook?). The participant took a picture from his own set of pictures, which were pre-arranged in an order corresponding to the order of the experimenter's pictures, and tried to identify the information requested by the experimenter. When the participant looked away from his picture, the experimenter repeated her question. The participant then answered the question in an SVO sentence (e.g. The girl cooks the carrot.).

Fifteen question-answer dialogues were embedded in the game to elicit fifteen SVO sentences in five focus conditions: narrow focus in sentence-initial position (NF-i), responding to who-questions and narrow focus in sentence-final position (NF-f), responding to what-questions, in addition to narrow focus in sentence-medial position (NF-m), responding to what-does-X-do-with-Y questions, contrastive focus in sentence-medial position (CF-m), correcting the experimenter's statement about the action, broad focus over the whole sentence (BF), responding to what-happens questions. The target SVO sentences were unique combinations of 3 verbs, 3 object-nouns and 6 subject-nouns. All words were highly familiar words to Dutch 4-year-olds. Each verb and object noun occurred once in each focus condition but never appeared with the same subject noun twice in the game.

2.1.2. Participants

Seventy-five 4- and 5-year-olds (age range: 4;1 to 5;11) participated in the production study. The children were all from monolingual Dutch-speaking families and were recruited from four primary schools in Utrecht Province. Nine adult female native speakers of Dutch took part in the experiment as controls. All participants had normal hearing and speaking ability. Three children did not finish the game. Each session was audio-recorded at a sampling rate of 44.l kHz with 16 bits resolution and video-recorded.

2.1.3. Data annotation

For each participant, the recording was first orthographically annotated in Praat [15]. Second, full-sentence responses were selected as usable responses if they were not plagued by any of the following factors: self-correction, use of pronouns, use of non-target words, detectable hesitation-induced silences, responding to a non-target question, elided responses, overlap with the experimenter's speech, and poor recording quality. Third, the usable full-sentence responses and the corresponding questions or statements were selected and extracted as individual .wav files.

2.2. The evaluation experiment

The usable full-sentence responses and corresponding questions or statements were combined into context-response dialogues with a 1000-ms interval between the question and the response in each dialogue. Subsequently, three intonationally-trained native speakers of Dutch listened to the dialogues and evaluated each response on how well its prosody fitted in the context on a five-point Equal Appearing Interval scale, with 1 standing for 'does not fit' for and 5 standing for 'fits perfectly'. In total, 105 dialogues from 42 children and 25 dialogues from 9 adults in the NF-i condition and 92 dialogues from 36 children and 25 dialogues from 9 adults in the NF-f condition were subjected to evaluation, together with the dialogues from the other focus conditions. The production of 32 of the children was evaluated in both the NF-i and NF-f conditions. To minimise variation in the scores due to comparisons between speakers, the dialogues were presented to the raters per speaker and the experiment was conducted in four 20- to 30-minute sessions. The raters could listen to each dialogue maximally three times before finalising the score. Inter-rater reliability analysis showed that there was a high inter-rater agreement (Cronbach's Alpha = .793; Interclass correlation coefficient = .774)

2.3. Results

Mixed modelling was used to examine the difference in the scores between the children and the adults in each focus condition. The fixed factor was 'age-group' (children vs. adults). The random factors included 'sentence' and 'participant'. The dependent variable was the mean score of the raters. In each analysis on the effect of 'age-group', two models were built, one with only the random factors, and one with both the random factors and the fixed factor. The two models were then compared to each other in an ANOVA test. A statistically significant difference between these two models indicated a main effect of the fixed factor. The p-values reported here were the p-values of the ANOVA tests. Our models showed that the fixed factor 'age-group' had a main effect on the scores in the NF-i condition but not in the NF-f condition. The difference between the children (mean: 2.56) and the adults (mean: 3.65) was thus statistically significant in the NF-f condition ($p < .001$), indicating that the children were not adult-like in their use of prosody in sentences with a focal subject. The differences between the children (mean score: 3.45) and the adults (mean score: 3.22) in the NF-i condition was not statistically significant, indicating that the children could use prosody as appropriately as the adults did in sentences with a focal object.

Taking results from [11, 12] into account, these results suggest that the adult listeners found accenting the post-focus object NPs, acceptable, as found by [16, 17]. But they found the children's failure to exploit phonetic realisation to

distinguish focus from non-focus in the NF-i condition less acceptable.

3. The comprehension study

The reaction-time experiment used in [9] was conducted on the four- to five-year-olds who took part in the production study about two weeks after the production experiment. We only describe the most essential details of the RT experiment here and refer the reader to [9] for further information.

3.1. The RT experiment

The experiment was presented to the children as a game. The children were told that in the game a boy was going to look at a number of pictures with his three pets, a parrot, a chicken, and a duck. The boy wanted to know whether his pets knew the pictures well and which of the pets knew the pictures best. To find this out, the boy showed one picture a time to one of his pets and asked the pet a question about the picture. The children could hear the conversations between the boy and the pets via a headphone set and see each picture on a computer screen together with the boy and his pets. The children were asked to listen to the dialogues between the boy and his pets carefully and judge whether the pets have given correct answers to the boy's questions or not. If they thought a pet gave a correct answer, they should press the green button of the pushbutton box. If they thought that the pet's answer was incorrect for some reason, they should press the red button.

Four lists of 24 experimental dialogues were created from 24 source answer–sentences together with the accompanying pictures. The answer sentences were lexically identical but appeared in different focus conditions (NF-i and NF-f) and prosody conditions in different lists. The prosody of the answer sentences was appropriate in half of the dialogues and inappropriate in the other half of the dialogues. In addition, 20 fillers were included. The answers in the experimental dialogues were all semantically correct but the answers in the fillers contained either a lexical error or a pronunciation error. The questions of the dialogues were recorded by a male speaker of Dutch and the answers by a female speaker of Dutch at a sampling rate of 44.l kHz with 16 bits resolution in the recording studio of the Max Planck Institute for Psycholinguistics. Prosodic analysis on the answer sentences confirmed that the prosody was as intended in both the appropriate and inappropriate prosody conditions. The answer sentences were similar in length across focus conditions and prosody conditions.

The children did the experiment individually in a quite room at their schools. The experiment was conducted by means of the Zep Experimental Control Application [18]. An approximately equal number of children were assigned to each of the list. The exact stimulus list a child got was however randomly decided. Each session lasted about 20 minutes starting with a practice session. In the practice session, the children were familiarised with the task and trained to use the push-button box properly.

The timeline of a trial was as follows: A target picture together with the picture of the boy and one of his pets appeared on the screen. At the same time, the boy said *Kijk* 'look' as an attention getter. 800 ms later, he named an entity in the picture. The 800-ms delay was built in to allow the participants to take a proper look at the picture. 1200 ms after the naming, the boy asked the question. 2200 ms after the end of the question, the pet answered the question. At the end of the answer sentence, a timer with 1 ms accuracy was activated and a picture of the push-button box appeared on the screen.

The RT was recorded from the end of the answer sentence until a button was pressed and the correct–incorrect judgment was automatically recorded. The children were instructed to press the button as quickly as they could, but not before the end of the answer sentence. A timeout device was set at 4 s after the end of a sentence.

Seventy-one out of the 75 children completed the experiment. Two measures were taken from each child, i.e. the 'correct–incorrect' judgement and the RT.

3.2. Statistical analysis and results

The children judged the answers on the experimental trials to be correct in nearly all cases, as expected. The reaction times obtained from the children whose production was evaluated in the NF-i condition and the reaction times obtained from the children whose production was evaluated in the NF-f condition were included for further analysis. The reaction times were log-transformed. Mixed-effect modelling was used to assess the effect of two fixed factors, 'prosody' (appropriate vs. inappropriate) and 'focus condition' (NF-i vs. NF-f) and their interaction on the log-transformed reaction times. Two random factors were included, 'participant' and 'list'. The models were built as described in section 2.2. Our models revealed a main effect the fixed factor 'focus' ($p<.0001$) and a significant interaction between 'focus' and intonation' ($p<.0001$). The main effect of 'focus condition' was such that object focus triggered a longer mean RT than subject focus. This was related to the fact that one had to wait till the end of the sentence in the NF-f condition to form his judgement [9]. The effect of the interaction between 'prosody' and 'focus condition' was such that inappropriate prosody triggered a longer mean RT than appropriate prosody in the NF-f condition but a shorter mean RT in the NF-i condition. This showed that the children were similar to adults in their processing of the focus-prosody interface only when focus was sentence-final.

4. Production and Comprehension

The discussion above centred on the general patterns that have emerged in the children as a group in production on the one hand and in comprehension on the other hand. The results revealed a production-comprehension symmetry whereby the children's production went in tandem with their comprehension. In this section, we address the question of whether there are individual differences in the production-comprehension symmetricity. We focus on the data from the children who were both evaluated for their production and completed the reaction experiment (40 in the NF-i condition and 36 in the NF-f condition; 32 of the children appeared in both conditions; mean age: 5;3).

To quantify individual differences in the production-comprehension interface, two kinds of scores were obtained for each child: 'production scores' that could reflect a child's ability to use prosody in different focus conditions in production; 'comprehension scores' that could reflect how a child processed inappropriate prosody compared to appropriate prosody in different focus conditions. Regarding the production scores, a score was computed for each child in each focus condition by first calculating the mean score of the raters for each response and then averaging the mean scores of all available responses in each focus condition. Regarding the comprehension scores, a score was computed for each child by calculating the ratio between the mean log-transformed reaction time in the inappropriate-prosody condition and that in the appropriate-prosody condition in each focus condition. If the child responded faster in the appropriate-prosody condition than in the inappropriate-prosody condition, the child should have a comprehension score bigger than 1.

We examined the correlation between the production scores and comprehension scores by conducting a Spearman's

correlation coefficient test in each focus condition. The Spearman's correlation coefficient was .113 in the NF-i condition and .034 in the NF-f condition. The significance of the correlation coefficient was .481 and .845 respectively. The results suggested that there was no significant relationship between production and comprehension.

Interestingly, examining the production and comprehension scores in each child, we observed notable individual differences among the children. Assuming that a comprehension score higher than 1 meant successful processing of the focus-prosody interface and a production score higher than 3 indicated relatively accurate use of prosody in focus marking, there appeared to be four sub-groups in the children regardless of the focus condition: (1) poor production, poor comprehension; (2) good production, good comprehension; (3) poor production, good comprehension; (4) good production, poor comprehension, as shown in Figure 1. Such notable individual differences account for the insignificant correlation coefficients in the correlation tests.

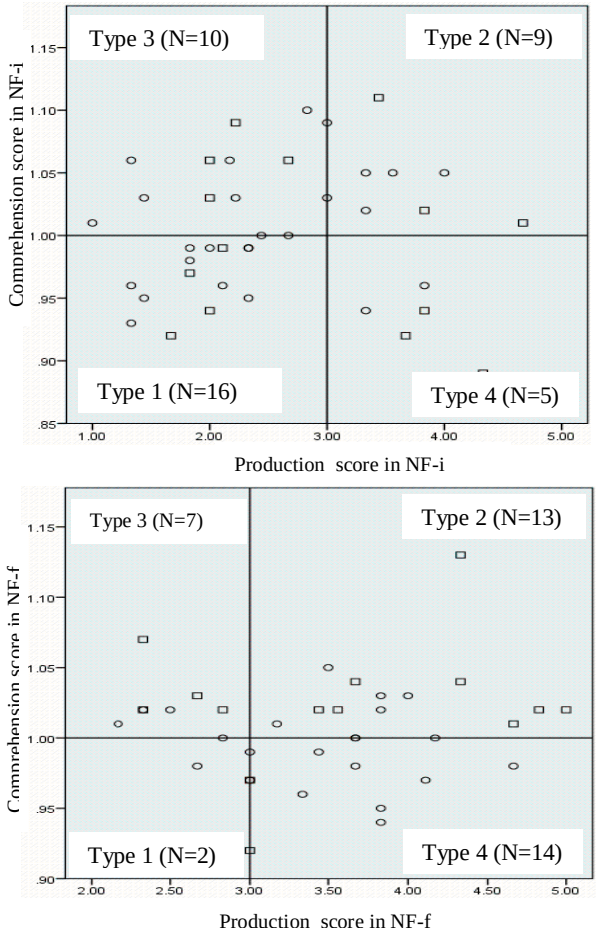


Figure 1. Children's production and comprehension scores (40 in the NF-i condition, 36 in the NF-f condition). Circle: female; square: male

5. Discussion and conclusions

In the current study, we have taken into account the use of multiple prosodic cues in focus marking and examined production and comprehension in a single group of Dutch 4- to 5-year-olds. The children were perceived to be able to mark focus prosodically like adults when focus was in sentence-final position, but not when focus was in sentence-initial position. This asymmetry in the children's production ability was in line with the results based on phonological and phonetic analysis on children's prosody [11, 12] and the tendency to accent given information in adult Dutch [15,16]. By evaluating the overall use of prosody in focus marking, we have obtained a more accurate picture of children's prosodic focus-marking in production. Our results however differed from the results reported in [9] regarding comprehension. The 4- to 5-year-olds in this study differed from their age-matched peers and adults in [9], who responded faster in the appropriate-prosody condition regardless of focus conditions. At first sight, these results may look perplexing. However, these results can be well explained by the children's own production. The children were adult-like in how they processed the focus-prosody interface in the NF-f condition but not in the NF-i condition. Interestingly, their use of prosody was judged to be as appropriate as the adults' use of prosody in the NF-f condition but not so in the NF-i condition (section 2.3). Thus, the children had difficulty with the focus-prosody interface in the NF-i condition in both production and comprehension. Our results thus show at a more fine-grained level that production and comprehension are symmetrical when we treat children as a homogenous group.

The picture is quite different regarding the relationship between production and comprehension in each child. There was not just one type of relationship between production and comprehension present in the data. Zooming in on the production and comprehension scores of each child, we have identified four kinds of relationships between production and comprehension: (1) poor production, poor comprehension; (2) good production, good comprehension; (3) poor production, good production; (4) good production, poor comprehension. These results suggest some degree of independence in learning to use prosody to mark focus in production and learning to process focus-related prosodic information in online language comprehension. Having the representation of the focus-prosody interface does not guarantee success in the actual production at the age of 4 or 5. Conversely, being able to use prosody accurately in focus marking does not entail the ability to integrate focus-related prosodic information into online language comprehension at this age. Possibly, the children who did well in both production and comprehension differed from those who did poor in both or did well in either production or comprehension in other aspects of language development and/or in other areas developmentally. Future research is needed to have a better understanding of the causes for the striking individual differences in the production-comprehension (a)symmetricity.

6. Acknowledgements

This study is supported by a VIDI grant (276-89-001) from the NWO (Netherlands Organisation for Scientific Research). A big thank-you goes to the children and teaching staff from Houten Montessori Primary School, De Ontdekkingsreis Primary School (Doorn), Soest Montessori Primary School, De Wegwijzer Primary School (Soest) for their indispensable cooperation in this research. We thank Paula Cox, Martine Veenendaal, Saskia Verstegen for administering the tests. We also thank Paula Cox for drawing the pictures, and Frank Bijlsma, Alex Manus, Sjef Pieters and Theo Veenker for technical support, Tom Lentz and Mattis van den Bergh, Huub van den Bergh for statistical advice, Ao Chen, Nivja de Jong, Xiaoli Dong, Zenghui Liu, René Kager, Anna Sara Romøren, Anqi Yang and Wim Zonneveld for their feedback.

7. References

[1] Clark, E. (1993). *The Lexicon in acquisition*. Cambridge: CUP.

[2] Hendriks, P., and Koster, C. (2010). Production/comprehension asymmetries in language acquisition. *Lingua*, 120 1887-1897.

[3] Lambrecht, K. (1994). *Information structure and sentence form: Topics, focus, and the representations of discourse referents*. Cambridge: CUP.

[4] Vallduví, E., and Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics*, 34(3): 459-520.

[5] Ladd, D. R. (2006). *Intonational Phonology*. Cambridge: CUP.

[6] Birch, S., Clifton, C.J., (1995). Focus, accent, and argument structure: effects on language comprehension. *Language and Speech*, 38 (4): 365–391.

[7] Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47: 292–314.

[8] Cutler, A., and Swinney, D. A. (1987). Prosody and the development of comprehension. *Journal of Child Language* 14: 145–67.

[9] Chen, A. (2010). Is there really an asymmetry in the acquisition of the focus-to-accentuation mapping. *Lingua*, 120: 1926-1939.

[10] Gussenhoven, C. (2004). *The phonology of tone and intonation*. Cambridge: CUP.

[11] Chen, A. (2011a). Tuning information packaging: intonational realization of topic and focus in child Dutch. *Journal of Child Lang*uage, 38: 1055-1083.

[12] Chen A. (2009). The phonetics of sentence-initial topic and focus in adult and child Dutch. In M. Vigário, S. Frota and M. J. Freitas (eds.), *Phonetics and Phonology: Interactions and interrelations* (pp. 91-106). Amsterdam: Benjamins.

[13] Hanssen, J., Peters, J., and Gussenhoven, C., (2008). Prosodic effects of focus in Dutch declaratives. In: Barbosa, P.A., Madureira, S., Reis, C. (Eds.), Proceedings of the 4th International Conferences on Speech Prosody, Editora RG/CNPq, Campinas, pp. 609–612.

[14] Wells, B., Peppé, S., and Goulandris, N. (2004). Intonation development from five to thirteen. *Journal of Child Language*, 31: 749–78.

[15] Boersma, P., and Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.60, retrieved from http://www.praat.org/

[16] Nooteboom, S. G., and Kruyt, J. G. (1987). Accents, focus distribution, and the perceived distribution of given and new information: An experiment. Journal of Acoustical Society of America, 82: 1512-1524.

[17] Terken, J., and Nooteboom, S, G. (1987). Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and cognitive processes*, 2: 145-163.

[18] Veenker, T.J.G. (2013). The Zep Experiment Control Application (Version *2.0*) [Computer software]. Beexy Behavioral Experiment Software. Available from http://www.beexy.org/zep/

Phonetic variations : Impact of the communicative situation

Sandrine Brognaux[1,2], *Thomas Drugman*[2]

[1]Cental, ICTEAM (Université catholique de Louvain), Belgium
[2]TCTS Lab (Université de Mons), Belgium
sandrine.brognaux@uclouvain.be, thomas.drugman@umons.ac.be

Abstract

While speech synthesis research is now focussing on the generation of various speaking styles or emotions, very few studies have considered the possibility of including phonetic variations according to the communicative situation of the targeted speech (sports commentaries, TV news, etc.). This paper proposes a phonetic analysis of large French corpora to assess the influence exerted by three situational 'traits': read/spontaneous, media/non-media and expressive/non-expressive. It shows that some variations, like elision, tend to be more frequent in spontaneous and non-media speech, conversely to liaisons which appear more often in read and media speech. Interestingly, no phonetic variation draws a clearcut distinction between expressive and non-expressive speech. Finally, a prosodic analysis indicates that the phonetic variations are not directly correlated with the rhythmic features of their corresponding situational 'trait'.
Index Terms: Phonetics, Rhythm, Speaking Style, Speech Synthesis

1. Introduction

Text-To-Speech (TTS) synthesis has reached in the last decades a fairly good level of quality and intelligibility for the generation of neutral read speech. The interest has now shifted to the production of speech corresponding to various speaking styles and emotions. Most studies focussing on this topic have considered modifications at the prosodic and voice quality levels only (see [1, 2]). Surprisingly, potential phonetic modifications of the sentence to synthesize are generally discarded. One rare exception is presented in [3], which modifies the pronunciation of final schwas according to the communicative situation (radio news, conversation, etc.).

This concern is particularly relevant in French, where words are characterized by a high amount of phonetic variants. Schwa elisions and liaisons are the most frequent phonetic modifications. The first consists in the optional pronunciation of a schwa vowel in the middle or at the end of a word (e.g. *petite* pronounced [ptit]). The second relates to latent consonant at the end of a word which can be pronounced when followed by a vowel or a mute h (e.g. *les enfants* pronounced [le͡zɑ̃fɑ̃]). Many linguistic studies have analyzed the modalities of appearance of these phenomena (see [4] and [5] for liaisons, [6] and [7] for epenthetic schwa, [8] and [9] for elisions). They show that the realization of the different variants can be explained by many factors: morpho-syntax [10], speech rate [11, 8, 5], word frequency [4, 5], word probability [12], degree of articulation [13], origin of the speaker [9, 14], age of the speaker [9], etc. Few linguistic analyses, however, have studied the influence exerted on phonetic variations by the communicative situation (TV news, political speech, text reading, etc.), also sometimes referred to as the 'phonogenre' [15, 16]. Yet, the potential interaction between both levels is widely acknowledged [8, 17] and was shown to be very influential for prosody [3, 17, 16]. Only the phonetic differences between read and spontaneous speech have aroused great interest [6, 4, 18].

Most speech synthesizers integrate basic phonetic variations. However, they are trained to produce a phonetic transcription corresponding to standard read speech. For optional variations, the most likely variant is generally produced, independently of the communicative situation. While research is now targeting the generation of expressive [1, 2] and media-related speech (e.g. sports commentaries [19]), the need for a broad study of the influence of these situational 'traits' (as further defined) on phonetics is striking.

Our study proposes an analysis of the influence exerted by the communicative situation on the phonetic realization. Because they cannot be easily ranked on a single scale, the various communicative situations are defined according to three binary 'traits': media/non-media, expressive/non-expressive and read/spontaneous. They will be referred to as 'situational traits' in the remainder of this paper. The main objective of the study is to offer an insightful description of the phonetic features of each 'trait' to outline what should be considered when synthesizing a certain communicative situation.

Our analysis has the advantage of relying on a very large corpus in French of about 300 minutes from 32 speakers and 10 communicative situations (sports commentaries, TV news, political speech, etc.). The study of the phonetic realization is based on a strategy making use of natural language processing (NLP) techniques. In a second stage, rhythm is considered as it was shown to be one of the prosodic correlates of phonetic variations [11, 5]. The potential correlation between phonetic and rhythmic features is then evaluated.

The paper is organized as follows. Section 2 presents the corpus and its annotation. The methodology exploited to carry out our study is detailed in Section 3. The main analysis, both phonetic and rhythmic, of the corpus is described and discussed in Section 4. Finally, Section 5 concludes the paper and discusses further works.

2. Corpus design

Our corpus is an extended version of C-PROM [20] including additional sub-corpora exploited for speech synthesis. A special focus is made on sports commentaries [21] which have also been added to the corpus. The phonetization of the speech files was done automatically and further corrected manually. The entire corpus was then automatically phonetically aligned

with EasyAlign [22] and Train&Align [23].

The corpus consists of 300 minutes from 32 French-speaking speakers (French, Belgian and Swiss) and ten sub-corpora corresponding to different communicative situations (interview, political speech, etc.). Each situation contains 2 to 7 speakers and is defined according to three binary situational 'traits': media, read and expressive. Expressive is here defined as an audible emotive implication of the speaker (e.g. excitation, anger, happiness, etc.), be it acted or not. It should be noted that this 'trait' gathers different kinds of expressivity. Emotion valence, for example, can be positive (e.g. happy) or negative (e.g. sad). This could lead to averaged effects in our analysis, and hinder the interpretation of the role played by the various aspects of expressivity.

A summary of the different sub-corpora is shown in Table 1. The situational 'traits' being continuums, some corpora were not classified (NC) if their nature regarding a dimension was ambiguous. The continuum between read and spontaneous speech, for example, goes through 'prepared', which could be assigned to conferences. Because text to speech corpora were not broadcast as such, but could be used for public announcements, they were not classified for the media 'trait'. For interviews, only the parts of the interviewee were kept. The number of speakers per 'trait' is rather balanced and ranges from 13 to 17 with an average length of about 2 hours of speech per trait.

Table 1: *Distribution of the sub-corpora according to the three studied situational 'traits' (TTS corpora were recorded for speech synthesis purposes).*

Communicative situation	Read	Media	Expressive
Sports commentaries	-	+	+
Conference	NC	NC	-
Political discourse	+	+	NC
Interview	-	+	+
Itinerary explanations	-	-	-
TV news	+	+	NC
Expressive speech TTS	+	NC	+
Neutral speech TTS	+	NC	-
Neutral reading	+	-	-
Narration	-	-	+

3. Methodology

For the phonetic analysis, we developed a specific methodology integrating NLP techniques. For each sub-corpus, the orthographic transcription was exploited to produce its automatic phonetization with the NLP tool ELite [24] designed for TTS purposes. It produced a 'standard' phonetization of the text, corresponding to neutral read speech. This transcription was then automatically aligned with the phonetic transcription really pronounced by the speaker (and which was manually checked).

This alignment relies on a slightly modified version of Levenshtein's edit distance [25]. Several adaptations were made:

- Each phoneme is represented by only one character,
- Some phonemes substitutions are not penalized (ø→œ, e → ɛ, etc.) as they might result from a subjective perception of the annotator,
- Insertions and deletions of silences are not penalized.

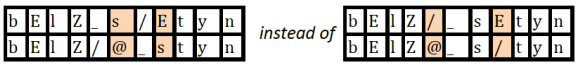
To retrieve the alignment, the matrix obtained by the algorithm is backtracked. Finally, all modifications are stored according to their type (insertion, deletion or substitution).

To avoid potential phonetization errors of the NLP, all sound files containing numbers (written as ciphers) were deleted. Proper names being very frequent in sports commentaries, their deletion would have resulted in a highly reduced sub-corpus. For that reason, we decided to consider only modifications of phonemes not belonging to proper names or to syllables just before and after a proper name.

A first analysis of the alignment highlighted recurrent errors made by the algorithm when two modifications occurred in a small phonetic context. This led to some refinements:

- The deletion and insertion of schwas being more frequent, they were favored and assigned a reduced cost.
- The cost was also reduced for substitution of [i] by [j], which is rather frequent.

This avoided alignments such as:



The main advantage of using NLP-produced phonetization is that it allows for an easy comparison of the pronunciation of the corpus with a so-called 'standard' pronunciation. The latter already considers most mandatory phonetic variations such as liaisons or elisions dictated by the linguistic context. It provides a more precise analysis than a comparison with a phonetized dictionary (see e.g. [4]) while being fully automatic.

Throughout this study, the statistical significance of the results is calculated via unilateral t-tests or Wilcoxon tests depending on the normality of the variable distribution. Correlations are evaluated using Spearman's coefficient.

4. Phonetic and prosodic analysis

4.1. Analysis of the phonetic variations

In this section, we first consider the overall proportion of phonetic variations for each situational 'trait' (4.1.1). We then focus on the analysis of four phonetic variations that were qualitatively assessed to be the most frequent in our corpus: schwa elision (4.1.2), epenthetic schwa (4.1.3), final consonant elisions (4.1.4) and liaisons (4.1.5).

4.1.1. Overall proportion of phonetic changes

Phonetic variations are analyzed by comparing the NLP-produced standard phonetization with the real pronunciation by the speaker. The overall proportion of phonetic changes is computed as the total amount of modifications (deletions, insertions or substitutions) divided by the maximal number of characters, i.e. the number of characters of the longest of both strings.

Table 2 shows significant differences in the amount of phonetic changes for both media and read dimensions (with respectively p=0.043 and p=1.2e-05). This indicates that spontaneous and non-media speech differ rather strongly from what produces a generic NLP. Conversely, read speech corpora exploited for speech synthesis display, on average, only 1.31% of phonetic changes. This finding partly implies that, while the NLP produces suitable phonetizations for neutral read speech, it requires non-negligible modifications to produce non-media

or spontaneous speech. Finally it is worth noting that no significant differences are found along the 'expressive' dimension. This might be due to the heterogeneity of this 'trait', which notably gathers emotions with different valences.

The next sections investigate typical phonetic phenomena for each 'trait'.

4.1.2. Schwa elision

Schwa elision is one of the most intricate phonetic variations in French. It relates to schwas which can be pronounced or not at the middle or the end of a word. Our analysis excludes final schwas which may rather be linked to liaisons and are further investigated in the next subsection. The percentage of elided schwa is here computed as the number of schwa deletions, inside words, divided by the total amount of schwas inside words. Figure 1 shows the significant role played by the distinctions spontaneous/read and media/non-media (with respectively p=0.005 and p=1.7e-04). It shows that more schwas are elided in spontaneous speech, corroborating earlier studies [4, 9, 6]. An interesting finding is that more schwas are also elided in non-media speech. This may be explained by the fact that media speech tends to belong to a higher level of language which has been said to be correlated with lower elision rates [26]. As in [8], we observe rather high inter-speaker variability.

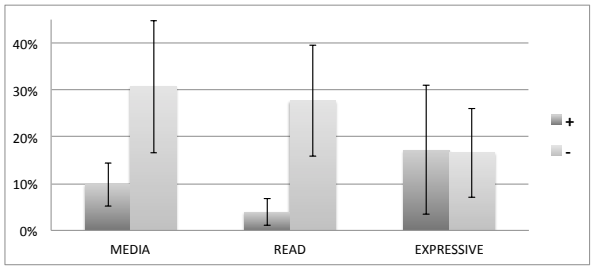


Figure 1: *Percentage of elided schwas in middle word position, together with their 95% confidence intervals.*

4.1.3. Epenthetic schwa

The epenthetic schwa is seen as the insertion of a final schwa on a word ending or not in -e (e.g. *match* pronounced [matʃə]) [6, 27, 28]. Candea [7] shows that its frequency has increased in the last decades and that it can occur independently of the phonetic or rhythmic context. While it was previously seen as a sign of informality, the sociolinguistic aspect is now fading out.

Our analysis focussed only on epenthetic schwas in words with no final -e. Interestingly, this variation is significantly more frequent in media compared to non-media speech, as shown in Figure 2 (p=0.013). This goes in line with [3] which assessed a higher rate of ending schwa pronunciations (all words considered) in radio news and political speech compared to conversational speech. This rate is also significantly higher in spontaneous and expressive speech (p=0.019 and p=0.008), most likely due to their high frequency in sports commentaries. High inter-speaker variability is however observed. While they are often studied on Parisian French, no difference was witnessed in our corpus between French, Belgian and Swiss speakers.

4.1.4. Final consonant elisions

When analyzing the alignment of both predicted and real phonetizations, we observed that the 'il' pronoun (meaning 'he' or 'it') is often pronounced [i], with elision of the final 'l'. This phenomenon occurring quasi-exclusively in front of phonetic consonants, we analyzed its appearance in that specific context (see Table 2). Only sub-corpora containing at least 3 occurrences of the pronoun in that context were kept. Both media and read 'traits' are significant, with higher elision rates in spontaneous and non-media speech (respectively p=6.1529e-05 and p=0.002) which goes in line with results obtained for the elision of schwa in Section 4.1.2.

Another type of consonant elision regards the elision of the final liquid when preceded by an obstruent (e.g. 'peut-être" pronounced [pøtɛt]) [29]. A first qualitative analysis shows that this phonetic variation highly depends on the phonetic context. The liquid is nearly always pronounced when followed by a vowel. Conversely, it may be dropped when followed by a consonant. Our analysis was carried out in this latter phonetic context only, on corpora containing at least 4 occurrences of such phonetic context. Table 2 shows that, here again, spontaneous and non-media corpora display significantly more elisions of the liquid than read and non-expressive speech (with respectively p=1.1e-05 and p=0.04).

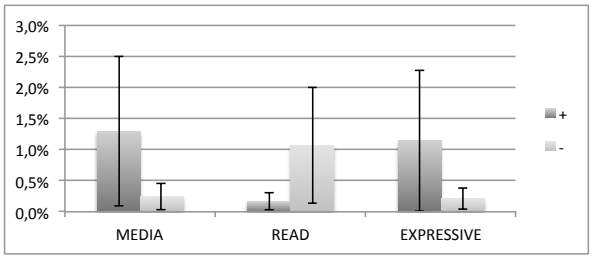


Figure 2: *Percentage of words, not ending in -e, pronounced with a final schwa, together with their 95% confidence intervals.*

4.1.5. Liaisons

We defined potential liaisons contexts as words ending in a French liaison consonant /t, n, z, R, p/ and followed by a vowel, as in [4, 5, 10]. It should be noted that these potential liaisons do not refer to so-called 'optional liaisons', all liaisons being considered, be they mandatory, facultative or prohibited. Table 2 shows that read speech displays a significantly higher rate of liaisons (p=4.6396e-05) which confirms findings in [4, 5, 18]. Interestingly, media speech also shows more liaisons, even if the difference is not significant. This might be due to the fact that media and read speech tend to be more formal, 'sustained' speech being more inclined to high liaison rates [30, 26].

4.2. Prosody: Any correlations between phonetic and rhythmic features?

Rhythm, and speaking rate in particular, has been shown to be correlated with schwa elisions, more elisions appearing in fast speech [11]. This section analyzes various rhythmic features to assess their correspondence with the situational 'traits' and evaluate their correlation with the aforementioned phonetic variations. It focusses on three rhythmic measures: the articulation rate, the mean duration of inter-pausal units (IPU) and the proportion of prominent syllables.

The analysis of the *articulation rate*[1] highlights significant

[1] We focus here on the articulation rate, i.e. the speaking rate (num-

Table 2: *Summary of the phonetic changes for the three situational 'trait' dimensions.*

	All changes		Elision of [l] in 'il"		Elision of liquid in obstruent+liquid		Liaisons	
Situational 'trait'	+	-	+	-	+	-	+	-
Read	1.78%	4.25%	8.33%	76.56%	3.65%	50.34%	59.33%	42.15%
Media	2.93%	4.11%	51.87%	96.67%	18.53%	49.99%	54.52%	44.77%
Expressive	3.57%	2.92%	52.53%	34.72%	33.54%	21.74%	44.53%	50.27%

differences for the read 'trait' only, with lower speaking rates in spontaneous speech (p=0.019). This can be explained by the presence of lengthened syllables, due to hesitations. As in [17], we also observe a significantly lower percentage of articulation for spontaneous speech (p=0.049). Conversely to existing studies (e.g. [11]), however, no significant correlation is found between speaking rate and schwa elision ($|Rho| < 0.09$), or any other phonetic variation. We have shown, on the contrary, that elisions are more frequent in spontaneous speech which displays a lower articulation rate. This difference might be due to the fact that most existing studies focus on one specific task (e.g. text reading) in which only speaking rate is modified to observe the frequency of elisions. In our corpus, too many factors are influencing the speaking rate, e.g. hesitations, communicative situation, speaker idiosyncrasies, etc.

The mean duration of IPU turns out to be significantly longer in spontaneous speech compared to read speech (p=0.04). This seems to indicate that spontaneous speech displays less but longer silences. A possible explanation is that short pauses are rarely silent in spontaneous speech and rather tend to be filled by disfluency markers. Finally, the *percentage of prominent syllables* is assessed with Prosoprom [31], an automatic algorithm for detecting prominent syllables on an acoustic basis. Only the media dimension seems to be influential for that measure, media speech containing more prominent syllables (see Figure 3). This distinction also stands out when looking at initial stresses only, media displaying 19.1% of prominent initial syllables against 16.4% in non-media speech. This confirms findings in [15]. However, the inter-speaker variability is rather high which makes this difference not significant (p=0.07).

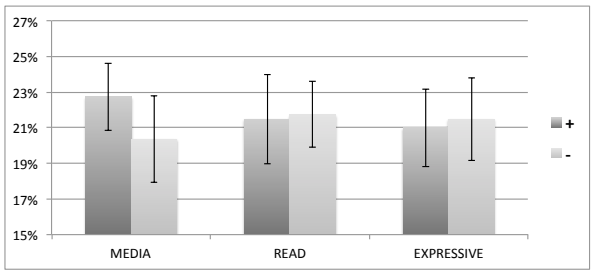


Figure 3: *Percentage of prominent syllables, together with their 95% confidence intervals.*

IPU duration is moderately correlated with schwa elisions and 'l' elisions of the 'il' pronoun (respectively Rho=0.59 and Rho=0.43). This may imply that more elisions are realized when longer sections of speech are uttered without pauses. This, however, does not hold for liquid elisions after obstruents. Interestingly, no correlation can be found between the percentage of prominences and phonetic variations.

ber of syllables per second) with silences excluded, the sub-corpora displaying different silence densities.

On the whole, correlations between rhythmic and phonetic variations are rather weak. This seems to indicate that phonetic variation is more directly dependent on the situational 'trait' itself than on the rhythmic features of the 'trait'. It should be noted that, as for phonetic variation, the expressive 'trait' is not characterized by any specific rhythmic feature.

5. Conclusion and perspectives

While speech synthesis of various speaking styles and emotions is now the focus of much research, very few studies consider potential phonetic variations according to the communicative situation (expressive, spontaneous, etc.). This paper proposed a phonetic analysis of a large corpus in French to assess the influence played by three situational 'traits': read/spontaneous, media/non-media and expressive/non-expressive. It first showed that spontaneous and non-media speech exhibit a significantly higher percentage of phonetic variations compared to standard read speech as produced by a conventional NLP for TTS. Regarding the various phenomena, we showed that spontaneous speech is characterized by a higher elision rate and less liaisons, which confirms results of earlier studies. Media speech usually follows the same phonetic tendencies as read speech, which may be due to the overall higher level of language compared to spontaneous and non-media speech. However, it displays much more epenthetic schwas, which seems to be particularly characteristic of sports commentaries. Interestingly, the expressive 'trait' is not associated with any specific phonetic feature. The diversity of the corpora in that 'trait' (e.g. emotions with different valences) should be further investigated to evaluate the role played by the different aspects of expressivity.

Rhythm was analyzed in a second stage and showed higher speaking rates and shorter inter-pausal units for read speech. Higher proportions of prominences were observed for media speech. Here again, no specific rhythmic feature was associated with the expressive 'trait'. Finally, low levels of correlation were found between the rhythmic parameters and the phonetic variations, except for a moderate correlation between the duration of the inter-pausal unit and some types of elisions. This implies that phonetic variations depend primarily on the situational 'trait' (read/spontaneous and media/non-media) and not on the rhythmic features of that 'trait'.

Further studies will focus on the perceptive analysis of those phonetic changes to assess whether they only constitute possible variants or if they influence the credibility of the message. Required phonetic variations will then be integrated in HMM-based synthesis according to the targeted communicative situation.

6. Acknowledgements

Authors are supported by FNRS. The project is partly funded by the Walloon Region Wist 3 SPORTIC. Authors are grateful to J.-P. Goldman for his insightful advice.

7. References

[1] J. Yamagishi, K. Onishi, T. Musuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IECE Transactions on Information and Systems*, vol. E88-D(3), pp. 502–509, 2005.

[2] R. Tsuzuki, H. Zen, K. Tokuda, T. Kitamura, M. Bulut, and S. Narayanan, "Constructing emotional speech synthesizers with limited speech database," in *ICSLP*, 2004, pp. 1185–1188.

[3] S. Roekhaut, J.-P. Goldman, and A. C. Simon, "A model for varying speaking style in TTS systems," in *Speech Prosody*, 2010.

[4] C. Fougeron, J.-P. Goldman, and U. H. Frauenfelder, "Liaison and schwa deletion in French: an effect of lexical frequency and competition?" in *Interspeech*, 2001, pp. 639–642.

[5] C. Fougeron, J.-P. Goldman, A. Dart, L. Gulat, and C. Jeager, "Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français," in *Actes of TALN*, 2001.

[6] A. Hansen, "The covariation of [schwa] with style in parisian french: an empirical study of 'e caduc' and prepausal [schwa]," in *ESCA Workshop on Phonetics and Phonology of Speaking Styles*, 1991.

[7] M. Candea, "Le e d'appui parisien : statut actuel et progression," in *XXIVe Journées d'Etudes sur la Parole, Université de Nancy II*, 2002, pp. 185–188.

[8] A. Burki, M. Ernestus, C. Gendrot, C. Fougeron, and U. H. Frauenfelder, "What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech," *The journal of the Acoustical Society of America*, vol. 130 (6), pp. 3980–3991, 2011.

[9] F. Hambye, "La prononciation du français contemporain en Belgique. Variations, normes et identités." Ph.D. dissertation, Université catholique de Louvain, Belgique, 2005.

[10] P. Boula de Mareuil, M. Adda-Decker, and V. Gendner, "Liaisons in French: a corpus-based study using morphosyntactic information," in *ICPhS*, 2003.

[11] A. Lacheret-Dujour, "Phonological variations in read speech, reduction phenomena and speaker classes: Do allophonic choices represent speaking style?" in *ESCA Workshop on Phonetics and Phonology of Speaking Styles*, Barcelona (Spain), 1991.

[12] D. Jurafsky, A. Bell, M. Gregory, and W. D. Raymond, "Probabilistic relations between words: Evidence from reduction in lexical production," *Typological studies in language*, vol. 45, pp. 229–254, 2001.

[13] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo and hyperarticulated speech," in *Speech Synthesis Workshop*, 2010.

[14] A. Martinet, *La prononciation du français contemporain*. Librairie Droz, 1971.

[15] J. P. Goldman, A. Auchlin, and A. C. Simon, "Discrimination de styles de parole par analyse prosodique semi-automatique," in *Interface Discours Prosodie (IDP)*, 2009.

[16] T. Prsir, J. P. Goldman, and A. Auchlin, "Variation prosodique situationnelle: étude sur corpus de huit phonogenres en français," in *Interface Discours Prosodie (IDP)*, 2013.

[17] A. C. Simon, A. Auchlin, M. Avanzi, and J. P. Goldman, "Les phonostyles: une description prosodique des styles de parole en français," in *Les voix des Français. En parlant, en écrivant*. Abecassi, M. & G. Ledegen, 2009.

[18] V. Lucci, *Etude phonétique du français contemporain à travers la variation situationnelle (débit, rythme, accent, intonation, e muet, liaisons, phonèmes)*. Publications de l'Université des Langues et Lettres de Grenoble Grenoble, 1983.

[19] B. Picart, S. Brognaux, and T. Drugman, "HMM-based speech synthesis of live sports commentaries: Integration of a two-layer prosody annotation." in *8th ISCA Speech Synthesis Workshop (SSW8)*, 2013.

[20] M. Avanzi, A. C. Simon, J. P. Goldman, and A. Auchlin, "C-PROM. An annotated corpus for French prominence studies." in *Speech Prosody*, 2010.

[21] S. Brognaux, B. Picart, and T. Drugman, "A new prosody annotation protocol for live sports commentaries," in *Interspeech*, 2013.

[22] J.-P. Goldman, "Easyalign: an automatic phonetic alignment tool under Praat," in *Interspeech*, 2011, pp. 3233–3236.

[23] S. Brognaux, S. Roekhaut, T. Drugman, and R. Beaufort, "Train&Align: A new online tool for automatic phonetic alignments," in *IEEE Workshop on Spoken Language Technologies*, 2012. [Online]. Available: http://cental.fltr.ucl.ac.be/train_and_align/

[24] V. Colotte and R. Beaufort, "Linguistic features weighting for a text-to-speech system without prosody model," in *Interspeech*, 2005, pp. 2549–2552.

[25] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals." *Soviet Physics Doklady*, vol. 10 (8), pp. 707–710, 1966.

[26] L. Warnant, *Orthographe et prononciation en français. Les 12000 mots qui ne se prononcent pas comme ils s'écrivent.*, Duculot, Ed., 1996.

[27] A. B. Hansen, "Etude du e caduc - stabilisation en cours et variations lexicales," *Journal of French Language Studies*, vol. 4, pp. 25–54, 1994.

[28] A. B. Hansen and M.-J. Hansen, *Structures linguistiques et interactionnelles dans le français parlé. Actes du colloque international*. Museum Tusculanum Press, 2003, ch. Le schwa prpausal et l'interaction.

[29] J. W. de Reuse, "La phonologie du français de la région de Charleroi (Belgique) et ses rapports avec le wallon," *La linguistique*, vol. 23, pp. 99–115, 1987.

[30] F. Argod-Dutard, *Eléments de phonétique appliquée*, Colin, Ed., 1996.

[31] J.-P. Goldman, M. Avanzi, A. Lacheret-Dujour, A. C. Simon, and A. Auchlin, "A methodology for the automatic detection of perceived prominent syllables in spoken French," in *Interspeech*, 2007, pp. 98–101.

Differences between the acoustic typology of autonomy-supportive and controlling sentences

Netta Weinstein, Konstantina Zougkou and Silke Paulmann

Department of Psychology & Centre for Brain Science, University of Essex, UK
netta@essex.ac.uk, kzougo@essex.ac.uk, paulmann@essex.ac.uk

Abstract

The current study was first to describe distinct patterns of prosody that discriminate motivationally laden speech. To do this we applied self-determination theory, a widely used motivational framework. Participants in the US and UK were asked to read out loud either autonomy-supportive sentences (that support choice and volition) or controlling (pressuring and coercive) sentences. Data analyses were conducted using a conservative hierarchical linear modeling approach to account for nesting of sentences within individuals. Across both countries and controlling for gender, autonomy-supportive sentences were read using lower pitch, less intensity, and a slower speech rate than were controlling sentences. Multiple regression analyses showed links between these patterns of prosody for each participant and his or her current level of motivation, providing additional validity to results. Findings inform both the motivation and prosody literatures and offer a first description of how different kinds of motivational speech may sound.

Index Terms: autonomy-support, controlled motivation, motivational prosody, social prosody, prosodic contour, sentence prosody

1. Introduction

Motivational language is a significant aspect of daily interactions across relationships; it is a fundamental component of people's influence on, and responses to, others. People regularly attempt to shape the behavior of others with motivationally laden language. A child may be told to "go and tidy your room" and while the verbal message plays an important role in shaping behavior, there is reason to believe that prosody is also critical; for example, listening to a calm but firm tone of voice saying "go and tidy your room now" will have greater impact on children's behavior than listening to the same words spoken in a shaky, low tone of voice.

Human communications are often intended to drive and shape others' behaviors, and in these cases motivational elements are imbued in language. The theoretical framework provided by self-determination theory (SDT) argues that two types of qualitatively different messages can drive behavior. Autonomy-supportive statements, such as "you may [do this] if you choose" improve relationships and well-being by supporting perceived choice and volition in the listener. In contrast, controlling statements such as "you must [do this]" are experienced as pressuring or coercive and can undermine well-being [1], but they may also be more effective for achieving desired behaviors in the short-term. The motivational qualities of an interaction have numerous impacts on the personal, relational, and behavioral outcomes of that interaction. Autonomy-supportive versus controlling environments have been shown to shape defiant behaviors [2], encourage interpersonal closeness [3], increase well-being [4], lead to more responsible decision-making [5], and shape successful or unsuccessful performance on important tasks [6], among other outcomes. Furthermore, these motivational communications are relevant in political communications, parent-child relationships, sports and exercise, education, the workplace, and healthcare [7], and as such this fundamental work on motivational communications can be extended across the gamut of human interactions.

Although social psychology has a long tradition of exploring how individuals deliver and understand motivationally laden ideas expressed through specific verbal messages (i.e. words), and research emerging from SDT has examined the words used in both types of messages [8], there is no research to date on how prosody distinguishes these qualities of motivation. Attention to vocal attributes of utterances is crucial in everyday life, particularly in situations when semantic cues are missing (e.g. "time to leave", can be said in either a controlling or autonomy supportive way) or when semantics and prosody mismatch. However, the basic semantic analyses commonly used do not capture the important effect prosody has on motivational language processing. Thus, the current study set out to explore the role prosody plays when communicating motivational language. Past research on another function of prosody, namely emotional prosody, has highlighted that the interplay of multiple acoustic parameters such as frequency variables (e.g., mean, range, contour of pitch), voice qualities (e.g., jitter, shimmer, spectral frequencies), intensity (loudness), and speaking rate, are reliably associated with specific emotional states. For instance, angry statements are often expressed with high intensity, high pitch, and fast speaking rate, while sad statements are conveyed with reduced intensity, low pitch, and slow speaking rate [9]. Thus, in spoken communication, listeners can rely on differences in temporal, pitch, and intensity cues to infer how someone feels; presumably, the same is true when individuals hear motivational messages. It is crucial to note that although basic emotions and motivational styles may correlate, for example, someone who is angry may use a controlling tone, the two constructs are conceptually and operationally distinct – someone who is angry may still use autonomy-supportive language. Accordingly, emotions and motivation should show related but distinct prosody patterns, so that understanding basic emotions does not translate to defining motivational prosody patterns. Specifically, we expected that control and autonomous motivational sentences would be defined by distinct acoustic configuration profiles, as reflected in pitch, temporal, and amplitude differences for the two sentence types.

2. Method

2.1. Participants and procedures

Participants were 100 students recruited from a small university in the North-East USA (n = 51) and from a

comparably sized university in the East of England (n = 49). Of these, 39 American and 31 British females took part. Participants were aged 18 to 32 (M = 22 years).

Participants were randomly assigned to an Autonomy-Supportive or Controlling condition (2-way between-subjects design), which determined the motivational content of the sentences to be read out loud. Instructions for this part of the study asked participants to: "*…read a number of sentences out loud… Please read them in a loud and expressive voice, and pronounce clearly. As if you really mean it…*" Sentences were then presented one at a time at the center of a computer screen at a rate of 12 seconds per sentence. Participants first practiced on a set of 10 motivationally neutral sentences, including "join me at the park" and "why don't we meet tomorrow?" Following this, participants were asked to say aloud the sentences specific to their allocated condition multiple times. The first reading of motivational sentences was designed to increase participant immersion in the particular motivational state, though we have focused on data collected during a second reading.

2.2. Materials

2.2.1. Autonomy supportive and controlling sentences

A series of 12 each autonomy-supportive and controlling sentences were selected based on theoretical considerations and previous research in SDT [10]. Examples of these are "you may do this if you choose" and "you're free to do this" (autonomy) and "you have to do it my way" and "you ought to do it" (control). Conditions were matched on the number of words – both sets of sentences ranged from 3 to 9 words.

To describe the acoustic typology of autonomy and controlling speech, the most commonly studied acoustic parameters, pitch and intensity (mean, minimum, maximum and range), were measured in the current study. Speech rate was also measured that reflected duration of reading out a sentence (in sec.) divided by the number of syllables in that sentence (sentences ranged from 3-10 syllables each).

A rating study was conducted to validate sentences. In particular, we aimed to confirm that autonomous sentences were perceived by others to be more supportive of choice, and that controlling ones were perceived by others to be more pressuring. To this end, an independent sample (n = 33) of British participants listened to recordings of both controlling and autonomy-supportive sentences. Sentences were presented in randomized blocks grouped by condition, and order of sentences was randomized within each block (within-subjects design). Following each sentence, participants were asked to report on the extent each sentence was pressuring (for the controlling condition) using a scale of 1 (*not pressuring at all*) to 5 (*very pressuring*) and on the extent speakers provided choice (for the autonomous condition) with a scale of 1 (*does not support choice*) to 5 (*supports choice very much*). Results showed that participants perceived controlling sentences to be more pressuring (M = 3.80) than supportive sentences (M = 2.16), $F(1, 32) = 133.57$, $p < .001$, and less supportive of choice (M = 2.06) than autonomy supportive sentences (M = 3.62), $F(1, 32) = 31.24$, $p < .001$.

Participants also reported on their current levels of controlled motivation after each block of sentences (blocks presented multiple sentences from one condition only). Findings showed exposure to multiple sentences altered motivational states in a way consistent with the framing of the sentence. After hearing controlling sentences participants reported more controlled motivation (M = 3.14) as compared to after hearing autonomy supportive sentences (M = 2.72), $F(1, 32) = 13.73$, $p = .001$.

2.2.2. State motivation

Participants taking part in the main study, i.e. participants who read aloud the different sentences, reported on their state levels of controlled motivation after a prompt delivered at the end of the study: "how much do you feel this way right now?" Three items assessed motivation, namely feeling "pressured", "coerced", and "choiceless". Each was paired with a seven-point scale ranging from 1(not at all) to 7(very much). Internal reliability was acceptable, $\alpha = .81$.

3. Results

3.1. Analytic strategy

Primary analyses were conducted with hierarchical linear modeling (HLM) [11], [12] because individual sentences (defined at level 1) were nested within speakers (defined at level 2). This method recognizes interdependence of sentences collected from the same participant as well as variation between participants and condition. We first conducted unconditional models to assess intraclass correlation (ICC); this analysis provided an estimate of the variability within-speakers (between sentences) and between-speakers and all parameters showed sufficient variability at both levels for conducting full models. Full models predicted major prosody parameters from gender, country of origin, and condition, as well as the interactions between predictors. The order in which we entered variables was determined by conceptual considerations. Level 2 variables were centered on sample means as recommended by Bryk and Raudenbush [11]; no predictors were specified at level 1. Degrees of freedom for these models are based on the number of participants and observations included in specific analyses, accounting for limited missing data (<5%) due to problems with recordings.

3.2. Preliminary findings

3.2.1. Manipulation check

Preliminary univariate analyses of variances (ANOVAs) predicted state levels of controlled motivation from assignment to condition, gender, country, and their interactions. Findings showed no effect of sex or country on state motivation, $F(1, 87) = .14$, $p = .71$, and $F(1, 87) = 1.01$, $p = .32$. However, those in the Autonomy condition reported less controlled motivation (M = 1.80) at the end of the study as compared to those in the Controlling condition (M = 2.49), $F(1, 87) = 6.26$, $p = .02$. There were no interactions between condition, gender, and country, $F(1, 87) < 2.80$, $ps > .10$. This finding confirms the effectiveness of the sentences used in the two conditions as a mean of motivation manipulation, indicating that assignment to condition did effectively shape speakers' motivational profiles.

3.2.2. Mean levels within countries and genders

Table 1 presents findings for mean scores in each of the two countries and for each gender. Since each of these two predictors may be expected to impact the effects of condition

on prosody indicators, both were controlled for in later analyses.

Table 1. *Means for primary variables of interest in both countries tested and for each gender.*

	M USA	*M* UK	*M* Men	*M* Women
Pitch mean	202.50	186.43	148.39	214.38
Pitch min	143.14	143.15	103.12	160.98
Pitch max	336.76	293.66	261.11	338.02
Pitch range	193.62	150.51	158.66	177.04
Intensity mean	69.35	56.62	63.30	62.83
Intensity min	54.34	32.20	40.26	44.47
Intensity max	82.22	70.09	76.45	76.08
Intensity range	27.88	37.89	36.19	31.61
Speech rate	0.20	0.23	0.21	0.22

3.3. Primary findings

Separate hierarchical linear models predicted mean, minimum, maximum, and range pitch and intensity, as well as speech rate per syllable from gender, country, and condition, and from their interactions.

3.3.1. Pitch measurements

Results showed women spoke in a higher pitch than men, b = 64.88, $t(91) = 8.73$, $p < .001$, though there was no difference across countries, $b = 7.34$, $t(91) = 1.20$, $p = .23$. Controlling for the two covariates, condition marginally affected pitch, b = -10.64, $t(91) = -1.90$, $p = .06$, such that those in the Autonomy condition used lower pitch than those in the Control condition (see Table 2 for a summary of findings and effect sizes). Neither gender, $b = 14.73$, $t(89) = 1.10$, $p = .28$, nor country, b = 14.52, $t(89) = 1.22$, $p = .23$, interacted with condition, indicating motivation was expressed in similar ways across subsamples. In separate models, it was found that autonomy-supportive sentences had marginally lower minimum pitch, b = -6.79, $t(91) = -1.86$, $p = .06$, while results for maximum pitch did not show a significant difference, $b = -22.61$, $t(91)$ = -1.49, $p = .14$; as well, range did not differ across conditions, b = -15.81, $t(91) = -1.08$, $p = .28$.

3.3.2. Intensity measurements

Generally, women spoke with lower intensity than did men, b = -2.01, $t(91) = -2.14$, $p = .04$, as did participants from the UK versus the US, $b = 13.02$, $t(91) = 15.84$, $p < .01$. Controlling for these gender and country differences, participants in the Control condition spoke with a louder tone of voice than those in the Autonomy condition, $b = -2.59$, $t(91) = -3.29$, $p = .002$. Condition did not interact with either gender, $b = -1.52$, $t(89)$ = -0.87, $p = .39$, or country, $b = -0.31$, $t(89) = -0.21$, $p = .83$. It was also found that controlling sentences were spoken with a higher maximum intensity, $b = -2.49$, $t(91) = -2.48$, $p = .02$, though there was no effect of condition on minimum intensity, b = -0.20, $t(91)$ = -0.20, p = .84. As well, autonomous sentences had marginally less range in intensity, $b = -2.29$, $t(91) = -1.98$, $p = .05$, controlling for country, $b = -9.12$, $t(91)$ = -8.28, $p < .001$, and gender, $b = -3.83$, $t(91) = -2.79$, p = .007. Condition did not interact with either country, b = 0.91, $t(91) = 0.47$, $p = .64$, or gender, $b = -2.34$, $t(91) = -0.93$, p = .36, in predicting range in intensity.

3.3.3. Speech rate

Results from full HLM models indicated participants from the US spoke at a faster rate, $b = -0.02$, $t(92) = -2.83$, $p = .006$, though no gender differences were apparent, $b = 0.01$, $t(91)$ = 0.93, $p = .36$. Autonomous sentences were spoken at a slower rate, $b = 0.03$, $t(92) = 4.79$, $p < .001$.

Table 2. *Effects of condition for all acoustic indicators: Results from primary HLM models.*

	b	*t*	*d*
Pitch mean	-10.64	-1.90◆	.40
Pitch min	-6.79	-1.86◆	.39
Pitch max	-22.61	-1.49	.31
Pitch range	-15.81	-1.08	.23
Intensity mean	-2.59	-3.29**	.69
Intensity min	-0.20	-0.20	.04
Intensity max	-2.49	-2.48*	.52
Intensity range	-2.29	-1.98*	.42
Speech rate	0.01	1.96*	.41

Note: effect size was computed using formulation for Cohen's *d*. ◆marginal significance, *$p \leq .05$, **$p < .01$.

3.4. State level outcomes of motivational sentences

To link prosody to state levels of motivation at the end of the study, a multiple linear regression analysis was used regressing the reported state levels of controlled motivation from speakers' tone of voice (prosodic profile) in the controlling condition, accounting for covariates. Mean pitch, intensity, and speech rate (reversed) were standardized and averaged to construct a controlling prosody profile; higher scores reflected more expression of a controlling prosodic tone of voice. Results showed use of controlling tones was linked to a more controlling state orientation at the end of the study, β = .24, $t(91) = 2.51$, $p = .01$; furthermore, in a second step of the analysis, this effect did not interact with covariates, βs < +/-.28, $t(89) < 0.59$, $p > .16$. These findings indicated that those who used a more controlling profile were then more likely to report a motivational style consistent with their tone, providing additional support that these profiles reflected an expression of one's motivational state.

3.5. Supporting analyses with exemplar files

Two independent research assistants blind to hypotheses and previous findings were asked to identify audiofiles that sounded natural, as if spoken and not read, and determined that twenty participants spoke in that ideally natural tone of voice. Sentences from this subset of participants were subjected to analyses of variances (ANOVAs) to attempt to replicate findings from the overall sample presented above. As before, analyses controlled for gender and country; findings for condition are presented below.

3.5.1. Pitch

Results of ANOVAs showed that autonomous sentences (M = 178.7) were spoken with lower pitch, $F(1, 246) = 5.52$, p =

.02, than controlled sentences (M = 188.3). Along with a lower mean, autonomous sentences showed lower minimum pitch, $F(1, 246) = 9.97$, $p = .002$ ($M_{autonomy}$ = 124.2 vs. $M_{control}$ = 135.2), as well as a lower maximum pitch, $F(1, 246) = 4.32$, $p = .04$ ($M_{autonomy}$ = 263.8 vs. $M_{control}$ = 297.6). As was the case for full HLM models, condition did not predict range in pitch, $F(1, 246) = 2.02$, $p = .16$ ($M_{autonomy}$ = 139.7 vs. $M_{control}$ = 162.4).

3.5.2. Intensity

Similarly to pitch, autonomous sentences (M = 62.1) were expressed using lower intensity, $F(1, 246) = 5.69$, $p = .02$, than controlled sentences (M = 63.2). Sentences showed higher minimum intensity in the Autonomy condition, $F(1, 246) = 8.55$, $p = .004$ ($M_{autonomy}$ = 44.1 vs. $M_{control}$ = 41.0), and no differences in maximum intensity, $F(1, 246) = 0.90$, $p = .35$ ($M_{autonomy}$ = 77.5 vs. $M_{control}$ = 78.2). In a separate analysis, autonomous sentences had a lower range of intensity, $F(1, 246) = 8.73$, $p = .003$ ($M_{autonomy}$ = 33.4 vs. $M_{control}$ = 37.2).

3.5.3. Speech rate

Finally, autonomous sentences also showed a slower speech rate than did controlled sentences, $F(1, 246) = 7.72$, $p = .006$ ($M_{autonomy}$ = .016 vs. $M_{control}$ = 0.18).

4. Discussion

The present study set out to explore the acoustic-perceptual underpinnings of motivational prosody using an experimental design. In contrast to studies investigating the acoustic profiles underlying emotional prosody (e.g. [9]), we used a large number of untrained speakers to develop an initial acoustic typology of motivational speech. Findings showed that autonomy-supportive messages such as "you're free to do this" were spoken with a lower mean pitch, lower mean intensity and were read more slowly than controlling sentences. These messages were also expressed using a smaller pitch range than controlling sentences. In contrast, latter messages such as "you ought to do it" were expressed with a higher maximum intensity (i.e. louder) than autonomy supportive sentences. The observed differences in pitch, amplitude, and temporal characteristics of our motivationally laden utterances suggest that speakers adopt specific prosodic speech patterns when communicating autonomy-supportive and controlling motivational sentences.

Everyday motivationally laden messages are used to shape listeners' behaviors. For instance, doctors encourage patients to adhere to preferred health and medical procedures, teachers aim to educate and socialize their students, and parents try to energize their children to engage in valued and avoid harmful behaviors. Even in non-vertical relationships between friends, romantic partners, or housemates, motivational messages are shared daily for important tasks (e.g. financial or lifestyle decisions that affect both partners) and small tasks (e.g. cleaning one's room). Previous work has examined which words are used to discriminate different motivations [8], but next to nothing is known about motivational communication beyond words, in the tone of voice used to express these messages. Our findings support the view that perceptually distinct vocal profiles are used to express motivational speech. Moreover, we show that listeners actually perceive auditorially presented autonomous sentences as being more autonomy-supportive or supportive of choice, whereas sentences from the controlling condition were perceived to be more pressuring. This suggests that expressive tones of voice can be formed and used to drive others to action, at least when prosodic and lexical-semantic cues are used concurrently.

Three major acoustic parameters were measured to identify vocal indicators of motivation: pitch, intensity, and speech rate. Interpretations about the direction of effects have to remain speculative at this point, but, arguably, controlling speech requires more 'effort' than autonomy controlled speech given that one is explicitly trying to shape someone else's behavior. Indeed, we find that the amount of energy used to produce controlling sentences is higher than when producing autonomy-supportive messages. Not only is this extra emphasis missing in the latter condition, autonomy-supportive speech is also characterized by slower speech rate, suggesting that the speaker might be less pressured and more flexible in their attempts to influence others' behaviors. In fact, it is well known that physiological modifications of the systems involved in speech production (i.e. respiratory, phonatory, articulatory) systematically alter a speaker's tone, and research in SDT has linked control to more physiological stress [13].

Our aim is that this research is seen as a starting point to outline different prosody patterns as a function of the speaker's motivational state and intention. Future studies can elaborate on this research in a number of ways. For example, this sample was relatively young and sentences read in a tightly controlled lab setting. This research should be replicated with older participants and those imagining themselves in real-life interactions. Furthermore, we tested a number of indicators of prosody, but future studies could further explore the exact prosody typology that characterizes autonomous and controlled motivation by examining a wider range of acoustic parameters. Finally, future studies will have to confirm acoustic profiles for the two different motivational states, independent of the content used to express a message. One way to do this is to use semantically neutral sentences intoned in either motivationally laden state.

5. Conclusions

The current study was aimed at classifying motivational speech patterns to develop an acoustic typology for autonomous and controlled motivations; ultimately, our goal is to advance a richer understanding of the nature of motivational communication. This is important for a number of reasons. First, prosody research has yet to be generalized to the large body of motivational research in SDT that manipulates motivation – the very manipulations used in the field can be modified and constructed systematically to reproduce expressive tones of voice that affect motivation and subsequent behaviors. Second, this research can be applied in analyzing real life recordings when individuals interact, to examine how motivation functions in daily life. Finally, speech inconsistencies and deceptive motivational influences can be studied for the first time once tone of voice is understood and defined as independent from words. More importantly, this research is fundamental to understanding the basic nature of autonomous and controlled motivations and how these can be communicated in tone of voice.

6. References

[1] Deci, E. L. and Ryan, R. M., "The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior", Psychol. Inq., 11:227-268, 2000.

[2] Vansteenkiste, M., Soens, B., Van Petegem, S. V. and Duriez, B., "Longitudinal associations between adolescent perceived degree and style of parental prohibition and internalization and defiance", Dev. Psychol., in press.

[3] Vansteenkiste, M., Ryan, R. M. and Deci, E.L., "Self-determination theory and the explanatory role of psychological needs in human well-being" in L. Bruni, F. Comim, & M. Pugno [Eds], Capabilities and Happiness, 187-223, Oxford: Oxford University Press, 2008.

[4] Deci, E. L., La Guardia, J. G., Moller, A. C., Scheiner, M. J. and Ryan, R. M., "On the benefits of giving as well as receiving autonomy support: Mutuality in close friendships", Pers. Soc. Psychol. B., 32:313-327, 2006.

[5] Huang, K-L., "The effect of different kind of risk-behaviors on driving and decision-making behavior using SDT", Thesis, Yun-Tech, 2006.

[6] Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M. and Deci, E. L., "Motivating learning, performance, and persistence: The synergistic role of intrinsic goals and autnomy-support", J. Pers. Soc. Psychol., 87:246-260, 2004.

[7] Weinstein, N., "Human motivation and interpersonal relationships: Theory, research, and applications", Dordrecht, NE:Springer, 2013.

[8] Ryan, R. M. and Deci, E. L., "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being", Am. Psychol., 55(1):68-78, 2000.

[9] Banse, R. and Scherer, K.R., "Acoustic profiles in vocal emotion expression", J. Pers. Soc. Psychol., 70(3):614-636, 1996.

[10] Deci, E. L. and Ryan, R.M., "Facilitating optimal motivation and psychological well-being across life's domains", Can. Psychol., 49:14-23, 2008.

[11] Bryk, A.S. and Raudenbush, S.W., "Hierarchical Linear Models in Social and Behavioral Research: Applications and Data Analysis Methods", [First Edition], Newbury Park, CA: Sage Publications, 1992.

[12] Raudenbush, S.W. and Bryk, A.S., "Hierarchical Linear Models" [Second Edition], Thousand Oaks: Sage Publications, 2002.

[13] Weinstein, N. and Ryan, R. M., "A motivational approach to stress response and adaptation", Stress & Health, 1: 4-17, 2012.

Congenital Amusia in linguistic and non-linguistic pitch perception: What behavior and reaction times reveal

Jasmin Pfeifer [1, 2], *Silke Hamann* [1], *Mats Exter* [2]

[1] Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, The Netherlands

[2] Institute for Language and Information, Heinrich-Heine-University, Düsseldorf, Germany

J.Pfeifer@uva.nl, Silke.Hamann@uva.nl, exter@phil.hhu.de

Abstract

Congenital Amusia is a developmental disorder that has a negative influence on pitch perception. While it used to be described as a disorder of musical pitch perception, recent studies indicate that congenital amusics also show deficits in linguistic pitch perception.

This study investigates the perception of linguistic and non-linguistic pitch by ten German amusics and their matched controls. To test the influence of amusia on linguistic pitch perception, the present study parametrically varied pitch differences in steps of one semitone in resynthesized statement-question pairs. In addition, we looked at the influence of stimulus duration, continuity of pitch and direction of pitch change (statement or question). Performance accuracy and reaction times were recorded. Behavioral results show that amusics performed worse than controls over all conditions. The reaction time analysis supports these findings, as amusics were significantly slower across all conditions. Both groups were faster in discriminating statements than questions. Performance accuracy supports these findings, as questions were also harder to discriminate. The present results warrant further investigation of the linguistic factors influencing amusics' perception of intonation.

Index Terms: congenital amusia, pitch, perception disorder

1. Introduction

Congenital amusia is a lifelong disorder defined by difficulties with the perception of tonal differences in music as well as speech. Affected individuals (henceforth: *amusics*) are faced with impairments in the musical domain. Their symptoms can range from an inability to discriminate notes of different pitches, an inability to recognize well-known songs without lyrics or an inability to recognize out of tune singing to an inability to recognize music as such. In the most extreme cases, it causes extreme discomfort and headaches [1-3]. Insufficient exposure to music, a hearing deficiency or brain damage have been excluded as causes [4], while the exact underlying deficit is still unknown. A fine-grained pitch processing deficit has long been assumed as underlying cause [3-6] but has recently been rejected as the sole cause of congenital amusia [7, 8]. Other proposed underlying deficits are a learning disability with respect to statistical learning [9, 10], a working-memory deficit specific to non-verbal sequences [7, 11, 12] or problems with rapid auditory temporal processing [13]. There has been no conclusive evidence for any of these hypotheses and a combination of underlying deficits is now being considered [7].

While it has been proven that congenital amusia negatively affects the musical domain, there has been uncertainty whether it is domain-specific to music or whether it also affects language. It was presumed that language is spared since it employs bigger pitch differences [2, 4, 5] but there is mounting evidence proving that language is also affected [14-17]. Patel et al. ([14]) investigated the pitch perception of English and French speaking amusics with an AX discrimination task using natural statement-question pairs, edited in a way that they differed acoustically in the final region of the intonation contour only. Tonal analogs of these statement-question pairs were also used. They found that 30% of amusics had difficulty discriminating statements from questions and that they performed better for the tonal analogs.

Liu et al. ([15]) also investigated the pitch processing of amusics using an AX discrimination (same-different) task with statement–question pairs, nonsense speech and tonal analogs. As in the study by Patel et al. [14], the stimuli retained the final pitch of naturally produced statements or questions. In this study all amusics performed significantly worse across all three stimuli types. Furthermore, amusics performed better on gliding tones than on natural speech, thereby demonstrating that congenital amusia impairs intonation perception.

The above-mentioned studies thus indicate that the view of congenital amusia as a music-specific disorder has to be reconsidered. Further studies show that amusics have problems distinguishing subtle intonational differences [15], emotional prosody [18], and lexical tones in tonal languages [16, 17, 19, 20]. It is therefore justified to say that congenital amusia is not limited to the musical domain. In light of these findings, further investigations concerning amusics' impairments in language perception are in order.

The present study investigates congenital amusia in speakers of German, a group that has not previously been studied. Our goal is to amass more evidence that congenital amusia does negatively affect language, more specifically, intonation perception, and to investigate possible factors which might influence amusics' perception. This study examines the discrimination of linguistic pitch and two types of tonal analogs by ten amusics and 30 matched controls. In contrast to earlier studies [5, 14, 15], the present study employs a parametric manipulation of small pitch differences from one to seven semitones. Furthermore, it tests the influence of three parameters – the length of stimuli, the continuity of the pitch curve and the direction of pitch change – on amusics' perception. The influence of stimuli length – amusics perform worse for longer stimuli – was shown in studies proposing a memory deficit [11, 12]. The influence of the continuity of the pitch curve was shown to be relevant for amusics in an earlier study [16], and the influence of the direction of pitch change was indicated by a case study [2], the latter showing that the tested amusic only detected rising pitch changes. The present study did not only consider performance accuracy, but also reaction times since amusics have been shown to react more slowly than controls [8, 21, 22].

	Age	Gender	Years of education	MBEA scale	MBEA contour	MBEA interval	MBEA rhythm	MBEA average
Amusic								
Mean	30.7	8 F	16.1	22.0	20.5	20.4	22.6	21.4
SD	10.6	2 M	3.6	2.9	2.1	2.3	3.0	1.4
Control								
Mean	29.3	26 F	16.9	28.4	-	-	24.9	26.6
SD	9.6	4 M	2.8	1.2	-	-	2.9	1.7
t-test								
t	0.394		-0.686	-6.743			-2.109	-8.482
p	0.698		0.497	**< 0.001**			**0.042**	**< 0.001**

Table 1. *Subject characteristics. Descriptive statistics and results of t-tests comparing amusic and control participant characteristics and mean scores of both groups on subtests of the Montreal Battery of Evaluation of Amusia (MBEA). F: female; M: male; SD: standard deviation; t: test statistic of the independent samples t-test; p: probability value; bold face indicates significant results.*

2. Method

2.1. Participants

10 amusic participants and 30 matched control participants were included in this study. None had neurological or psychiatric disorders. All were German native speakers with normal hearing (defined as a mean hearing level of 20 dB or less in both ears), which was assessed before the experiment by pure tone audiometry at 250–8000 Hz. All participants were recruited via advertisement and screened with the *Montreal Battery of Evaluation of Amusia* (MBEA; [23]), the main diagnostic tool to assess amusics, which consists of six subtests testing melodic organization (scale, contour and interval subtests), temporal organization (rhythm and meter subtests) and melodic memory (memory subtest). A mean score of 22 (or lower) out of 30 on the first four subtests was used to diagnose amusia in the present study (cf. [13, 15] and [24] for a detailed discussion). In addition, participants also had to answer a questionnaire about their musical background. The control group was matched for age, handedness, gender and years of education (cf. Table 1). Controls were screened with a shortened version of the MBEA, which contained only the scale and the rhythm subtest, to assess their musical abilities. All participants received a small monetary reimbursement for their participation.

2.2. Stimuli

The experiment was conducted in German. A male native speaker read four statement-question pairs (with a mean fundamental frequency 106.6 Hz) that were embedded in a story. These productions were recorded in a sound-attenuated booth with a Sennheiser ME 62 microphone and a Sound Devices MixPre microphone preamplifier/mixer onto a Marantz PMD570 solid-state recorder with a sampling frequency of 44.1 kHz. The two sentences in each pair were lexically identical but differed in the final region in the direction of the intonation contour (cf. [14]), i.e. statements had a falling pitch and echo questions a rising one. The sentences were constructed so that they differed in two further phonetic parameters. The first parameter was the continuity of the pitch curve, i.e. half of the sentences consisted only of voiced sounds, resulting in a continuous pitch contour, while the other half contained voiceless obstruents, yielding a discontinuous pitch contour. The second parameter was length: half the sentences were short (sentences 1 to 4 had 3 to 6 syllables with a mean duration of 1.05 s and a SD of 0.08 s) and the other half long (sentences 5 to 8 had 7 to 10 syllables with a mean duration of 1.59 s and a SD of 0.24 s).

Praat [25] was utilized to extract, stylize and simplify the pitch contour of each target sentence (for details see [24]). The resulting simplified pitch contour was then used to replace the original pitch contour, yielding synthesized stimuli. Seven further pitch contours were created for every target sentence by moving the final pitch region of the stimulus either upwards (for the statements) or downwards (for the questions) in one-semitone steps, thereby yielding a set of eight different stimuli per target sentence (cf. Figure 1) and 64 stimuli in total. The question was manipulated downwards towards a statement, but never reached the pitch level of the recorded statement, and the statement was manipulated upwards. The logarithmic semitone scale was used instead of a linear scale (as e.g. in [5]), thereby making the result more comparable to pitch detection thresholds of amusics in the literature [3, 26].

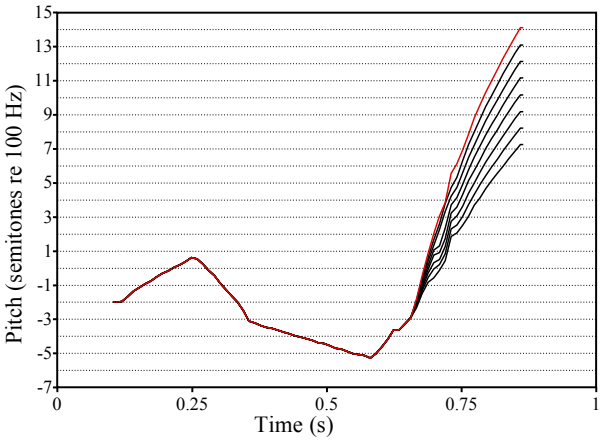


Figure 1. *Pitch contours of stimuli. Original simplified contour in red and seven manipulations in black.*

In addition to the speech stimuli, two types of tonal analogs, consisting of sinusoidal waves and pulse trains respectively, were created, resulting in a total of 192 different stimuli.
The sinusoidal wave analogs (short: sine) were created by converting the synthesized sentences into sinusoidal waves whose frequency exactly followed the sentences' pitch contour (cf. [14, 27]). The pulse train analogs (short: pulses) were created by converting the sine analogs into sequences of pulses, with the distance between pulses inversely proportional to the frequency of the pitch curve (a higher frequency corresponds to a smaller distance). These two types of tonal analogs were chosen since they differ in acoustic complexity. While the sinusoidal waves have a relatively

simple acoustic signal, the pulse trains have a more complex acoustic signal. The three different stimulus types therefore vary in the presence or absence of linguistic material and in acoustic complexity from sine, as the simplest one, to pulses, as the intermediate one, to speech, as the most complex one.
Since an AX discrimination task was used, stimuli had to be paired. Each stimulus with an altered final pitch region (black in Figure 1) was once paired with itself and once with the unaltered version (red in Figure 1), while a separation between the stimuli created from different sentences and between the different stimulus types was maintained (i.e. sine was not mixed with pulse or speech stimuli etc.). This yielded seven 'different' pairs and one 'same' pair per stimulus. For counterbalancing reasons, an equal number of 'same' and 'different' pairs was included in the experiment, i.e. 14 stimuli pairs. In total, 336 experimental stimuli pairs were included, consisting of 8 original sentences (4 statement-question pairs) x 3 different stimulus types (speech stimuli and tonal analogs) x 14 stimuli pairs (7 'same' pairs with unaltered final pitch regions and 7 'different' pairs with altered final pitch regions).
In addition to the 336 experimental trials, nine practice and 12 catch stimuli were created in the same way. Catch trials were 'different' stimuli pairs in which the final pitch region of one stimulus was altered by 24 semitones. These catch trials were included in the experiment to ensure participants paid attention and performed the task correctly. Controls and amusics perceived all catch pairs as different, thus no one had to be excluded on these grounds. Practice trials consisted of the three short, continuous, statement stimuli which were paired either with themselves (resulting in three 'same' stimuli pairs), or with contours that were raised finally by 2.5 and 5.5 semitones (resulting in six 'different" stimuli pairs). These nine practice trials were used in the practice session before the experiment.

2.3. Design and procedure

The 336 stimuli pairs differed in 5 conditions:

- type (voice, sine, pulses),
- length (short, long),
- continuity (continuous, discontinuous),
- direction (question, statement),
- interval (0-7 semitones).

They were used in a same-different discrimination task with a blocked design. The 14 stimuli pairs that shared all conditions were presented within a block in a randomized order. Across blocks, the order was pseudo-randomized so that blocks with more than two conditions in common did not immediately follow each other. The order of blocks was counter-balanced across participants to compensate for fatigue. There were two breaks, one after every eight blocks.

The experimental sessions took place in the phonetics laboratory at the University of Düsseldorf and lasted approximately 60 minutes. Participants were seated in a sound-attenuated booth, and the stimuli were presented over AKG K 601 headphones using Praat on a Windows XP computer. Participants could adjust the volume to a comfortable level. They were asked to listen carefully to each trial and to decide whether the two stimuli were the same or different. They were told to respond as quickly as possible by pressing labeled buttons on the keyboard. Behavioral results and reaction times were recorded with Praat.
Each trial followed the same pattern: A warning signal was followed by one second of silence followed by the stimulus pair with an inter-stimulus interval (ISI) of one second. This ISI was chosen since Williamson et al. [13] pointed out that longer ISIs may interfere with possible pitch memory deficits in amusics, while shorter ISIs might cause problems due to a rapid auditory processing deficit in amusics. The experiment was preceded by a practice session to familiarize participants with the experimental procedure, the different types of stimuli and the semitone intervals. Feedback was provided during the practice session but not during the experiment.

3. Results

Responses were scored as hits when different stimuli pairs were correctly identified as different, and as misses when they were not correctly identified. Conversely, correctly identified same-pairs were scored as correct rejections and as false alarms when they were incorrectly identified.

3.1. Reaction time (RT) analysis

RTs were measured from the offset of the second stimulus of each pair. Outliers, here defined as negative RTs or RTs slower than 3 SD of the group mean, were excluded. In a first step, the mean RTs for controls and amusics over all conditions were compared (cf. Figure 2). The RTs of amusics (M = 1121, SE = 13.2) and controls (M = 946, SE = 4.6) differed significantly ($t(998) = 12.5$, $p < 0.001$) only for hits. This represented a medium-sized effect r = 0.37. RTs are conventionally analyzed for hits only, therefore the following analysis takes only RTs of hits into account.

The next step consists of analyzing the RTs per variation step in semitones. Controls and amusics differed significantly at all semitone steps, except step 1 (Table 2). Next, RTs were submitted to an ANOVA with group (amusic, control) as the between-participant factor and length, continuity, direction and type as within-participant factors. All parameters had significant main effects except for continuity, which failed to reach significance (p = 0.29). Since continuity was also not included in any significant interactions, it was excluded from the statistical analysis in order to increase the power of the analysis. An ANOVA was run again without continuity.

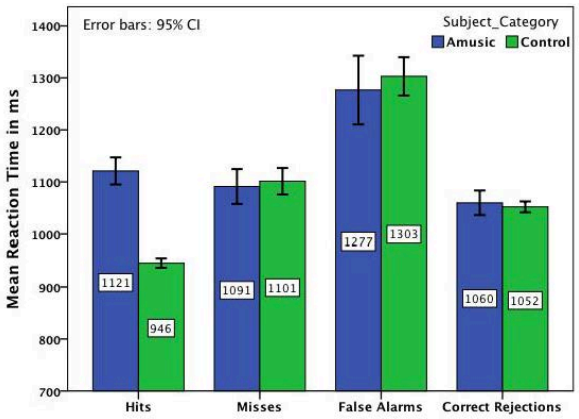


Figure 2. *Mean reaction times of amusics and controls for hits, misses, false alarms and correct rejections.*

A significant main effect was found for group ($F(1, 35)$ = 4.89, p = 0.034), indicating that amusic participants were slower to respond than control participants.

Variation Step	1	2	3	4	5	6	7	average
Amusic								
Mean (in ms)	1215	1208	1115	1079	1142	1109	1090	1121
SD (in ms)	482	384	394	343	377	360	355	374
Control								
Mean (in ms)	1112	1035	982	936	915	901	885	946
SD (in ms)	336	305	320	280	273	268	244	290
t-test								
t	1.414	4.188	3.094	4.442	6.697	6.922	7.070	12.501
p	0.163	**< 0.001**	**0.002**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**	**< 0.001**

Table 2. *Descriptive statistics and results of t-tests comparing amusics' and controls' reaction times across variation steps; SD: standard deviation; t: test statistic of the independent samples t-test; p: probability value; bold face indicates significant results.*

Significant main effects were also found for: Length ($F(1, 35) = 11.70$, $p = 0.002$), indicating that the discrimination of long stimuli ($M = 1008$, $SE = 33.2$) was faster than of short stimuli ($M = 1048$, $SE = 31.7$); direction ($F(1, 35) = 50.81$, $p < 0.001$), indicating that the discrimination of statements ($M = 979$, SE 32.8) was faster than that of questions ($M = 1077$, $SE = 32.6$); and type ($F(2, 70) = 7.82$, $p = 0.001$), indicating that the discrimination of voice stimuli ($M = 997$, $SE = 33.5$) was faster than of pulses ($M = 1028$, $SE = 30.5$) or sine stimuli ($M = 1059$, $SE = 35.3$). This last main effect reflected the significant difference between voice and sine stimuli ($p < 0.001$). There were no significant interactions between group and any of the other parameters. The absence of any significant interactions with group shows that the investigated parameters influenced the reaction times of amusics and controls in the same way, while still maintaining the main effect of group, i.e. that amusics are generally slower. There was a significant interaction between length and direction but since this interaction does not involve a group difference, it is not analyzed further at this point.

3.2. Performance accuracy analysis

The hit rate was calculated by dividing the number of correct responses for different trials by the total number of different trials. Across all conditions, amusics ($M = 0.42$, $SD = 0.20$) and controls ($M = 0.80$, $SD = 0.23$) differed significantly ($t(12) = -3.35$, $p = 0.006$). An analysis using Signal Detection Theory, a psychophysical approach of measuring performance, while taking the individual's ability to discriminate and their response bias into consideration [28], was also conducted. All data were analyzed using a regression analysis. A detailed discussion of this analysis with a subset of the participants can be found in [24]. Due to lack of space, only the results of the hit rate analysis will be reported here briefly. Main effects of group ($p < 0.001$), direction ($p < 0.001$), type ($p < 0.001$) and interval ($p < 0.001$) were found and interactions between group and type ($p < 0.001$) and group and interval ($p < 0.001$) were also found.

4. Discussion

The present study amasses more evidence that congenital amusia negatively influences speech perception by showing that amusics performed behaviorally worse and slower than controls for non-linguistic as well as linguistic stimuli. This supports earlier studies [14-17] claiming that amusia is not limited to the musical domain. While Patel et al. [14] found that only a subset of amusics had an impaired discrimination of linguistic material, in the present study all amusics were impaired (cf. also [15]).

Furthermore different parameters influencing amusics' perception were considered. It was shown that even at a distance of seven semitones, amusics were still impaired in comparison to controls. Their hit rate was not only lower, their reaction times were also significantly slower (see also [8, 21, 22]). Continuity of the stimuli did not significantly influence perception even though amusics performed slower/worse for discontinuous stimuli, supporting the findings by [16] where amusics performed worse for discrete stimuli. Questions, i.e. rising pitch changes, were discriminated slower and worse by amusics and controls. This is in contrast to the findings of [2], where the one tested amusic could only detect rising pitch changes. We hypothesize that statements might have been easier to discriminate since they appear more often than questions in real speech.

There were also dissociations between reaction times and performance accuracy: For length, there was no influence on the performance accuracy, but on RTs: surprisingly, long stimuli were discriminated faster. And while amusic and controls performed faster for linguistic stimuli, controls performed significantly better for non-linguistic stimuli, which was not the case for amusics. Amusics' performance accuracy did not differ for linguistic and non-linguistic stimuli. This supports findings by [16], who also found that the presence or absence of linguistic material did not influence amusics' performance, while controls performed better for non-linguistic material. This dissociation between reaction times and performance accuracy as well as the influence of the pitch change direction need to be investigated further.

5. Conclusions

In the present study, we investigated the pitch perception of amusics. Our goal was to gather further evidence that amusia does indeed affect intonation perception and to gain insight into factors that might have an influence. We found that amusics performed significantly worse across the entire experiment. Their pitch perception was impaired for speech stimuli and tonal analogs even at a distance of seven semitones, which is in line with earlier studies ([14-17]). This further substantiates the hypothesis that congenital amusia is not domain-specific but rather a general perceptual impairment. Concerning possible factors influencing amusics' perception, stimuli length and direction of pitch change were shown to play a role while continuity of stimuli did not. These and other linguistic parameters require further investigation.

6. References

[1] Stewart, L., "Fractionating the Musical Mind: Insights from Congenital Amusia", Current Opinion in Neurobiology, 18: 127-130, 2008.

[2] Peretz, I., Ayotte, J., Zatorre, R., Mehler, J., Ahad, P., Penhune, V., *et al.*, "Congenital Amusia: A Disorder of Fine-Grained Pitch Discrimination", Neuron, 33: 185-191, 2002.

[3] Foxton, J. M., Dean, J. L., Gee, R., Peretz, I., and Griffiths, T., D., "Characterization of Deficits in Pitch Perception Underlying "Tone Deafness"", Brain, 127: 801-810, 2004.

[4] Ayotte, J., Peretz, I., and Hyde, K., "Congenital Amusia - a Group Study of Adults Afflicted with a Music-Specific Disorder", Brain, 125: 238-251, 2002.

[5] Hutchins, S., Gosselin, N., and Peretz, I., "Identification of Changes Along a Continuum of Speech Intonation Is Impaired in Congenital Amusia", Frontiers in Psychology, 1: 1-8, 2010

[6] Hyde, K. and Peretz, I., "Brains That Are out of Tune but in Time", Psychological Science, 15: 356-360, 2004.

[7] Williamson, V. J. and Stewart, L., "Memory for Pitch in Congenital Amusia: Beyond a Fine-Grained Pitch Discrimination Problem", Memory, 18: 657-669, 2010.

[8] Omigie, D., Pearce, M. T., and Stewart, L., "Tracking of Pitch Probabilities in Congenital Amusia", Neuropsychologia, 50: 1483-1493, 2012.

[9] Loui, P. and Schlaug, G., "Impaired Learning of Event Frequencies in Tone Deafness", Annals of the New York Academy of Sciences, 1252: 354-360, 2012.

[10] Peretz, I., Saffran, J., Schön, D., and Gosselin, N., "Statistical Learning of Speech, Not Music, in Congenital Amusia", Annals of the New York Academy of Sciences, 1252: 361-366, 2012.

[11] Gosselin, N., Jolicœur, P., and Peretz, I., "Impaired Memory for Pitch in Congenital Amusia", Annals of the New York Academy of Sciences, 1169: 270-272, 2009.

[12] Tillmann, B., Schulze, K., and Foxton, J. M., "Congenital Amusia: A Short-Term Memory Deficit for Non-Verbal, but Not Verbal Sounds", Brain and Cognition, 71: 259-264, 2009.

[13] Williamson, V. J., McDonald, C., Deutsch, D., Griffiths, T. D., and Stewart, L., "Faster Decline of Pitch Memory over Time in Congenital Amusia", Advances in Cognitive Psychology 6: 15-22, 2010.

[14] Patel, A., Wong, M., Foxton, J., Lochy, A., and Peretz, I., "Speech Intonation Perception Deficits in Musical Tone Deafness (Congenital Amusia)", Music Perception, 25: 357-368, 2008.

[15] Liu, F., Patel, A. D., Fourcin, A., and Stewart, L., "Intonation Processing in Congenital Amusia: Discrimination, Identification and Imitation", Brain, 133: 1682-1693, 2010.

[16] Liu, F., Xu, Y., Patel, A. D., Francart, T., and Jiang, C., "Differential Recognition of Pitch Patterns in Discrete and Gliding Stimuli in Congenital Amusia: Evidence from Mandarin Speakers", Brain and Cognition, 79: 209-215, 2012.

[17] Liu, F., Jiang, C., Thompson, W. F., Xu, Y., Yang, Y., and Stewart, L., "The Mechanism of Speech Processing in Congenital Amusia: Evidence from Mandarin Speakers", PLoS ONE, 7: e30374, 2012.

[18] Thompson, W. F., Marin, M. M., and Stewart, L., "Reduced Sensitivity to Emotional Prosody in Congenital Amusia Rekindles the Musical Protolanguage Hypothesis", Proceedings of the National Academy of Sciences, 109: 19027-19032, 2012.

[19] Nan, Y., Sun, Y., and Peretz, I., "Congenital Amusia in Speakers of a Tone Language: Association with Lexical Tone Agnosia", Brain, 133: 1- 8, 2010.

[20] Jiang, C., Hamm, J. P., Lim, V. K., Kirk, I. J., Chen, X., and Yang, Y., "Amusia Results in Abnormal Brain Activity Following Inappropriate Intonation During Speech Comprehension", PLoS ONE, 7: e41411, 2012.

[21] Albouy, P., Schulze, K., Caclin, A., and Tillmann, B., "Does Tonality Boost Short-Term Memory in Congenital Amusia?", Brain Research, 1537: 224-232, 2013.

[22] Albouy, P., Mattout, J., Bouet, R., Maby, E., Sanchez, G., Aguera, P.-E., *et al.*, "Impaired Pitch Perception and Memory in Congenital Amusia: The Deficit Starts in the Auditory Cortex", Brain, 136: 1639-1661, 2013.

[23] Peretz, I., Champod, S., and Hyde, K., "Varieties of Musical Disorders: The Montreal Battery of Evaluation of Amusia", Annals of the New York Academy of Sciences, 999: 58-75, 2003.

[24] Hamann, S., Exter, M., Pfeifer, J., and Krause-Burmester, M., "Perceiving Differences in Linguistic and Non-Linguistic Pitch: A Pilot Study with German Congenital Amusics", in Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, Thessaloniki, Greece, 398-405, 2012.

[25] Boersma, P. and Weenink, D., "Praat: Doing Phonetics by Computer," 5.2.25 ed, 2011, retrieved 12 May 2011 from http://www.praat.org/.

[26] Hyde, K. L. and Peretz, I., ""Out-of-Pitch" but Still "in-Time"", Annals of the New York Academy of Sciences, 999: 173-176, 2003.

[27] Patel, A. D., Foxton, J. M., and Griffiths, T. D., "Musically Tone-Deaf Individuals Have Difficulty Discriminating Intonation Contours Extracted from Speech", Brain and Cognition, 59: 310-313, 2005.

[28] Green, D. M. and Swets, J. A., Signal Detection Theory and Psychophysics. New York: Wiley, 1966.

Computational annotation-mining of syllable durations in speech varieties

Jue Yu[1], *Dafydd Gibbon*[2], *Katarzyna Klessa*[3]

[1] School of Foreign Languages, Tongji University, China
[2] Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
[2] Institute of Linguistics, Adam Mickiewicz University in Poznań, Poland
gibbon@uni-bielefeld.de, erinyu@126.com, klessa@amu.edu.pl

Abstract

There are many techniques for modelling properties of speech duration patterns, including models of rhythm as oscillation, partial models of rhythm types as departures from isochrony, models of tempo acceleration and deceleration, and models of duration hierarchies and their relation to hierarchies in word and phrase structure. Except for oscillator modelling, many approaches use data extraction from speech annotations, often with mainly manual methods. We employ computational data-mining for phonetic research, as opposed to phonological research on the one hand or speech technological research on the other, and explore the potential of the computational annotation data-mining paradigm for improving efficiency and scope of analysis. We show consistent variation in syllable duration patterns in selected speech varieties in English, Chinese and Polish, chosen for their known different prosodic typological properties. Results include a possible limen of 50ms for relevant timing patterns. For data-mining we use the *Time Group Analysis* (*TGA*) methodology, directly in the *TGA* online tool and integrated into the *Annotation Pro+TGA* desktop software.

Index Terms: prosody, syllable duration, speech style, register, dialect, annotation mining, English, Chinese, Polish

1. Introduction: domain and methods

Inter-variety differences in speech duration patterning have been studied mainly in the context of prosodic typology and native-foreign pronunciation. The present sociophonetic contribution addresses the issue of intra-language variation in syllable durations at the phonology-phonetics interface, in pilot case studies of different registers or speech styles in the same dialect (English, Polish) and different dialects in the same register (Mandarin) using the computational annotation-mining paradigm [1], [2]. We concentrate on syllable duration patterns in interpausal time groups. Annotation practice in this field has been criticised for lack of a precisely specified empirical basis [3], [4], so pause and syllable annotation criteria require comment.

Pauses are typically defined acoustically with a minimal duration criterion such as 100, 150, or 200 ms ([5], [6], [7], [8], [9], [11], [12], [13], [14], [15], [16]): auditorily by holistic annotator perception (whether actual silence, or associated with final lengthening and other item-final features), or functionally (with syntactic boundary, hesitation). The annotations used here are grounded in a heuristic combination of acoustic, visualised and auditory criteria, with actual acoustic pause lengths sometimes much less than the commonly proposed minimum of 100 ms. Explicit functional criteria were not used, in order to avoid circularity in later studies of the relation between interpausal groups and grammatical constituents. The segmental content of so-called 'filled pauses' was not treated as a pause.

Syllable annotation is based on more language-dependent criteria than for pauses. The initial criterion of word boundary as syllable boundary is relatively straightforward for Mandarin and also for English. For Polish the criterion is more complex (cf. proclitics): a modified *Maximal Onset Principle* was used, with two constraints on onset structure: non-decreasing sonority, and attested actual occurrence as word onset [17]. Word-internally, ambisyllabic consonants were annotated as onsets of the following syllable in English and Polish; the issue does not arise in Mandarin phonotactics. We are not concerned with syllable-internal boundaries.

The literature reveals many techniques for measuring properties of speech duration patterns, including acoustic models of rhythm as oscillation, and annotation mining with partial models of rhythm types as degrees of isochrony, models of acceleration and deceleration, and of duration hierarchies and their relation to word and phrase hierarchies. We focus on annotation mining (cf. Section 2). Section 3 presents the results of the three pilot case studies on English, Mandarin and Polish, and Section 4 contains a summary and conclusion, and outlines future research and applications.

2. Annotation data-mining

We define speech annotation data-mining as the extraction of structured information from speech annotations, and use computational annotation mining tools for efficiency, consistency and handling large corpora: the *Time Group Analysis* (*TGA*) online tool [18] and *TGA* functions integrated into *Annotation Pro* [19], [20]. We apply the tools to reliable statistical distributional analysis of syllable duration relations and patterns, many of these properties being relatively inaccessible to manual approaches, at least for large data sets.

2.1. Annotations

Annotations are in general modelled as two-dimensional constructs structured as parallel symbolic information streams (tiers, layers) of event tags. Event tags are represented as label-interval pairs, where the intervals Δt can be represented in a number of ways: (1) as single time-stamps t for the event start or end with a second time-stamp implicitly provided by an adjacent event tag (ESPS, HTK, BOSS); (2) time information pairs: (a) event beginning and end time-stamps t_1, t_2 (Praat, Transcriber, WaveSurfer, ELAN, Anvil), or (b) event beginning time-stamp t_1 and duration Δt for $\Delta t = t_2 - t_1$ (Annotation Pro [20]). Time-stamps are typically represented (1) as sample numbers (with sample-rate stored in the annotation file metadata for conversion to time values),

(2) as clock time. Rarely, time-stamp triples may be defined for event beginning, centre (or peak) and end (SAM).

The parallel symbolic information streams offer two levels of complexity in annotation data-mining: *intra-stream relations* of sequence and hierarchy in single streams, and *inter-stream relations* of overlap or synchronisation [21], with hierarchy as a special case of overlap. Formal models from graph theory [22], event logic [23], automata theory [24] and interval calculus [25] are available for representing and computing with annotations.

Two kinds of intra-stream information can typically be mined from annotations: (1) distributional properties of label sequences and intervals (cf. Section 3.1); (2) interval duration $I=\Delta t$ and duration difference relations $\Delta I=\Delta\Delta t$ such as interval duration dispersions (or inversely: regularity or isochrony[1]), and interval acceleration and deceleration slopes. We focus on interval duration distributions and duration patterns in interpausal syllable sequences.

2.2. Duration dispersion or isochrony

Measures of duration dispersion (or its inverse, relative isochrony) which have been used for various phonetic units include standard deviation and the models shown in Table 1: *Pairwise Irregularity Measure*, *PIM*; *Pairwise Foot Difference*, *PFD*; *raw* and *normalised Pairwise Variability Index*, *rPVI* and *nPVI*; cf. references and discussion in [1]. *PIM* is a ratio model, *PFD* is a simplified variance model. The *nPVI* is a difference limen model $100*\Delta I/(I/2)$. (rather than the usual $\Delta I/I$, yielding an asymptote of 200 for the *nPVI*, not the usual 1.0), and eliminates rate change effects by comparing neighbours, not all intervals.

Table 1: *Definitions of PIM, PFD, PVI measures.*

$PIM(I_{1,\dots n})$	=	$\sum_{i\neq j}\lvert\log\frac{I_i}{I_j}\rvert$
$PFD(foot_{1\dots n})$	=	$\frac{100\times\sum\lvert MFL-len(foot_i)\rvert}{len(foot_{1\dots n})}$ where MFL = 'mean foot length'
$rPVI(d_{1\dots m})$	=	$\sum_{k=1}^{m-1}\lvert d_k-d_{k+1}\rvert/(m-1)$
$nPVI(d_{1\dots m})$	=	$100\times\sum_{k=1}^{m-1}\lvert\frac{d_k-d_{k+1}}{(d_k+d_{k+1})/2}\rvert/(m-1)$

The models are typically used for specific label types (foot, vocalic and consonantal intervals, syllables), but the formulae are neutral in this respect and may be used for any intervals. The models are not equivalent: Figure 1 shows correlations between the measures for utterances of 5 speakers of Brazilian Portuguese [1]; there is considerable inter-speaker variation in the correlations, and while *corr*(*SD*,*PFD*) is predictably high, *corr*(*PFD*,*nPVI*) and *corr*(*SD*,*nPVI*) are lower, though similar to each other. In general, *PIM* does not relate well to the other measures. The models have been called 'rhythm metrics', but they only fulfil one necessary rhythm condition of *relative* ('*fuzzy*', '*sloppy*') *isochrony*. They fail on the equally necessary condition of *rhythmic alternation*, since the use of absolute (unsigned) values does not distinguish between negative and positive duration changes [1], [2]. The *SD*, *PFD* and *nPVI* measures (though not *PIM*) are still useful models of relative isochrony (regularity, 'smoothness', 'evenness') of intervals, however.

[1] Organisation of an event sequence into equal time intervals; in data transmission engineering (sometimes incorrectly spelled 'isochryny') a particular kind of synchronisation.

Like the formulae, the annotation data-mining approach itself, shown in the pilot case studies in Section 3, is domain-neutral, in that it may be applied to annotations of any segments in speech, to annotations of visual head, hand and postural gesture streams, to both of these combined (or indeed to any comparable empirical time-function). We focus on syllables.

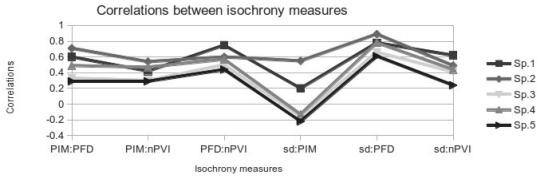


Figure 1: *Correlations between so-called 'rhythm metrics' for 5 speakers of Brazilian Portuguese.*

While the basic manual annotation mining studies noted above have been discredited as rhythm measures and have been conceptually overtaken by oscillator models [26], they have proved their worth as irregularity measures. We go a step further in using computational annotation mining for efficiency, consistency and data quantity, and for processing additional complex empirical parameters.

3. Varietal duration patterns: case studies

Contrary to *nPVI* studies, which ignore speech rate and filter out speech rate change, the case studies on English and Polish speech styles focus on acceleration, deceleration and, for Polish, also speech rate. The Mandarin study investigates relations between duration patterns and grammatical items in Beijing and Hangzhou Mandarin. At this stage, the three studies are designed to show the potential of computational annotation mining techniques with typologically different languages applied to relatively large data sets, rather than to pursue typological studies, because language variety corpora of adequate size are not readily available.

3.1. Case study 1: British English genres

The annotation data for pilot studies of annotation mining techniques with British English are taken from a subset of the Aix-MARSEC [27] database of radio speech, covering a range of sub-genres: *A* (*Commentary*), *B* (*News broadcast*), *C* (*Lecture aimed at general audience*), *D* (*Lecture aimed at restricted audience*), *E* (*Religious broadcast including liturgy*), *F* (*Magazine-style reporting*), *G* (*Fiction*), *H* (*Poetry*), *J* (*Dialogue*), *K* (*Propaganda*) and *M* (*Miscellaneous*). The genres *A*, *B*, *C*, *F* and *K* were included in this study due to their relatively similar discourse types. The Aix-MARSEC repository also contains annotation data-mining tools, but not for parameters investigated here.

Figure 2 shows averages of several metrics for the genre data (values are scaled to permit visualisation in the same graph). Except for genres *F* and *K*, values of the measures are consistently very similar, even though the speakers in each case are different and in some cases several speakers per genre are present in the corpus.

Genres *F* and *K* are outliers with regard to slope. The explanation may lie in a difference in discourse functions: genres *A*, *B*, *C* and *D* are typically formal read-aloud or rehearsed genres, while *F* is associated with more spontaneous speech, and *K*, whether read or not, would be expected to contain persuasion oriented rhetorical prosodic

features, including syllable lengthenings. Higher positive slope values mean increasing average duration, i.e. speech rate deceleration in interpausal units, contrasting with more constant speech rate in the read-aloud genres. Slope may thus be a useful discourse type marker, along with other prosodic parameters which were not represented in the annotations.

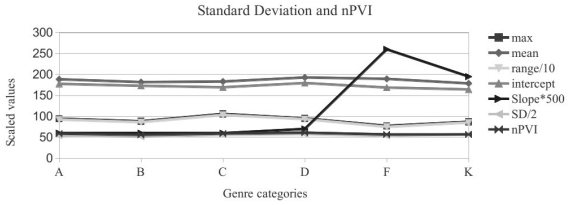


Figure 2: *Scaled annotation mining measures of six sub-genres of British radio speech.*

In addition to the analysis of duration dispersions, the *positive and negative polarities of duration differences* between neighbouring syllables were represented as *tokens*, retaining alternation (unlike the dispersion metrics), and *token sequence* distributions were registered at different *duration difference thresholds*. This technique is a first approximation to identifying actual rhythmic alternation independently of oscillator models, and in contrast to the dispersion measures shown in Table 1.

The token sequences of Table 2 (from the first utterance in the Aix-MARSEC database) show a high proportion of alternations at difference thresholds below about 50 ms; above 50 ms the difference threshold overrides many smaller alternation differences. Whether this transition at 50 ms is perceptually or functionally relevant needs more study.

Table 2: *Ranks of duration change n-grams* ($2 \leq n \leq 5$) *at thresholds 0...60* (\: *increase;* /: *decrease;* =: *same*); +, #: *word boundaries and pauses.*

Thr = 0		**Thr = 20**		**Thr = 40**		**Thr = 60**	
% (*n*)	**Seq**	**% (*n*)**	**Seq**	**% (*n*)**	**Seq**	**% (*n*)**	**Seq**
24 (65)	/\	20 (55)	/\	15 (41)	/\	17 (46)	==
23 (61)	\/	18 (48)	\/	13 (34)	\/	11 (29)	=\
13 (36)	\\	9 (24)	\#	9 (24)	\=	10 (26)	/=
17 (39)	\/\	13 (31)	\/\	9 (21)	\/\	8 (2)	===
13 (31)	/\/	10 (23)	/\/	7 (17)	/\/	6 (13)	==\
9 (21)	/\\	6 (13)	/\\	5 (11)	=/\	5 (12)	\/=
10 (20)	\/\/	7 (14)	\/\/	5 (10)	/\/\	4 (8)	====
9 (18)	/\/\	7 (14)	/\/\	4 (9)	\/\/	3 (7)	===\
5 (11)	\/\\	4 (8)	=\/\	3 (7)	=\/\	3 (7)	==/\
6 (10)	\/\/\	5 (9)	\/\/\	4 (6)	\/\/\	3 (5)	==/\/
5 (9)	/\/\/	4 (7)	/\/\/	3 (5)	\=/=\	3 (5)	+====

Preliminary studies show that a similar transition at around the 50 ms duration difference threshold can also be found in other languages and language varieties (cf. Section 3.2). However no hard and fast evidence can be given at this time. If this threshold transition at about 50 ms turns out to be generally valid, the result casts doubt on the validity of the raw duration data of previous duration dispersion studies.

3.2. Case study 2: Chinese regional accent

An issue which has not received detailed empirical attention in recent years is the relation between timing in syllable sequences and grammatical units such as words and phrases. A pilot annotation mining experiment was undertaken with recordings of 6 speakers (3 from the Hangzhou area and 3 from Beijing) reading a Mandarin Chinese translation of the IPA standard text 'The North Wind and the Sun' from the CASS corpus [28], [29]. *Time Tree* relations [21], [30] between syllable relations in interpausal groups, and words (one or more characters/syllables) were investigated.

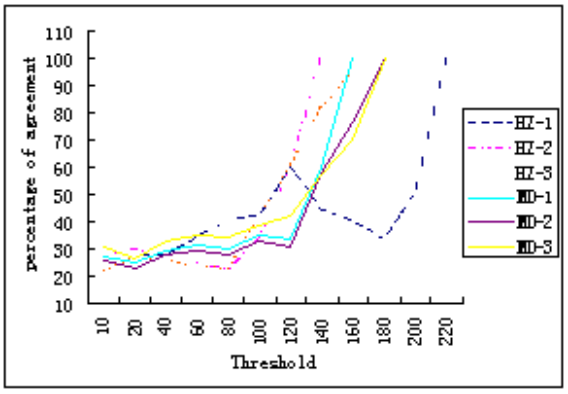


Figure 3: *Relations between duration-based syllable groupings and words for speakers of Beijing and Hangzhou varieties of Mandarin Chinese.*

The constituents were induced automatically from long-short duration patterns, where shorter constituents are prepended to longer constituents with a recursive quasi-iambic (*weak-strong*) *Time Tree* algorithm setting:

1. A syllable is a constituent.
2. A shorter constituent prepended to a longer following neighbour constituent is a constituent.
3. Nothing else is a constituent.

The algorithm was applied with all integer ms thresholds for duration differences from 0 ms to around 200 ms, where the correspondence ratio starts dropping. The following example shows a quasi-iambic *Time Tree* (represented as bracketing) of the Mandarin utterance "zhe4 shi1hou5, lu4 shang5 lai2 le5 ge4 zou3 daor4 de5" (*at that time, on the street came a traveller*), and a grammatical bracketing:

Quasi-iambic Time Tree: (((zhe4 (shi2 hou5)) (((lu4 shang5) (lai2 (le5 (ge4 zou3)))) daor4)) (de5 PAUSE))

Grammatical bracketing: ((zhe (shi hou)), (lu shang) ((lai) (le) (ge) (zou daor de)))

The groups (shi2 hou5) and (lu4 shang5) correspond to words; (ge4 zou3) is not a grammatical constituent. Also, factoring out the effect of the pause, (lai2 le5 ge4 zou3 daor4 de5) corresponds to a grammatical constituent. The correspondences between the syllable groupings and words are shown in Figure 3 (cf. also [31]).

Below a duration difference threshold of about 50 ms, correspondences between syllable groups and words are low, and are comparable among speakers. Correspondences gradually increase, and begin to diverge until about 100 ms, where they rapidly increase and interesting patterns emerge. Correspondences for Beijing Mandarin remain similar as thresholds move beyond 50 ms, while for the Hangzhou variety they are more diverse, as would be expected in a comparison between a standard accent (Beijing Mandarin) and a non-standard regional accent (Hangzhou Mandarin).

Whether the threshold limit of 50 ms is related to the limit found for English at a similar threshold order of magnitude (Section 3.1) needs further study.

3.3. Case study 3: Polish speech styles

In order to analyse syllable durations in Polish, recordings of read speech and dialogues from the Paralingua corpus [32] were used. The aim of the analysis was to look at timing patterns in speech recordings of 20 speakers in three stages of a recording session: (A) read speech produced at the very beginning of the session; (B) telephone conversations (task-oriented dialogues over the telephone); (C) read speech produced at the very end of the session after participating in a dialogue with time constraints imposed.

The time constraints were imposed only in the dialogue part of the experiment while for the final reading there were no time limits and the instruction was exactly the same as with session-initial reading (i.e. in both cases the speakers were requested to read the text in their normal, habitual way). Segmentation into interpausal time groups assumed minimal significant pause duration to be about 100 ms (cf. e.g. [5] for German). However, in some cases the minimal value of only ca. 50 ms was used, based on auditory perception and visual inspection of spectrograms by two experienced annotators. The study investigates the question whether consistent influence of the recording procedure on syllable timing could be associated with durational variability between time groups (see also [33]).

The most significant differences between the three types of speech were observed for slope, as with the English genres (Section 3.1). The overall mean values are included in Table 3. The overall mean of slopes for the dialogue recordings was significantly higher than for read speech of both types. Also, the overall means of intercepts and *nPVI* measures appeared to be highest for conversational speech.

Table 3: *Overall means of duration difference slope, intercept and nPVI for recording stages A, B and C.*

	Slope	Intercept	nPVI
A. Read 1	0,0925	145,05	43,83
B. Dialogue	0,2121	177,00	48,28
C. Read 2	0,0829	145,94	42,50

More detailed information on the individual differences between speakers in the three recording session stages can be found in Figure 4: the plots show the variability of mean slopes, intercepts and *nPVI* values for each of the 20 speakers. As shown in the figure, only in case of three speakers (23, 24, 29) was the mean slope close to the values for read speech while all the remaining speakers differentiated their slopes between read speech and dialogue.

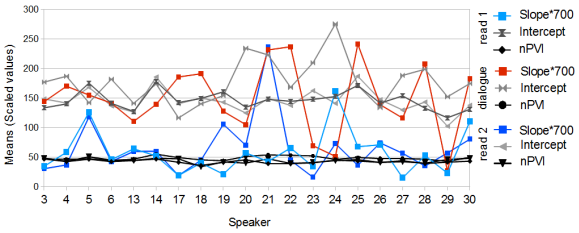


Figure 4: *The variability of selected timing properties in three speaking styles for 20 speakers of Polish.*

The two read speech tasks were very similar as regards the overall values. However, when individual results for particular speakers were investigated it was observed that the values for the initial reading tended to vary more among speakers than those for the final reading. This might suggest that after participating in the preceding tasks (20-30 minutes altogether including the inital reading and three dialogue tasks), inter-speaker differences in speech acceleration or deceleration tended to be less significant than in the earlier stages of the recording session, especially in the dialogues.

Also, speech rate (syll/sec) was measured for each recording stage. The observed overall mean value was slightly higher for the final reading (5.5. syll/sec) than for the two preceding parts (5.3, 5.4 syll/sec, respectively), but the differences were not statistically significant. Individual rate differences between the two reading tasks were mostly negligible, being exactly or almost the same for most speakers. The majority of speakers (except 13 and 17) clearly differentiated read and spontaneous tempi, but using faster or slower rate for a particular type of speech appeared to be individual rather than style-dependent. Overall rate means were in line with values observed for normal reading rate [33] and for dialogues [34] in Polish; [35] reported higher means around 6.9 syll/sec for Polish dialogues, possibly due to the different data type (fluent and coherent utterances with no unintelligible parts, false starts or hesitation sounds).

When comparing these observations with the results of slope variability measurements it was found that the two speakers whose mean rates were the same in both read and spontaneous speech still exhibited different patterns for acceleration-deceleration, as represented by the mean slope values for the two speech styles.

4. Conclusions

We have shown how new computational annotation-mining procedures can be deployed to examine a variety of interesting speech duration parameters in the sociophonetic context of speech genre, style and regional accent variation in typologically different languages. Despite the typological differences of phonology and morphology, the languages showed similarities: in a duration difference (Δt) threshold transition around 50 ms emerged (in different English and Mandarin contexts; not investigated for Polish), and in duration difference slope (English and Polish; not investigated for Mandarin). While the corpora used in the present studies were much larger than the small corpora used in previous manual annotation mining studies, in order to exploit computational annotation mining techniques fully and to move to machine learning techniques, larger annotated corpora for more languages are needed.

An interesting topic for future work is the minimal duration difference of 50 ms found in our production data for English and Mandarin. Present results do not yet permit the formulation or confirmation of relevant difference limen models of the $\Delta I/I$ type, where I is the average time-stamp difference based duration interval Δt, and ΔI is the average interval difference $\Delta\Delta t$.

We foresee applications of our computational annotation mining techniques in foreign language learning and testing studies, in modelling interfaces between phonetics in studies of phonology, prosody, grammar and discourse structure, and in evaluating naturalness in speech synthesis [36].

5. Acknowledgments

Creation of the Paralingua corpus was supported from the Polish financial resources for science in the years 2010-2012 as a development project (O R00 0170 12). The Mandarin corpus is used by kind permission of the Chinese Academy for Social Sciences. The TGA online tool is currently hosted by the Bielefeld University.

6. References

[1] Gibbon, D. and Fernandes, F. R., "Annotation-Mining for Rhythm Model Comparison in Brazilian Portuguese", Proc. Interspeech 2005, 3289-3292, 2005.

[2] Trippel, T., Gibbon, D. and Fernandes, F. R., "A BLARK extension for temporal annotation mining", Proc. LREC 2006, Genoa, 2006.

[3] Gut, U., "Rhythm in L2 speech", in Gibbon, D., Hirst, D., Campbell, N. [Eds], Rhythm, Melody and Harmony in Speech. Speech and Language Technology: Studies in Honour of Wiktor Jassem, 14/15:83–94, 2011/2012.

[4] Arvaniti, A., "The usefulness of metrics in the quantification of speech rhythm", Phonetica 66:46-63, 2009.

[5] Butcher, A., "Aspects of the speech pause: phonetic correlates and communicative function", Arbeitsberichte 15, Institut für Phonetik, Universität Kiel, 1981.

[6] Cruttenden, A., Intonation, Cambridge University Press, 1986.

[7] Dankovicova, J., "The minimum pause duration in spontaneous speech", PROPH – Progress Reports from Oxford Phonetics 5, 17–24, 1992.

[8] Dankovicova, J., Pigott, K., Wells, B., Pepp, S., "Temporal markers of prosodic boundaries in childrens speech production", Journal of the International Phonetic Association, 34 (1), 17-36, 2004.

[9] Duez, D., "Perception of silent pauses in continuous speech", Language and Speech 28.4 (377-389), 1985.

[10] Heldner, M., and Edlund, J., "Pauses, gaps and overlaps in conversations" Journal of Phonetics 38.4: 555-568, 2010.

[11] Hieke, A. E., Kowal, S and O'Connell, D. C., "The trouble with 'articulatory' pauses", Language and Speech 26.3: 203-214, 1983.

[12] Klatt, D. H. & Cooper, W. E., "Perception of segments duration in sentence contexts", in Cohen, A. & Nooteboom, S. [Eds], Structure and Process in Speech Perception, 69–89, Heidelberg: SpringerVerlag, 1975.

[13] Lehiste, I., Suprasegmentals, M.I.T. Press, Cambridge MA, 1970.

[14] Makashay, M. J., Individual differences in speech and non-speech perception of frequency and duration. PhD dissertation, Ohio State University, 2003.

[15] Megyesi, B. and Gustafson-Capkova, S., "Production and perception of pauses and their linguistic context in read and spontaneous speech in Swedish", Proc. Interspeech, 2002.

[16] Zvonik, E. and Cummins, F., "The effect of surrounding phrase lengths on pause duration", Proc. Interspeech, 2003.

[17] Demenko, G., Klessa, K., Szymański, M., Breuer, S. and Hess, W., "Polish unit selection speech synthesis with BOSS: extensions and speech corpora", in International Journal of Speech Technology, Volume 13 (2), 85-99, 2010.

[18] Gibbon, D., "TGA: a web tool for Time Group Analysis". Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013. (Cf. ref. there to online TGA tool: http://wwwhomes.uni-bielefeld.de/gibbon/TGA/)

[19] Klessa, K. and Gibbon, D., "Annotation Pro + TGA: automation of speech timing analysis". Proceedings of Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 26-31 May 2014.

[20] Klessa, K., Karpiński, M. and Wagner, A., "Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features", Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013.

[21] Gibbon, D., "Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data." in Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augurzky, P., Mleinek, I., Richter, N. and Schließer, J. [Eds], Methods in Empirical Prosody Research. Walter de Gruyter, 281–209, 2006.

[22] Bird, S. and Liberman, M., A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Linguistic Data Consortium, University of Pennsylvania, 1999.

[23] Bird, S. and Klein, E., "Phonological Events", Journal of Linguistics 26, 33–56, 1990 .

[24] Carson-Berndsen, J., Time Map Phonology: Finite State Models and Event Logics in Speech Recognition, Springer, 1997.

[25] Allen, J. F., "Maintaining knowledge about temporal intervals", in Communications of the ACM, 26 November 1983.

[26] Inden, B., Malisz, Z., Wagner, P., and Wachsmuth, I., "Rapid entrainment to spontaneous speech: A comparison of oscillator models", in Miyake, N., Peebles, D. and Cooper, R. P. [Eds], Proceedings of the 34th Annual Conference of the Cognitive Science Society, Austin, TX, Cognitive Science Society, 2012.

[27] Auran, C., Bouzon, C. & Hirst, D. J., "The Aix-MARSEC project: an evolutive database of spoken English", in Bel, B. and Marlien, I. [Eds], Proceedings of the Second International Conference on Speech Prosody, Nara, Japan, 561-564, 2004.

[28] Li A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V. and Chen, X., "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech", in Proc. Interspeech 2000, 485-488, Beijing, 2000.

[29] Yu, J., "Timing analysis with the help of SPPAS and TGA tools". Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013.

[30] Gibbon, D., "Corpus-based syntax-prosody tree matching". in Proc. Eurospeech, Geneva, 2003.

[31] Yu, J. and Gibbon, D., "Criteria for database and tool design for speech timing analysis with special reference to Mandarin", in Proc. O-COCOSDA 2012, 41-46, Macau, 2012.

[32] Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M., "*Paralingua* – a new speech corpus for the studies of paralinguistic features", in Vargas-Sierra, Ch. (Ed), Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th Int. Conf. on Corpus Linguistics (CILC2013), Procedia – Social and Behavioral Science 95. (48-58), 2013.

[33] Gibbon, D., Klessa, K., and Bachan, J., "Duration and speed in speech events", in Mikołajczak-Matyja, N., Karpiński, M. [Eds], Studies in Phonetics and Psycholinguistics. A Festschrift for Prof. Piotra Łobacz, Poznań, to appear 2013.

[34] Karpiński, M., Klessa, K., Czoska, A., "Local and global alignment in the temporal domain in Polish task-oriented dialogue". Proceedings of 7th Speech Prosody Conference, Dublin, Ireland, 20-23 May 2014.

[35] Malisz, Z., Speech rhythm variability in Polish and English: a study of interaction between rhythmic levels. PhD Thesis, Faculty of English, Adam Mickiewicz University, 2013.

[36] Gibbon, D., Moore, R. and Winski, R., [Eds], Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, 1997.

Sentence type and prenuclear contours in Brazilian Portuguese: production and perception

Izabel C. Seara[1], Juan Manuel Sosa[1], Vanessa G. Nunes[2]

[1] Laboratory of Applied Phonetics - FONAPLI, Universidade Federal de Santa Catarina, Brazil
izabels@linse.ufsc.br, sosa@sfu.ca
[2] Laboratory of Applied Phonetics – FONAPLI-UFSC and Universidade Federal de Sergipe, Brazil
vanessagnunes@yahoo.com.br

Abstract

In this paper we examine the intonation of the interrogative sentence mode in Brazilian Portuguese (BP), and how questions differ from their declarative counterparts. With this purpose, we characterize the nuclear patterns, as well as the prenuclear contours. Our aim is to identify which specific prosodic features, including prenuclear pitch range values, are systematically associated with the interrogative mode of enunciation. In the interdialectal comparison, we contrast how the speakers from Blumenau (SC) in the South and Aracaju (SE) in the Northeast, distinguish themselves from speakers of other varieties in their prenuclear patterns. These are significantly higher for the yes/no interrogatives than for the declarative types, which is not the case within other dialects in our study, including the standard varieties of Rio de Janeiro and Sao Paulo. Perception tests have corroborated the production results.

Index Terms: Phonology and phonetics of prosody, Prenuclear contours, Brazilian Portuguese

1. Introduction

In this paper, we analyse how the interrogative sentence mode is encoded in some dialects of Brazilian Portuguese (BP), and how yes/no questions differ from their neutral declarative counterparts. Our aim is to identify which specific prosodic features, including prenuclear pitch range values, are systematically associated with the yes/no interrogative mode of enunciation.

The distinction between statement and question intonation has been widely studied and has been claimed to be universal. The link between sentence mode and intonational contours has been established and the use of rising question intonation in yes–no questions has been reported for the great majority of languages, including those that are tonal.

In a number of Romance languages, such as Spanish and Brazilian Portuguese (BP), intonation is the sole method of distinguishing a yes–no question from a declarative statement. Patterns are generally assumed to be falling in statements and rising in questions, as they are in English, but as we show, this is an oversimplification. In the case of BP, more accurate is the statement by Bolinger [1] of "higher pitch somewhere in the utterance", mostly but not necessarily, rising intonation.

Indeed, the typical patterns for yes-no interrogatives in BP, although high in pitch, have been reported to have a final 'circumflex' falling intonation, and not a rising one. Lucente and Barbosa [2] transcribe the intonation of yes-no questions in BP as L+H*L%, stating that "there is a peak in the middle of the accented vowel and its fall coincides with the end of the vowel."

Moraes [3] has characterized the nuclear contours for declaratives and yes-no interrogatives in BP as H+L*L% and L+H*L% respectively, that is, both are falling intonations. The final peak for declaratives utterances happens categorically on the pre-stressed syllable, whereas the peak for the yes-no interrogative is on the accented syllable. In this same study the final contour of yes-no questions in BP is characterized as a nuclear tonal rise on the stressed syllable, followed by a fall in the following unstressed syllables. The AM notation used is the 'hat pattern' L+H*L%.

Although the unmarked, neutral yes-no questions in BP have been described as falling, as we saw, this does not exclude rising final contours with H% boundary tones. In our reseach, we have found a number of such rising contours in more marked interrogatives such as confirmatory or incredulous questions [4]. Also, some rising yes-no questions arise from truncation, that is, the incomplete rendition of the rising-falling pattern due to lack of segmental material; this occurs in utterances with stress on the final syllable, and also in cases of final vowel deletion.

It has also been reported that many dialects of BP regularly use a rising contour for the unmarked yes-no questions; for instance, in the South (Porto Alegre-- RS) [5], Lages – SC [6] and the Northeast (Aracaju - SE) [7]. In each of these dialects, a nuclear pattern L H*H% has been proposed for the nuclear region.

Some studies, such as the one in [8] have even proposed a kind of division of varieties of BP based on the final contour of interrogatives, with an isogloss that would divide the country into two halves: the one using rising contours in North, the one using the falling, 'circumflex' contour in the South.

In our data of different varieties of BP, we have found a number of such rising contours, but the analysis of their specific contexts and pragmatic uses (USE) is still in progress.

Our current findings indicate that, for the interrogative tonal nucleus in some southern regions, the pronunciation of the speakers from the cities of Florianopolis and Blumenau show a circumflex contour (L+H*L%), with or without truncation [4,6]. The performance of the two speakers from the city of Lages showed two terminal contours: L+H*H% and L+H*L%, the latter being the most frequent. We noted how the production of these utterances by the Aracaju informant had a recurrent rising contour L+H*H%, in spite of also exhibiting the L+H*L% contour [7].

Our research has also uncovered a kind prenuclear contour that is different from the one described in the literature for BP, which was like the prenuclear reported for Spanish, i. e., higher for questions than for declaratives.

In order to present the results of this research, in the following Section 2 we describe the data collection, analysis and results for the prenuclear contours that resulted from the production tests. In section 3, we present the methods used and results obtained in the perception experiments; and finally we present the conclusions we can draw from our results.

2. Production of prenuclear contours

Studies in languages such as Spanish [9] have established that sentence-initial f0 peaks of interrogatives are significantly higher than those in statements, by values that are strikingly regular in terms of tonal targets. It is for this reason, it has been claimed, that the Spanish spelling system uses the inverted question mark "**¿**" in order to indicate the beginning of the question.

This phenomenon of higher initial peaks has also been noted in languages such as Danish and Swedish and Bengali, but it does not seem to occur in English or French. But what about BP? Given some preliminary observations about this phenomenon, we investigated whether the prenuclear contour marked as well.

Moraes [3] for instance, has observed that the neutral yes-no question is characterized by a melodic rise on its first accented syllable, which is slightly higher than that observed in statements; this rise often reaches the post-stressed syllable. However, he also remarks that perceptual tests for the Rio de Janeiro dialect, have not shown that the distinction observed in relation to statements concerning the pre-nuclear accent, does plays a role in the auditory recognition of the two modalities [10]. Thus it seems that, at least for the Rio variety, it is on the nuclear accent that the contrast is concentrated. Other studies have confirmed that for the Florianópolis variant, the prenuclear region is not significantly different [11].

As we argue, yes/no questions in Brazilian Portuguese differ from declaratives based largely on the tonal structure of the nuclear contour. However, in a number of dialects such as the varieties of BP spoken in Blumenau, Santa Catarina (South), and Aracaju, Sergipe (Northeast), the increased height of prenuclear peaks seems to be a recurrent feature that characterizes yes-no questions (as well), as it occurs in the other languages referred to such as Spanish [9].

In order to test the occurrence of the wider interrogative prenuclear register span, we compared neutral declarative utterances and neutral yes/no questions in these dialects of BP.

2.1. The production experiments

Our data was collected from the AMPER-POR Project [12], with informants of both genders from the target cities in the state of Santa Catarina, Brazil: Florianopolis, Lages and Blumenau; and from the state capital of Sergipe, Aracaju

Our goal was to observe if there were differences in the prenuclear, as well as in the nuclear contours of yes-no questions between the speakers of those localities. For this, we compared the intonations of neutral declarative utterances and their corresponding (the) yes-no questions.

We analised a total of 382 neutral declarative sentences, and 382 yes-no questions. The samples were analysed automatically by the interface software of the AMPER Project [12] which generated figures that overlap the pitch curves of the declarative and interrogative modes, on the basis of three repetitions of the same sentence, as in Figure 1.

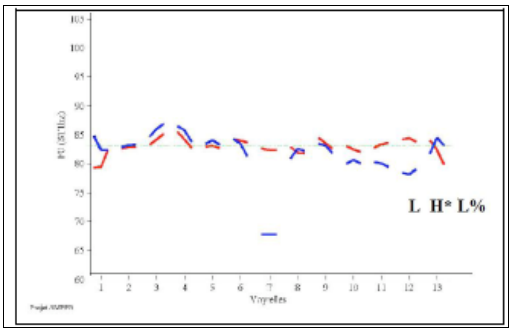


Figure 1: *Yes/no Interrogative (blue) and declarative (red) renditions of the sentence: "O Renato gosta do pássaro nadador" of the Florianopolis speaker.*

The statistical tests were done with SPSS, to verify if there were significant differences in the declarative and interrogative modes. For the statistical analysis we used SPSS Program (SPSS Statistic 17.0. Polar Engineering and Consulting, copyright 1993-2007). The dependent variable was the fundamental frequency (*f*0) and the independent variables were the position of the vowel in the prenuclear region (stressed and post-stressed) and the sentence type (declarative and interrogative). Since the data for each variable group was rather small, around 20 items in each, we chose the non-parametric Mann-Whitney U, which compares two independent groups [13]. With the value $p<0.05$, we could see whether the differences were or not.

2.2. Production results

Now we describe the initial part of sentences in order to characterize the prenuclear contour.

The male speaker from Florianópolis did not have substantial prenuclear differences between declaratives and interrogatives, as we see in Figure 1. For the male speaker from Lages, the prenuclear contour was also virtually identical for declaratives and interrogatives.

For the male speaker from Blumenau on the other hand, the first stressed syllable is much higher in interrogatives than declaratives; this was statistically significant, as shown in Table 1.

Table 1. *Mean values and standard deviation patterns of the fundamental frequency (F0) in Hz, found in the stressed and post-stressed positions of the prenucleus of the declarative and interrogative types, for males and females, and comparison of these types for the Blumenau (SC) data.*

		Declarative		Interrogative		Test*
Position	Subject	Mean (SD)	N. data	Mean (SD)	N. data	
Stressed	Male	125 (3)	20	151 (10)	20	Z = -5,42, ***p*<.001**
	Female	204 (11)	20	283 (47)	20	Z = -3,89, ***p*<.001**
Post-stressed	Male	147 (6)	20	138 (6)	20	Z = -4,64, ***p*<.001**
	Female	253 (29)	20	319 (37)	20	Z = -4,77, ***p*< .001**

*The statistic test was *Mann-Whitney U*, applied to compare the types of sentences (declarative x interrogative). SD = standard

deviation; Z = value of statistical test; *p* = significance; significant results are in bold (*p*<,05).

Our first relevant finding was that in the interdialectal comparison, the speakers from Blumenau, both male and female, distinguish themselves from the speakers of the other varieties in their prenuclear patterns, significantly higher than the declarative counterparts (as in Figure 2). This does not appear to happen with the speakers from Florianópolis [11] and Lages (as in Figure 3).

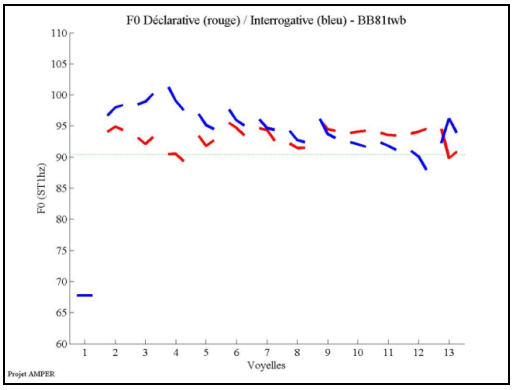


Figure 2: *Contours of yes-no questions and neutral declaratives: production of a Blumenau speaker of a yes-no question (blue) and a neutral declarative (red) of the sentence "O Renato gosta do pássaro nadador"* [7].

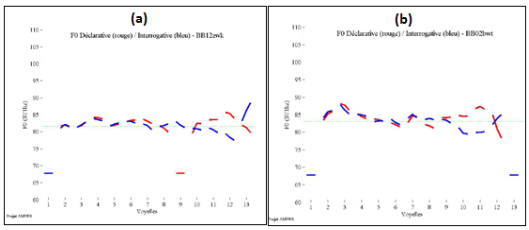


Figure 3: *Contours of yes-no questions and neutral declaratives: in (a) the production of a yes-no question (blue) and a neutral declarative (red) of the sentence "O Renato bêbado gosta do bisavô" by the Lages speaker; and in (b) the production of a yes-no question (blue) and a neutral declarative (red) of the sentence "O pássaro nadador gosta do Renato"* [7].

We found that the Aracaju data also presented this behavior (as in Figure 4), and that difference between declaratives and interrogatives in the prenuclear contour for the Aracaju data are statistically significant, as shown in Table 2.

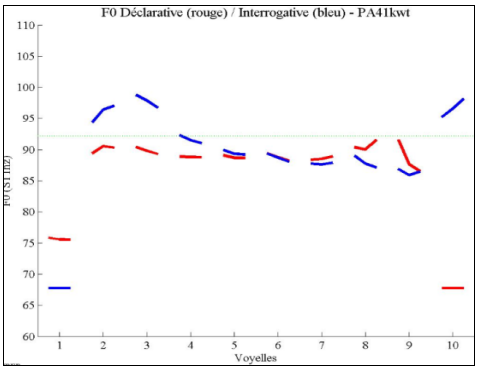


Figure 4: *Yes-no question (blue) and declarative (red) of the sentence "O bisavô gosta do Renato" by the male informant from Aracaju* [7].

Table 2. *Mean values and standard deviation of the fundamental frequency (F0) in Hz, found in the stressed and post-stressed prenuclear positions for the declarative and interrogative, of the Female group, and comparison of the two types for the Aracaju (SE) data.*

	Declarative		Interrogative		Test
Position	Mean (DP)	N. data	Mean (DP)	N. data	
Stressed	191 (22)	20	230 (54)	25	Z=3,30, **p=.001**
Post-stressed	198 (27)	20	231 (51)	22	Z=2,79, **p=.005**

These results were then verified in perception experiments with stimuli that tested whether listeners could perceive the differences in the prenuclear region, as is explained in the next section.

3. Perception of prenuclear contours

3.1. The perception experiments

We conducted three different perception experiments. In these tests, 40% of the items were distractors and 60% corresponded to sentences that showed prenuclear differences between the declarative rendition and the interrogative. The tests were set and applied as follows.

The first one (1) was to verify whether the sentence types would be recognized with only the f0 contour of the whole sentence. In this test, the subjects only heard the f0 contour, without any segmental information (filtered utterance); (2) the second was to verify if the sentence types would be perceived with only the subject NP (the first prenuclear peak) of the sentences with the actual words without hearing the end of the sentence. In this test, the subjects heard the beginning of a sentence in natural speech, either a yes/no question or a neutral declarative utterance. The listener had to decide whether he/she heard the stimulus as a statement or a question, with only the subject NP prenuclear contour to be heard; (3) the third aimed to verity if the listeners could perceive the mode of the sentence hearing only the f0 contour of the subject NP of the sentence. In this test the subjects only heard the filtered beginning of each sentence, and again

had to decide whether he/she heard the stimulus as a statement or a question, purely with the tonal information of this subject NP (see Figure 5).

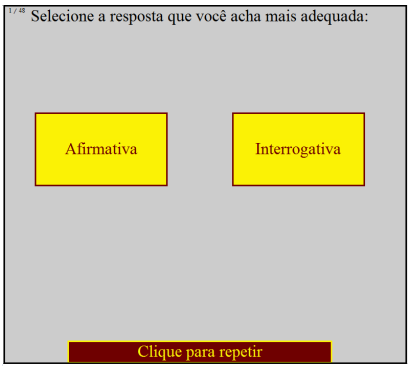


Figure 5: *Sample of perceptual discrimination test of dialects and sentence types.*

For each of the tests there were 116 stimuli for a total of 348. Each of the three types of tests were taken by six listeners. 116 stimuli X 3 tests X 6 listeners = 2088 stimuli in total. This test was staged with Praat and the results were collected automatically, also with Praat. Results proved to be consistent and significant when related to speakers from Blumenau (SC).

3.2. Perception results

Results of the perception tests were far more consistent when related to speakers from Blumenau than those from Florianopolis and Lages. Consistency was considered in terms of the number of correct responses of the subjects in relation to the type of sentence to which the subject NP belongs.

Test 1, which evaluated whether the sentence was declarative or interrogative by means of solely the information given by the filtered stimuli, resulted in the greatest score of correct responses, 67% (Figure 6).

Test 2, which evaluated the stimuli that consisted of the words that integrated the prenucleus of the sentences without any filtering, resulted in 57% of correct responses (Figure 6), with three of the listeners scoring a percentage above 60%.

Test 3, which consisted of the filtered prenuclear contour, only received 51% of correct responses (Figure 6), although two listeners scored a percentage of correct responses close to 60%.

This last result seemed to show more random responses. This was likely due to the very limited information presented to the listeners; in addition to not having the segmental information, the stimuli were very short in duration.

Thus, we can consider the results of test 2 to be more consistent regarding the initial part of the utterance, which presented real words that integrated the prenucleus of the declarative and interrogative sentences. These perceptual results seem to confirm the significant differences found in the prenuclear region in the production of the Blumenau speakers.

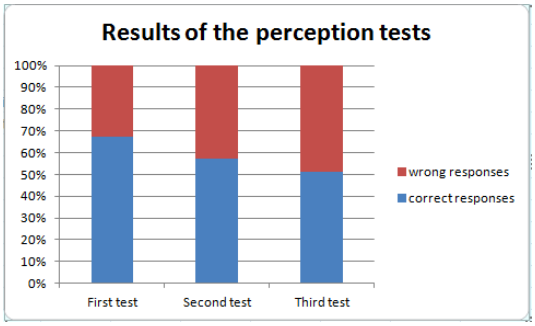


Figure 6: *Graph with the responses to the three perception tests.*

4. Conclusions

The results of our research in Brazilian Portuguese, suggest that at least in some dialects, the vertical, quantitative dimension is used significantly to distinguish sentence modes (Tables 1 and 2), as well as categories of pragmatic meaning. There is overwhelming evidence that it is in the nuclear contour that the distinctive clues are encoded, that is, in the trajectory and direction of the terminal melodic line.

The typical contours of these sentences have been shown in BP to differ in significant ways, based not only on contour shape but also on pitch-height-related phenomena. The final contour can be either rising or falling, but the peaks tend to be consistently higher in interrogatives than in declaratives.

The different sentence types are typically represented by a specific tune, or variety of tunes. Yes-no interrogatives tend to be remarkably regular in terms of the tonal design, as well as in the value of the tonal targets. We have, however, identified more than one typical tune, according to the dialect. These differences are also perceived in the prenuclear contour. We conclude that the prenuclear contour, although significant in at least those dialects of BP we described here, do not seem to be used in all the varieties we have studied thus far.

What the analysis shows is that there are significant differences for Blumenau as well as for Aracaju, and that the percentage of identification of the two sentence types, especially for Test 2, was nearly 60%.

More results and further discussion will be available as our research progresses.

5. Acknowledgements

We are grateful to Conselho Nacional de Pesquisa of Brazil – CNPq -, and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES for funding this research.

We are also grateful to Eva Christina Orzechowski Dias for her help with the statistical analysis.

6. References

[1] Bolinger. D. L. Intonation across languages. In J. Greenberg, Ed., Universals of Human Language, vol. 2. Phonology. Palo Alto, CA: Stanford University Press, pp. 471-524, 1978.

[2] Lucente, L.; Barbosa, P.A. Notação entoacional do português brasileiro em corpora de fala semi-espontânea e espontânea. Revista Intercâmbio, vol. 16. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2007.

[3] Moraes, J. A. The pitch accents in Brazilian Portuguese: analysis by synthesis, Speech Prosody 4, Online: http://www.isca-speech.org/archive/sp2008/papers/sp08_389.pdf, 2008.

[4] Sosa, Juan Manuel; Nunes, Vanessa G.; Seara, Izabel C. Variação prosódica das sentenças totais no falar catarinenses: um estudo experimental. Leitura (UFAL), n. 52, 2014.

[5] Gomes Chaves, Raquel. (2013). O comportamento entoacional das interrogativas totais e parciais na fala porto-alegrense (mscript), UFSC. 2013.

[6] Nunes, Vanessa G. Análises entonacionais de sentenças declarativas e interrogativas totais nos falares florianopolitano e lageano. Dissertação (Mestrado em Linguística) Universidade Federal de Santa Catarina, 2011.

[7] Nunes, Vanessa G. Características entonacionais de sentenças interrogativas totais nos falares catarinense e sergipano. Monografia de qualificação de doutorado. Programa de Pós-graduação em Linguística – UFSC, dez. 2013.

[8] Castelo Silva, Joelma A Prosódia regional em enunciados interrogativos espontâneos do português do Brasil. Revista Gatilho, ano VII, v.13, set., p.1-13. 2011.

[9] Sosa, J. M. La entonación del español. Su estructura fónica, variabilidad y dialectología. Madrid: Cátedra. 1999.

[10] Moraes, J. A; Abraçado, M. A descrição prosódica do português do Brasil no AMPER, Geolinguistique – Hors série – no. 3, 2005, p. 337- 345.

[11] Seara, I. C.; Silva, M. C. F. ; Berri, André . A entoação do SN-Sujeito no PB falado em Florianópolis: sentenças declarativas e interrogativas totais. Revista Internacional de Linguística Iberoamericana, v. IX, p. 157-168, 2011.

[12] Projeto Atlas Multimídia Prosódico do Espaço Dialetal Românico. http://pfonetica.web.ua.pt/AMPER-POR.htm

[13] Martins, Carla. Manual de análise de dados quantitativos com recurso ao IBM SPSS. Braga: Psiquilibríos Edições, 2011.

Use of suprasegmental information in the perception of Spanish lexical stress by Spanish heritage speakers of different generations

Ji Young Kim

Department of Spanish, Italian, and Portuguese
University of Illinois at Urbana-Champaign, USA
jkim315@illinois.edu

Abstract

The present study examines the perception of Spanish lexical stress by Spanish heritage speakers of different generations and compares their performance to that of Spanish native controls and English second language (L2) learners of Spanish. Previous studies have shown that English L2 learners experience great difficulty in perceiving Spanish lexical stress. Such difficulty is argued to be derived from English listeners using different strategies from Spanish listeners in the perception of stress. Given that Spanish heritage speakers share the same dominant language with English L2 learners (English), but differ from them with regard to the first language (Spanish), the present study intends to seek whether heritage speakers show similar or different patterns when compared with L2 learners. The present study also intends to account for the heterogeneity among heritage speakers by comparing heritage speakers of different generations. Using a forced-choice identification task with stressed minimal pairs of paroxytone and oxytone verbs, results showed that while 1st generation US-born heritage speakers pattern like Spanish native controls by paying more attention to the acoustic cues of the stimuli, 1.5/2nd generation US-born heritage speakers pattern like English L2 learners by showing bias towards paroxytone verbs.

Index Terms: speech perception, lexical stress, heritage language phonology

1. Introduction

Broadly speaking, heritage speakers are people who grow up exposed to both a majority language and an ethnic minority language. Despite their heterogeneity, heritage speakers share some characteristics in common. Generally speaking, heritage speakers, more specifically heritage speakers in the US, are early bilingual speakers of English (majority language) and a non-English home language (minority language) who have lived most or all of their lives in the US. The parents of these speakers are native speakers of the home language (heritage language), thus, heritage speakers acquire this language as their first language (L1) and, for many of them, systematic exposure to English, which is their second language (L2), does not occur until they enter institutional settings such as kindergarten and elementary school. Since English is the majority language and the heritage language is a minority language, heritage speakers' use of English increases as they grow up, whereas their use of the heritage language becomes limited to familial settings. This subsequently results in a gradual shift of language dominance from the heritage language (L1) to English (L2). Therefore, while heritage speakers have very strong command of English, their command of the heritage language is usually short of the native speaker level of their parents or peers raised in their home countries [9]. Heritage speakers are similar to L2 learners in that they are more dominant in English than the heritage language, but are different from them in that while heritage speakers learn the heritage language as the L1, L2 learners learn it as a L2. Studies on L2 phonology have shown that English L2 learners of Spanish experience great difficulty perceiving Spanish lexical stress [11]. Thus, the present study intends to see whether Spanish heritage speakers show similar patterns to English L2 learners when they perceive Spanish lexical stress.

2. Lexical stress in Spanish and English

Stress may be defined as prominence in a syllable of a word resulting from extra muscular energy that is acoustically manifested in higher pitch, longer duration, and higher intensity, in some cases with segmental effects [7]. Although researchers agree that these three universal parameters are important indices of stress, languages differ from one another with regard to what parameter functions as the primary correlate of stress. Spanish and English are typologically similar in that stress is phonologically contrastive in both languages, but they differ in the realization of stress [5]. With regard to the acoustic correlates of stress in the two languages, recent studies, have shown that when pitch accent is controlled, the primary correlate of stress is duration in Spanish [10], while it is vowel quality in English [1]. Unlike English, in which unstressed vowels undergo vowel reduction, resulting in a schwa [ə], in Spanish vowels maintain their vowel quality regardless of whether they are stressed or unstressed [5]. Thus, Spanish listeners would not be sensitive to changes in vowel quality, but rather to changes in duration and other suprasegmental cues to identify stressed vowels.

English L2 learners of Spanish are shown to have great difficulty in identifying the position of stress in Spanish even after explicit instructions [11], most likely due to influence from their native language (English). Although lexical activation is sensitive to all the information that is available in speech signals, whether it is segmental or suprasegmental, it is likely that listening strategies that are applied in the native language influence the way people listen to foreign language input [2]. English is more segmentally-based [12] and generally suprasegmental cues only provide redundant information [3]. Thus, English listeners are not accustomed to paying attention to suprasegmental information when this is the only cue available in the speech signal, unlike Spanish listeners, who are sensitive to even small differences in suprasegmental cues [3]. Therefore, the goal of the present study is to examine whether Spanish heritage speakers who are dominant in English also show difficulty in perceiving stress contrasts in Spanish due to influence from English, and whether differences are found among heritage speakers, who are found to be highly heterogeneous. Among various factors, the present study will focus on the effect of heritage speakers' generation on their perception of Spanish lexical stress.

3. Experiment

3.1. Participants

In total, 89 subjects participated in the present study: 25 Spanish native speakers (NS) (18F, 7M), 17 Spanish heritage speakers (HS) (13F, 4M), and 47 English L2 learners of Spanish (L2) (36F, 11M). The NSs were recruited in a North-Central region of Mexico. They reported that, although they have learned languages other than Spanish, this did not happen until later in life and they use only Spanish most of the time (avg. 93%). Thus, even though these speakers are strictly speaking not monolinguals, the present study will consider them as such, because they do not use the other languages functionally. The L2s are beginner to intermediate-level Spanish major or minor students recruited from a Spanish grammar course at a university in the Midwest, US. The L2 learners reported that they did not learn Spanish before the age of 9 (avg. 11.62 years) and their daily use of Spanish is less than 10% (avg. 7.34%). The HSs were born and raised in the US, mostly in the Chicago area, to Mexican families. Based on their (1) age of acquisition, (2) language use, and (3) language proficiency, it was determined that the heritage speakers were all English-dominant. That is, the heritage speakers (1) are early bilinguals in that they acquired both Spanish and English before the age of 5, which is before the period when foreign accent starts to appear if a language is not learned by then [4], (2) currently use English more frequently than Spanish, and (3) self-rated their proficiency in Spanish lower than English. The HSs were sub-divided into two groups based on whether they are 1st generation US-born Mexican-Americans (both parents are from Mexico) (HS1: 8F & 3M) (HS1) or 1.5 generation (only one parent is 1st generation US-born) or 2nd generation US-born Mexican-Americans (both parents are 1st generation US-born) (HS2: 5F & 1M) (HS2). 1.5 and 2nd generation heritage speakers were grouped together due to small number of participants (three speakers each) and based on the assumption that these speakers would behave differently from HS1s, given that they are exposed to Spanish less frequently than HS1s and they consider Spanish as a L2, unlike HS1s who consider it as a native language.

3.2. Test materials

32 stress minimal pairs that differ only in the location of lexical stress were used in the present study. The stress pairs consisted of disyllabic Spanish regular *–ar* verbs in the first person singular of the present indicative (e.g., *PAso* 'I pass') and the same verbs in the third person singular of the (simple) past perfective tense (e.g., *paSO* 'he/she/you(formal) passed'). The former case always has stress on the penultimate syllable (paroxytone) and the latter case always has it on the last syllable (oxytone). Apart from the target items, 68 fillers with verbs of both present and past tense were included. Each item was inserted in the second-to-last position of a meaningful sentence with narrow focus on the last word (subject). Whereas English has relatively fixed word order and greater flexibility in assigning the nuclear stress to words in different locations, Spanish has greater syntactic freedom and new information items are generally moved to the end of the sentence, where they receive nuclear stress [5, 6]. In the present case, by locating the subject in the last position, *Por la plaza paso yo* 'I pass through the plaza', the subject carries the narrow focus. Thus, by using such sentence structure, it was possible to separate lexical stress from pitch accent, given that the word that carries the nuclear accent is the subject (*yo*), not the target word (*PAso*). All the stimuli were produced by a male native speaker of Mexican Spanish and were recorded in a sound-attenuating booth using an AKG C520 head-mounted microphone and a Marantz PMD570 solid state recorder. The last word of each sentence was removed, leaving the sentences incomplete (*Por la plaza paso…*). Since transitions always occurred from /o/ (back vowel) to either /e/ or /j/ (front vowel/glide), using the spectrogram display in *Praat*, the cut-off point was determined as the moment when the F2 value started to make a noticeable increase. After removing the last words, all the tokens were played to make sure no transition information was included in the speech signals.

3.3. Procedures

In order to avoid priming effect, two lists were used, each containing only one member of each of the minimal stress pairs, with paroxytones and oxytones randomly distributed between the two lists. Half of the participants completed one list and half of them completed the other. The participants sat in front of a computer and listened to the incomplete sentences through headphones. In Mexico, the stimuli were presented through a Dell Vostro 230 desktop computer with KOSS UR-20 headphones and, in the US, they were presented through a Samsung SENS R410 laptop computer with Sony MDR 7506 headphones. While listening to the stimuli, the participants saw two subject words (e.g., *yo* "I" vs. *él* "he") on each side of the computer screen. Then, they decided which of the two options the following word had to be by clicking on either the left or the right key, which were indicated with colored stickers on the keyboard. The incomplete sentences were presented in pseudo-randomized order and the order of the subject words on the screen was counter-balanced. The presentation of the stimuli was done using *E-Prime*.

4. Results and Discussion

Participants' accuracy and response time (RT) were automatically collected through *E-Prime*. With regard to accuracy, all correct responses were coded as "1" and all incorrect responses were coded as "0". The effect of Group (NS/HS1/HS2/L2), Stress Pattern (paroxytone/oxytone), and the interaction of the two factors on participants' Accuracy (correct/incorrect) was statistically analyzed using logit mixed effects modeling with subject and item as random factors. The *glmer* function in the *lme4* package in R was used for the analysis. The two fixed factors (Group and Stress Pattern) were centered using contrast-coding. The best fitting model according to backwards selection included random intercepts for subject and item with no slope terms. Figure 1 shows the accuracy rate for paroxytones, oxytones, and fillers of each group. All four groups performed close to ceiling when perceiving the fillers, which confirms that the errors that occurred could not be due to problems regarding Spanish verb conjugation, which could be considered as a possible confound due to the experimental design of the present study (i.e., identify the subject of the verb). Results showed that NSs (the baseline group) performed significantly better than HS2s ($\beta = -2.1618$, SE = 0.4121, $z = -5.245$, $p < 0.001$) and L2s ($\beta = -2.8936$, SE = 0.2812, $z = -10.291$, $p < 0.001$), while the difference between NSs and HS1s was only marginally significant ($\beta = -0.6444$, SE = 0.3744, $z = -1.721$, $p = 0.0852$). A significant main effect was found in Stress Pattern ($\beta = 2.1644$, SE = 0.6195, $z = 3.494$, $p < 0.001$), indicating that

overall participants were better at perceiving paroxytones. However, a significant interaction was also found between Stress Pattern and HS2 (β = -3.2335, SE = 1.1367, z = -2.845, p < 0.01) and between Stress Placement and L2 (β = -4.3969, SE = 0.7048, z = -6.239, p < 0.001). This suggests that NSs' accuracy difference between paroxytones and oxytones was significantly different from that of HS2s and L2s. As shown in Figure 1, while NSs performed slightly better when perceiving oxytones, HS2s and L2s performed worse in this stress pattern. Such difference in direction explains why significant interactions have occurred. With regard to HS1s, no significant interaction was found between Stress Pattern and HS1, suggesting that the accuracy difference between the two stress patterns was not different between NSs and HS1s. In fact, HS1s behaved similarly to NSs, with the accuracy rate of oxytones slightly higher than paroxytones.

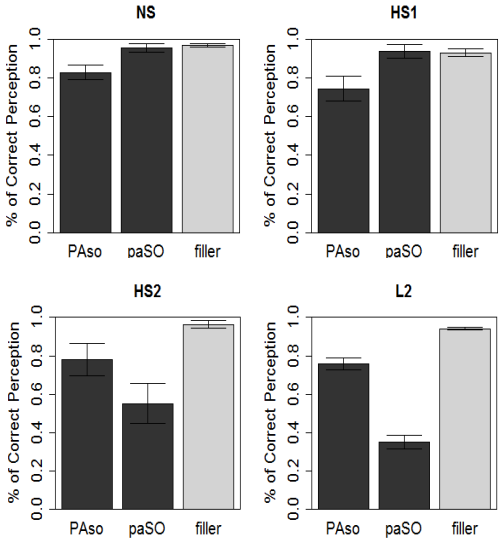


Figure 1: *Accuracy rate of paroxytones (left), oxytones (middle), and fillers (right) in each group*

It is interesting to note that while HS2s' and L2s' perception of paroxytones was target-like, their accuracy rates for oxytones were either close to chance level or lower, implying that HS2s and L2s have a bias towards selecting paroxytones. Participants' sensitivity and response bias towards paroxytones were analyzed by calculating their d-prime scores and response criterion (C scores), respectively. HIT was considered as the case in which participants selected a paroxytone when the stimulus was a paroxytone and FALSE ALARM was considered as the case in which participants selected a paroxytone when the stimulus was an oxytone. A one-way ANOVA with Group as independent variable was conducted on participants' d-prime scores and C scores. A main effect was found for both d-prime scores ($F(3,85) = 54.13$, $p < 0.001$) and C scores ($F(3,85) = 19.31$, $p < 0.001$). TukeyHSD post-hoc analyses showed that for both scores, the L2s differed from both the NSs and H1s ($p < 0.001$), while no significant difference was found between the L2s and H2s. Likewise, no significant difference was found between NSs and HS1s. Figure 2 shows that HS2s and L2s performed differently from NSs and HS1s in that their d-prime scores were close to zero, indicating that they had very low sensitivity in distinguishing the two stress patterns. The negative C scores confirm that this low sensitivity is due to bias towards selecting paroxytones. It is important to note that the perception pattern shown in the HS2 group was not consistent for all speakers in this group. Among the 6 HS2s that participated, half of them showed a strong bias toward paroxytones (paroxytone: avg. 83.33% accuracy; oxytone: avg. 20.83% accuracy). Among the remaining three speakers, two showed similar accuracy rates in the two stress patterns (paroxytone: avg. 81.25%; oxytone: avg. 84.38%) and one showed 100% accuracy rate in paroxytones, while in oxytones the accuracy rate was only chance level (56.25%). According to the HS2s' language profile, the three participants that showed a strong bias toward paroxytones had not been exposed to Spanish before avg. 7.33 years of age, while the other three acquired Spanish from birth. This finding may explain why different patterns were found among HS2s.

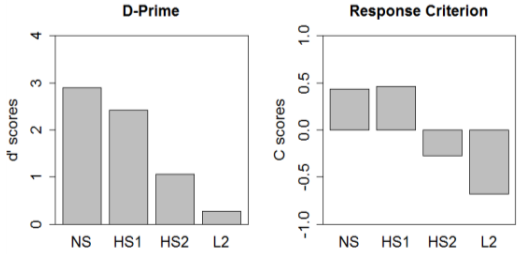


Figure 2: *D-prime (sensitivity) and response criterion (response bias) scores in each group*

Given that all the participants listened to the same stimuli, the different patterns found between NSs and HS1, on the one hand, and (half of) HS2s and L2s, on the other, can be explained through the way participants attended to the acoustic cues of the stimuli. Further analyses were conducted on the acoustic properties of the stimuli used in the experiment. Regarding the vowel quality of stressed and unstressed vowels, which is considered to be the critical acoustic cue in English, but not in Spanish, raw F1 and F2 values of the stressed and unstressed vowels were extracted and normalized using Lobanov normalization. The effect of Stress (stressed/unstressed) on the normalized F1 and F2 values was analyzed using linear mixed effects modeling with item as random factor. Tokens that were produced with creaky voice were removed from the analysis. Results showed that there was no significant effect of Stress in any of the three vowels used in the study (/a/, /i/, and /u/). Thus, vowel quality could not have had an effect on participants' response.

Apart from vowel quality, the differences in pitch, duration, and intensity between stressed and unstressed vowels were measured by subtracting the pitch, duration, and intensity values of the unstressed vowel from those of the stressed vowel in each item. As Figure 3 shows, the pitch differences were lower than zero in paroxytones, whereas in oxytones they were higher than zero, indicating that regardless of stress pattern the pitch was always higher in final syllables than in penultimate syllables. With regard to intensity differences, stress vowels had higher intensity than unstressed vowels in both paroxytones and oxytones (i.e., the values were higher than zero) and such tendency occurred to a similar degree in the two stress patterns. However, the information obtained from pitch and intensity differences does not explain why NSs and HS1s perceived oxytones with higher accuracy than paroxytones, because higher pitch in final syllables in

paroxytones would have misled their perception to oxytones and similar intensity differences in the two stress patterns would have led to similar accuracy rates. Rather, it is more likely that NSs and HS1s were sensitive to duration, which is argued to be the critical parameter for stress in Spanish [10]. The duration differences of the stimuli were slightly higher than zero in paroxytones, while in oxytones they were much higher than zero. This indicates that in oxytones the stressed vowels were much longer than the unstressed vowels, while in paroxytones, they were only slightly longer, although this difference was statistically significant (β = 60.747, SE = 13.051, t = 5.464, $p < 0.001$). This is a possible explanation of why NSs and HS1s had higher accuracy rates in oxytones. However, despite larger duration differences in this stress pattern, (half of) HS2s and L2s had lower accuracy rate, suggesting that HS2s and L2s may not pay much attention to the acoustic cues in order to understand the message, supporting [2]. Rather, it is likely that they have a bias toward selecting paroxytones, supposedly because these verbs are the first form that is learned in the classroom (present tense), hence the "easier" form to which they are more accustomed.

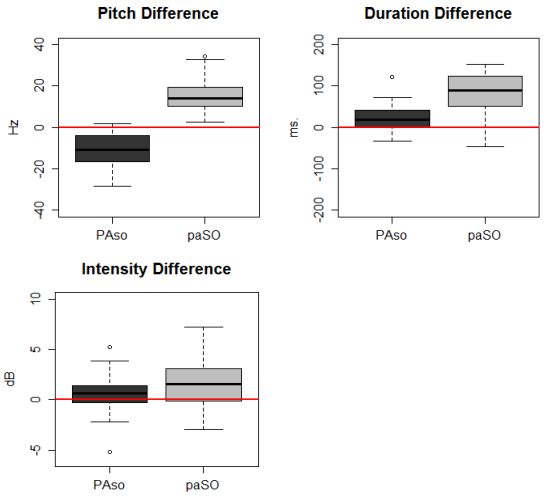


Figure 3: *Difference in pitch, duration, and intensity between stressed and unstressed vowels of stimuli*

RT was calculated as the amount of time (ms) elapsed after the onset of the target word. Only RTs of correct responses were considered in the analysis. The effect of Group, Stress Pattern, and the interaction of the two factors on participants' RT was statistically analyzed using linear mixed effects modeling with subject and item as random factors. Results showed that NSs responded significantly faster than HS2s (β = 829.71, SE = 260.76, t = 3.182) and L2s (β = 620, SE = 141.28, t = 4.388). There was no significant difference between NSs and HS1s. Regarding Stress Pattern, no main effect was found and no significant interaction was found between Stress Pattern and any of the three (non-baseline) groups, suggesting that paroxytones and oxytones were perceived with similar RTs across all four groups. That is, although the accuracy rate for oxytones was very low, HS2s and L2s did not take a particularly longer processing time when perceiving these words. This suggests that the low accuracy rate found in oxytones in (half of) HS2s and L2s is not derived from processing difficulty. Rather it is more likely to be due to bias toward paroxytones.

5. Conclusions

In the present study, the perception of Spanish lexical stress by Spanish heritage speakers and English L2 learners of Spanish was examined. Heritage speakers are similar to L2 learners in that English is the dominant language for both groups, but they are also different from each other in that while heritage speakers are exposed to Spanish at an early age, L2 learners do not learn Spanish until later. Thus, it is considered that heritage speakers have better command of Spanish phonology than L2 learners, yet not to the point that they are comparable to Spanish monolinguals, due to influence from English. The present study sub-divided the heritage speakers based on generation (1st generation US-born: HS1 vs. 1.5/2nd generation US-born: HS2). The findings indicate that while HS1s performed similarly to Spanish native speakers, HS2s were more similar to L2 learners. Given that HS2s receive less input in the heritage language than HS1s, such findings may suggest that the amount of input has an effect on heritage language phonology. However, when further analyzing the perception patterns in this group, it was found that regardless of whether a person is 1.5 or 2nd generation, those who acquired Spanish from birth showed a different pattern from those who have not been exposed to Spanish until later. This finding implies a strong effect of age of acquisition on heritage speakers' prosody, which is a linguistic aspect that is argued to be acquired from infancy [8]. However, in order to generalize such argument, it is necessary to further analyze such effect by examining a larger number of heritage speakers while taking into account factors such as language use and generation.

6. References

[1] Campbell, N. and Beckman, M. E., "Stress, prominence and spectral tilt", in A. Botinis, G. Kouroupetroglou, and G. Carayiannis [Eds], Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications, 67-70, Athens: ESCA and University of Athens, 1997.

[2] Cutler, A., "Native listening: Language experience and the recognition of spoken words", Cambridge, MA:MIT Press, 2012.

[3] Delattre, P., "A comparison of syllable length conditioning among languages", International Review of Applied Linguistics in Language Teaching, 4(14): 183–198, 1966.

[4] Flege, J. E., "The intelligibility of English vowels spoken by British and Dutch talkers", in R. D. Kent [Ed], Intelligibility in Speech Disorders: Theory, Measurement, and Management, Studies in Speech Pathology and Clinical Linguistics, 157-232, Amsterdam, The Netherlands: John Benjamins Publishing, 1992.

[5] Hualde, J. I., "The sounds of Spanish", Cambridge: Cambridge University Press, 2005.

[6] Ladd, D. R., "Intonational phonology (2nd edition)", Cambridge, UK: Cambridge University Press, 2008.

[7] Ladefoged, P., "A Course in Phonetics (4th ed.)", Fort Worth, TX: Harcourt College Publishers, 2001.

[8] Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C., "A precursor of language acquisition in young infants", Cognition, 29: 143–178, 1988.

[9] Montrul, S., "Incomplete acquisition in bilingualism: Re-examining the age factor", Amsterdam: John Benjamins, 2008.

[10] Ortega-Llebaria, M., "Phonetic cues to stress and accent in Spanish", in M. Díaz-Campos [Ed], Selected Proceedings of the 2nd conference of Laboratory approaches to Spanish Phonology, 104-118, Somerville, MA: Cascadilla Press, 2006.

[11] Saalfeld, A. K., "Stress in the beginning Spanish classroom: an instructional study", Doctoral dissertation, University of Illinois at Urbana-Champaign, 2009.

[12] Soto-Faraco, S., Sebastián-Gallés,N., and Cutler, A., "Segmental and suprasegmental mismatch in lexical access", Journal of Memory and Language, 45: 412 – 432, 2001.

Applying a fuzzy classifier to generate Sp_ToBI annotation: preliminar results

David Escudero[1], *Lourdes Aguilar*[2], *César González*[1], *Valentín Cardeñoso*[1], *Yurena Gutiérrez*[2]

[1]Department of Computer Science, University of Valladolid, Spain
[2]Department of Spanish Philology, Universitat Autònoma de Barcelona, Spain
descuder@infor.uva.es

Abstract

One of the goals of the Glissando research project[1] is to enrich a radio news corpus [1] with Sp_ToBI labels. In this paper we present the application of the automatic predictions of a fuzzy classifier to speed up the labeling process. The strategy is proposed after completing the following steps: a) manual annotation of a part of the Glissando corpus with Sp ToBI labels and checking of the coherence of the labels; b) training of the automatic system; c) validation or correction of the automatic system's predictions by a human expert. The automatic judgments of the classifier are enriched with confidence measures that are useful to represent uncertain situations concerning the label to be assigned. The main aim of the paper is to show that there exists a correspondence between the uncertain situations that are identified during an inter-transcriber experiment and the uncertain situations that the fuzzy classifier detects. Labeling time reduction encourages the use of this strategy.
Index Terms:Prosodic labeling, fuzzy classifier, Sp_ToBI

1. Introduction

Prosodic labeling aims to enrich spoken utterances with labels that are representative of the relationship between the prosodic form and function of the constituents of the message. Although prosodic labeling systems establish clear rules and protocols, the difficulty of the task and the inherent subjectivity of the labelers' judgments introduces a high number of inconsistencies. Prosodic labeling systems assume that uncertain situations could appear and reserve special symbols for representing them (like the symbol '?' in ToBI [2] and RaP [3] or the explicit computation of the transcriber disagreement in RPT [4]). Leaving aside these well-known difficulties of the task of manual annotation, the inter-transcriber tests of consistency have identified cases where two different transcribers decide to assign different labels to the same prosodic event. [5] suggested the use of alternative tiers for capturing ambiguities. These facts suggest the existence of an area of uncertainty across the categories, due to the perceptual and acoustic similarity of some pair of labels [6]. Fuzzy sets theory [7] has been widely used to represent those situations where it is difficult to classify a given element into the different possible categories.

It must be noted that recognizing uncertainty in the task of identification of some labels is not equivalent to saying that the prosodic categories used in the ToBI framework are fuzzy categories, since they are based on the description of the intonational phonology of the language, and, as a consequence, each of them is related to a clear phonological content. However, the process of annotation (either manually or with the aid of semi-automatic tools) has shown that the resulting labels can carry uncertain information when they have to be associated to the acoustic signal. In [8] we have shown how fuzzy sets can be used to represent situations where assigning a class to a given prosodic unit is difficult because of the high degree of uncertainty. The BURNC corpus [9] was used in the experiments, which is one of the most important references for studies on automatic ToBI prosodic labeling in English, and in this paper, we will present the application of the same fuzzy classifier to the subset of news of the Glissando corpus[1], which aims to be a reference for studies on Spanish prosody.

Since manual annotation is a time-consuming process and very costly in terms of human resources, efforts have to concentrate on developing tools for automatic prosodic labeling or, at least, to aid the experts to speed the process [10]. The state of the art on automatic prosodic labeling reports identification rates higher than 90% in binary decisions, such as the presence or absence of accent, boundary or break. However, when the system is faced with the classification of pitch accents, boundary tones or level of breaks, the rates dramatically decrease to about 70% (see [11] for a review of the state of the art). In [12], we showed that the reasons for these low accuracy rates are the high similarity among some pairs of classes and the imbalanced nature of the prosodic corpora. As expected, the difficulties of manual annotation are reflected in how successful automatic approaches to prosodic labeling are. This paper aims to show that the use of a fuzzy classifier considerably increases the performance when soft classification is performed, and that there exists a correspondence between the uncertain situations that are identified during the inter-transcriber experiments and the uncertain situations that the fuzzy classifier detects, a reason to consider the application of the tool as a good strategy to speed the process of manual annotation.

The strategy implies fulfilling the following steps: a) manual annotation of part of the Glissando corpus with Sp ToBI labels and checking of the coherence of the labels; b) training of the automatic system; c) correction of the automatic system's predictions by a human expert. Section 2 describes the process of manual labeling of the training corpus and the quality assessment procedure that has been applied. Section 3 presents the architecture and methods of the fuzzy classifier. The small number of editing operations in the revision process (as detailed in Section 4) evidences the quality of the fuzzy predictions. Additionally we show that the most uncertain predictions correspond to labels that are also problematic in the inter-transcriber consistency tests. We end with conclusions and the future work of this ongoing research.

[1]This work has been partially supported by Ministerio de Ciencia e Innovacion, Spanish Government (Glissando projects FFI2011-29559-C02-01,02

CORPUS	L	W	S	Pitch Accents	Boundary Tones	Breaks
Sp_ToBI (this work)	4	108	2	0.68/78.35%	0.70/85.05%	0.76/88.63%
Cat_ToBI[6]	10	264	4	0.462/61.17%	0.69/86.10%	0.68/77.14%
Am_ToBI(fe)[13]	4	644	2	0.69 / 71%	0.84 / 86%	0.65 / 74%
Am_ToBI(ma)[13]	4	644	2	0.67 / 72%	0.76 / 82%	0.62 / 74%
E_ToBI[14]	26	489	4	na / 68%	na / 85%	na / 67%
E_ToBI[15]	2	1594	1	0.51 / 86.57%	0.79/ 89.33%	na / na

Table 1: Global inter-transcriber agreement results for Sp_ToBI compared with results reported for other ToBI systems. Columns labelled *Pitch Accents*, *Boundary Tones* and *Breaks* separate results according to the respective ToBI events that have been considered. The figure in the cells are the κ index and the pairwise inter-transcriber rate (as a percentage). **L** is the number of labelers, **W** is the size of the corpus in words and **S** is the number of styles. *(fe)* is female, *(ma)* is male and *(na)* means the information is not available.

2. The Sp_ToBI manually labeled subcorpus

The Glissando news subcorpus contains recordings of eight different Spanish speakers, each of them reading more than 36 news items [1]. For our purposes, two of these speakers were chosen, taking into account differences in gender (i.e. male and female) and reading style (i.e. radio speaker and advertisement actor). The labeled corpus consists of 1100 seconds of reading of news speech recorded by two professional speakers: 12 news read by a radio professional (female voice) and 12 news read by an advertising professional (male voice). These news items include a total of 3202 words (7091 syllables) labeled with 2058 pitch accents, 1115 boundary tones and 1029 breaks.

The news data-set has been annotated using the Sp_ToBI labels proposed in [16, 17], with the modifications advanced in [18] and some adjustements needed for the speaking style, contained in the guidelines distributed in `http://veus.glicom.upf.edu/`. The tonal inventory is adapted to the specific phenomena pertaining to declarative utterances of a news data set in terms of a reduction of the tonal inventory and the definition and representation of boundary tones. In particular, the tag =% is associated to those cases where the pitch keeps the previous tone value (i.e. sustained pitch), and the parentheses stands for allotonic variations of L+H* and L+>H* (that is, when the fall is not perceived in the pre-stressed syllable (L+)H* and (L+>)H* are used).

The procedure was perceptually based: the transcriber was encouraged to focus preferentially on perception: her task consisted in listening carefully to the utterance in order to (a) mark the subjective sense of disjuncture between each pair of words and before each pause (break tier) and (b) mark prominences and tonal events (tone tier).

Since the ToBI framework is phonologically-driven, various methods of estimating the consistency and stability of the labels assigned to the corpus were conducted: (i) periodical meetings to define guidelines to annotate read news; (ii) discussion and resolution of differences in transcription throughout a six-month period and (iii) validation of consistency among transcribers with an interreliability experiment.

In order to measure the confidence of the annotation, four experts labeled independently the same news (108 words) read by a professional speaker. Pair-wise comparisons and values of the kappa index support the stability of the labels among transcribers in the main categories, but they also show that there is confusion among others. The results of the inter-transcriber consistency test can be seen in the table 1. Values of the kappa index between 0.6 and 0.8 like the ones we obtained are commonly considered as substantial agreement. These consistency rates are comparable with the ones reported in similar studies for the prosodic labeling of other corpora in different languages (see table 1). Uncertainty exists, which is the main argument that supports the use of a fuzzy classifier.

3. Automatic labeling with uncertainty

We face the automatic labeling of ToBI events following the multi-class classification approach. The multi-class classification problem has the goal to assign a ToBI label to a given prosodic unit that is, typically, a word or a syllable. The multi-class classification approach contrasts with binary classification where the goal is to determine whether an accent or a boundary is present or not in the given prosodic unit.

In [11] we showed that multi-class identification of ToBI labels can be efficiently done by using *pairwise coupling classification*. The complex multi-class classification problem is divided into several simpler problems, by means of pairwise coupling. The basic idea is that it is easy for the machine to assign a label when only two classes are considered to be possible. For example it is easy to assign the label L+H* or the label L* to a prosodic unit when the only alternatives are these two classes. However, assigning the label L+H* when the alternatives are H*, !H*, H+!H*, L+!H*, L* and L*+H is a much more challenging task. Our proposal is to combine several two-class classifiers (one for every pair of possible labels) in order to achieve the multi-class classification because two-class problems provide higher accuracy results.

Furthermore, in [12] we observed that different *types of classifiers* behave differently in the classification of different tones. Thus, decision trees seem to be specialized in the identification of the most populated classes while neural networks tend to balance the number of predicted labels for each of the classes. The compromise in term of the number of samples in every class is important in automatic prosodic labeling because prosodic corpora are naturally imbalanced. For example in the BURNC corpus, 77% of the words are labeled with two out of the eight possible labels[12]. In this work, complementarity between artificial neural networks (NN), decision trees (DT) and support vector machines (SVM) classifiers has been exploited to improve the final system, combining their outputs using a fusion method.

In order to combine the decision scores that result from the three classifications modules (DT, NN and SVM), we used the comprehensive fuzzy technique proposed in [8]. The fuzzy integral technique has proven useful for combining classifiers in several contexts [19, 20, 21, 22, 23, 24]. We use the implemen-

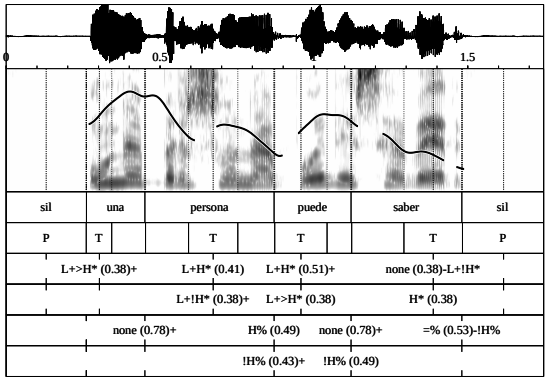


Figure 1: Multi-tier prediction interface that allows a human labeler to check the output of the automatic fuzzy classifier. Word and syllable segmentation appear in the upper tiers (T stands for lexically stressed syllable). The following tiers correspond to the predicted labels for pitch accents (aligned at the middle of the stressed syllable) and boundary tones (aligned at the end of the word).

tation of the Sugeno fuzzy integral [25] as described in [26]. As result, each ToBI category is assigned to the words of the corpus with a degree of certainty. This degree of confidence is a numeric value in the [0,1] interval (the highest the value the more the certainty). The α-cut approach permits to reduce the set of candidate labels assigned to every word. The α-cut value is empirically assigned as explained in [8].

In order to train the classifiers, we have applied a multifold approach that divides the corpus into two sections: 90% training and 10% test. Some categories show a very low number of instances, so we decided to group them with similar types thereby creating particular classes. To do that, we display the inter-label distance into a *Multidimesional Scaling* (MDS) 2D plot following the perspective adopted in [27]. This MDS map is built with the confusion matrix of a decision tree classifier: the more the inter-class confusion the closer the labels in the map. This plot allows experts to make a decision regarding the different categories. The closest categories are good candidates to be collapsed into an alternative category. As result, we use the following Sp_ToBI pitch accents: H*, L* = {L* ∪ L*+H ∪ H+L* }, L+>H*, L+H* ={L+H* ∪ (L+)H*}, L+!H* ={L+!H* ∪ (L+)!H* ∪ !H*}, L+¡H* ={L+¡H* ∪ (L+)¡H* ∪ ¡H*}; and the following boundary tones: L%, H%, =%, !H%, LH% ={LH% ∪ L!H%}. Additionally, the class "none" represents the absence of tone. After performing that clustering, classification rates improved considerably.

The input of the classifier is composed of acoustic information (F0, energy and duraction features) and POS tags as detailed in [11]. More details about this system can be found in [28].

4. Procedure of revision of the automatic system's predictions

The fuzzy classifier has been applied to unseen samples of the Glissando news subcorpus read by different voices than those used in the manual annotation. A total subset of 18 news (6 news read by 3 different voices, 2 female and 1 male) has been annotated by means of the predictions of the fuzzy classifier and a human expert has reviewed all the tags.

	Label	Intertranscriber agreement per symbol	Unique label predictions
Pitch Accent	**none**	53.5%	50.3%
	L+H*	28.6%	37.5%
	L+!H*	6.5%	7.1%
	L+>H*	4.3%	1.7%
	H*	3.2%	0.4%
	L*	3.0%	2.9%
	L+¡H*	0.9%	0.2%
Boundary Tone	**none**	78.1%	77.3%
	L%	8.1%	10.4%
	H%	7.2%	3.1%
	!H%	4.6%	7.5%
	LH%	1.1%	0.5%
	=%	0.9%	1.1%

Table 2: *Inter-transcriber agreement per symbol* is the number of times (in percentage) that two of the transcribers agree assigning the same symbol to the same prosodic unit. *Unique label predictions* is the number of times (in percentage) that the fuzzy classifier predicts only one symbol per prosodic unit.

Figure 1 illustrates the graphical interface used to present the automatic system's predictions so as to a human expert can verify or correct them. The visual interface aligns the tags predicted by the fuzzy classifier with each prosodic event: pitch accents aligned with the stressed syllables and tone boundaries aligned with the end of the word. The classifier predicts presence or absence of break (with the tag "none"), and if a prosodic rupture exists, the type of boundary tone. On the contrary, information related to the levels of prosodic rupture (break indices) is not present. This information can be inferred from the type of boundary tone, since in our results, there is a statistically significant correlation between BI3 associated to H%, !H%, =%, L!H% and BI4 associated to L% (Pearson's chi squared test X-squared = 742.1301, df = 5, p-value $<$ 2.2e-16). There are only marginal cases in which L% is associated with BI3 , and in which H%, !H%, =%, L!H% are associated with BI4.

Compared with conventional crisp classifiers, the main advantage of a fuzzy classifier is that it can provide more than one label per prosodic unit (to a maximum of three in our system), depending on the uncertainty of the predictions. Each tag is accompanied by a numerical value in the [0,1] interval, the higher the value the more the certainty, and tags are ordered from the highest to the lowest degree. At this point, it should be noted that the procedure, according to the fuzzy set theory, is not based on probabilities since the degree of certainty is independently assigned to each category. This is the reason why the values can sum up more than 1, that is, if three tags appear, we cannot infer that there are three complementary possibilities that sum up 1. On the contrary, having more than one tag in the output represents a difficult situation in which more than one label evidences a degree of certainty over the threshold set by the α-cut. Another situation that can be found is that even when only a tag is predicted, it is not necessarily accompanied by a complete confidence (marked with 1).

The task of the human expert is to evaluate the candidates: he/she checks if some of the proposed tags are the right one according to his/her perception and marks it with "+". If she/he

doesn't agree with any of the tags proposed, he/she attaches "-" and writes a new option.

Figure 1 illustrates different types of situations that the human expert can find in the process of reviewing the output of the fuzzy classifier and that reflects the uncertainty that implies using a phonologically-based system such as ToBI. With respect to tone boundaries, the most certain decision is the label "none" associated to the words "una" (a) and "puede" (can): only one label is predicted, with a high degree of certainty (0.78). In this case, the tag "none" (meaning absence of prosodic break and as a consequence, absence of tone boundary) proposed by the system is validated (+), since any tonal nor segmental cue signal a prosodic break.

On the other hand, at the end of the prosodic group "una persona" (a person) the system suggests two candidates, associated to a very similar degree of certainty: a high tone, H% (0.49) or a mid tone, !H% (0.43). Crucially, the transcriber has to decide if the difference of range is phonologically significant in this context: since the transcriber perceives that the tone decreases to a mid-tone from an L+H* nuclear pitch accent, the second option is selected.

As far as pitch accents is concerned, the difficulty in discriminating between rising accents with or without peak displacement is evidenced in the word "puede" (can). As observed in the pitchtrack, the F0 peak is situated in the syllable border, that is, both stressed and post-stressed syllables show a high tone, a fact that makes difficult the decision. The transcriber chooses the rising tone without peak displacement (L+H*), because it is generally accepted that the high tone should be completely placed in the post-stressed syllable [29].

It may also happen that the proposals of the fuzzy classifier are wrong. It is the case of the pitch accent corresponding to the word "saber" (know). The system proposes two tags: "none", meaning absence of pitch accent and a high tone (H*). Since the stressed syllable is tonally accented, but with a rising accent and not with a high tone, the transcriber dismisses both options and writes the correct one, which is L+!H*. The downstep relates to the immediately preceding high tone within the same prosodic group.

At this point, it should be said that the adequate tuning of the α-cut value (that is the number of candidates that are presented in the interface) is crucial for a correct system operation. Lowering the value of the α-cut yields a higher number of positive cases, understanding as positive those right cases found within the set of predicted labels (computed as the soft-classification rate in [8]). For the reviewer it is important to know that the probability to have the right tag in the set of candidates is really high, but on the other hand, the more the labels the harder the selection of the correct one. In our case, in the training stage we obtained soft classification rates of 82% for pitch accents and 85% for boundary tones . These rates are clearly higher than the accuracy rates that we obtain in classic non-fuzzy classification (69.2% and 81.2% respectively). The increase in the confidence rates is expected to improve the performance of the reviewing process.

In the reviewed subset of 18 news, only one label is predicted in the majority of the cases (60% for boundary tones and 45.3% for pitch accents), and few cases have three labels (3.4% for boundary tones and 2.5% for pitch accents).

5. Results

The results coming from the process of revision report that in most cases (81.8% for boundary tones and 72.6% for pitch accents) the labeler choices the first candidate of the fuzzy classifier as the right option. Only 9.2% of the boundary tones and 13.5% of the pitch accents labels needed to be edited, that means, to be corrected with a different label. In a preliminary test, the real time ratio of that labeling process has been observed to be 1:66 when the template of predictions is used. This ratio contrasts with the one obtained without any supporting template which was 1:80. These ratios have been obtained from the comparison of the time that a transcriber needs to manually label a news and the time that he/she needs to label a news of comparable size with the aid of the fuzzy classifier: in the first case, 3600 s were labeled in 44.79 s (including autochecking after the first annotation) whereas in the second case, 3000 s were labeled in 45.23 s (including as well an autochecking). The experiment has been done by the same transcriber in separate days, during a stretch of time without any interruption.

Another encouraging result is that we found a clear correspondence between the results obtained by the fuzzy automatic classifier and the results obtained in the inter-transcriber agreement tests. In the inter-transcriber test, the pair of labels that are most frequently confused by the human experts is the pair L+H* vs. L+!H* for pitch accents and the pair "none" (absence of a break) vs. !H% for boundaries (26.9% and 34.4% of the total disagreement respectively). These pairs also have the highest frequency of appearance when the fuzzy classifier predicts more than one label per word (21.8% and 29.9%, respectively). To sum up, Table 2 shows that in the range of certainty the percentages are also similar: those cases where the fuzzy classifier predicts an unique label and the frequency of intertranscriber agreement per symbol are similar for each tag.

6. Conclusions and future work

We have presented the results of an experiment in which an automatic system has been applied for the Sp_ToBI labeling of the Glissando corpus. The automatic system can generate more than one candidate label per prosodic unit according to its degree of confidence. The revision of the candidate labels provides an alternative to speed up the labeling process.

The efficiency of this strategy to speed up the labeling process is supported by the fact that only a small proportion of the predicted labels is edited. Furthermore, in most cases, the reviewer selects the first label out of the set of predictions.

The use of fuzzy labels adequately resembles the uncertainty that characterizes the human prosodic labeling process in many situations. This is evidenced by the fact that most uncertain situations for the automatic classifier correspond with labels that are the most frequently confused in manual inter-transcriber tests.

This is part of ongoing work in which an iterative training and testing process is being applied in order to improve the predictions. The reviewed labels are reintroduced in the training stage of the classifier so that the knowledge of the system increases iteratively. We are currently investigating definitions of quality metrics that measure the goodness of this iterative approach. The improvement of the revision template interface (currently in $praat$) following the suggestions of the transcribers is also current and future work.

The results of the labeling process, manual, reviewed and predicted label in the different stages, are expected to be freely available for research purposes in the web page of the project `http://veus.glicom.upf.edu/`.

7. References

[1] J.-M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. de-la Mota, C. González, C. Vivaracho, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, E. Estebas, M. Cabrera, and A. Bonafonte, "Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan," *Language Resources and Evaluation*, vol. 47, no. 4, pp. 945–971, 2013.

[2] M. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford University Press, New York, 2005, pp. 9–54.

[3] L. Dilley and M. Brown, "The rap (rythm and pitch) labeling system," Massachusets Institute of Technology, Tech. Rep. http://tedbla.mit.edu/rap.html, 2005.

[4] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[5] A. Brugos, N. Veilleux, M. Breen, and S. Shattuck-Hufnagel, "The alternatives (alt) tier for ToBI: Advantages of capturing prosodic ambiguity," in *Proceedings of Speech Prosody 2008*, 2008, pp. 273–276.

[6] D. Escudero, L. Aguilar, M. d. M. Vanrell, and P. Prieto, "Analysis of inter-transcriber consistency in the Cat_ToBI prosodic labeling system," *Speech Commun.*, vol. 54, no. 4, pp. 566–582, May 2012.

[7] D. DuBois and H. Prade, *Fuzzy sets and systems: theory and applications*. Academic Pr, 1980, vol. 144.

[8] D. Escudero-Mancebo, C. González-Ferreras, C. Vivaracho-Pascual, and V. Cardeñoso Payo, "A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling," *Computer Speech and Language*, vol. 28, no. 1, pp. 326 – 341, 2014.

[9] M. Ostendorf, P. Price, and S. Shattuck, "The Boston University Radio News Corpus," Boston University, Tech. Rep., 1995.

[10] A. K. Syrdal, J. Hirshberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Communication*, no. 33, pp. 135–151, 2001.

[11] C. Gonzalez-Ferreras, D. Escudero-Mancebo, C. Vivaracho-Pascual, and V. Cardeñoso Payo, "Improving automatic classification of prosodic events by pairwise coupling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2045 –2058, sept. 2012.

[12] C. González-Ferreras, C. Vivaracho-Pascual, D. Escudero-Mancebo, and V. Cardeñoso Payo, "On the automatic ToBI accent type identification from data," in *Proceedings Interspeech*, 2010, pp. 142–145.

[13] A. Syrdal and J. McGory, "Inter-transcriber reliability of ToBI prosodic labeling," in *Proceedings of ICSLP*, vol. 3, 2000, pp. 235–238.

[14] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the ToBI framework," in *Proceedings of ICSLP*, 1994, pp. 123–126.

[15] T. Yoon, S. Chavarría, J. Cole, and M. Hasegawa-Johnson, "Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI," in *Proceedings of Interspeech, Jeju*, 2004, pp. 2729–2732.

[16] E. Estebas Vilaplana and P. Prieto, "La notación prosódica en español. una revisión del sp_tobi," *Estudios de Fonética Experimental*, vol. XVIII, pp. 263–283, 2009.

[17] ——, "Castilian Spanish Intonation," in *Transcription of Intonation of the Spanish Language*, P. Prieto and P. Roseano, Eds. Lincom Europa, München, 2010, pp. 17–48.

[18] G. Elordieta, "Transcription of intonation of the Spanish language," in *Estudios de Fonética Experimental*, vol. XX, 2011, pp. 273–293.

[19] J. A. Benediktsson, J. R. Sveinsson, J. I. Ingimundarson, H. Sigurdsson, and O. K. Ersoy, "Multistage classifiers optimized by neural networks and genetic algorithms," *Nonlinear Anal., Theory, Meth., Applicat.*, vol. 30, no. 3, pp. 1323–1334, 1997.

[20] S.-B. Cho and J. H. Kim, "Combining multiple neural networks by fuzzy integral and robust classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 25, pp. 380–384, 1995.

[21] S. B. Cho and J. H. Kim, "Multiple network fusion using fuzzy logic," *IEEE Trans. Neural Networks*, vol. 6, pp. 497–501, 1995.

[22] P. D. Gader, M. A. Mohamed, and J. M. Keller, "Fusion of handwritten word classifiers," *Pattern Recogn. Lett.*, vol. 17, pp. 577–584, 1996.

[23] A. Verikas, A. Lipnickas, K. Malmqvist, M. Bacauskiene, and A. Gelzinis, "Soft combination of neural classifiers: A comparative study," *Pattern Recogn. Lett.*, vol. 20, pp. 429–444, 1999.

[24] D. Wang, J. M. Keller, C. A. Carson, K. K. McAdoo-Edwards, and C. W. Bailey, "Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion," *IEEE Trans. Syst., Man, Cybern.*, vol. 28B, pp. 583–591, 1998.

[25] M. Grabisch and M. Sugeno, "Multi-attribute classification using fuzzy integral," in *IEEE Int. Conf. Fuzzy Systems*, 1992, pp. 47–54.

[26] L. I. Kuncheva, ""fuzzy" versus "nonfuzzy" in combining classifiers designed by boosting," *Transactions on Fuzzy Systems*, vol. 11, no. 6, pp. 729–741, december 2003.

[27] D. Escudero-Mancebo and E. Estebas-Vilaplana, "Visualizing tool for evaluating inter-label similarity in prosodic labeling experiments," in *Proceedings Interspeech 2012*, 2012.

[28] D. Escudero, L. Aguilar, C. González, Y. Gutiérrez, and V. Cardeñoso, "On the use of a fuzzy classifier to speed up the Sp_ToBI labeling of the Glissando Spanish corpus," in *Proceedings of International Conference LREC*, Reykjavik, Iceland, May 2014, p. in press.

[29] T. Face and P. Prieto, "Rising accents in Castilian Spanish: a revision of Sp-ToBI," *Journal of Portuguese Linguistics*, vol. 6.1, pp. 117–146, 2007.

The production and perception of L1 and L2 Dutch stress

Marie-Catherine Michaux[1], *Sandrine Brognaux*[2,3], *George Christodoulides*[1]

[1]Centre Valibel, Institute for Language and Communication, University of Louvain, Belgium
[2] Cental, University of Louvain, Belgium [3] TCTS, University of Mons, Belgium

marie-catherine.michaux@uclouvain.be, sandrine.brognaux@uclouvain.be, george@mycontent.gr

Abstract

This study aims at exploring the production and perception of Dutch word stress by Francophone learners of (Belgian) Dutch. For this purpose a production experiment was first carried out. In line with other studies, it was hypothesized that participants would show a tendency to stress the final syllable. Even though this hypothesis was confirmed, there was also a substantial lack of agreement between the five labellers who perceptually annotated the data for stress position. To further investigate this matter, acoustic measures were extracted. The data suggest that both groups of speakers do not use acoustic correlates to signal prominence in the same way, the Dutch group using intensity, vocalic nucleus duration and pitch movement more, while the French group prefers duration and pitch movement. This study also led us to develop tools to phonetise, syllabify and facilitate the acoustic analysis of Dutch speech.

Index Terms: L2 prosody, Dutch as a Foreign Language, speech perception, speech production

1. Introduction

Due to discrepancies between the Dutch and French prosodic systems and to the lack of attention paid to pronunciation in Dutch didactics, Dutch word stress can be problematic for Belgian Francophone learners of Dutch as a Foreign Language (DFL). Dutch on the one hand is a variable-stress language where stress is a lexical property of words [1] that can be used contrastively (e.g., *voorkomen,* 'to happen', vs. *voorkomen,* 'to prevent'). On the supra-lexical level, Dutch uses 'accents' to signal the informational status (linked to the concept of 'focus') of words. Stress is determined by the linguistic system, whereas accent depends on the communicational aims of a speaker [2: 41]. Dutch word stress (measured on words out of focus) is acoustically correlated (mainly) with duration, spectral tilt, and to a lesser extent overall intensity and timber. Dutch accent is mainly rendered by abrupt changes in f_0, duration, spectral tilt and overall intensity [2].

French, on the other hand, does not have lexical contrastive stress: the standard 'primary accent' typically falls on the last syllable of 'accentual' word groups [e.g. 3, 4, 5]. Rather than being contrastive, this 'primary' accent has a demarcative function [6]. Besides this primary accent, French has several secondary accents falling on any syllable of the word group and covering rhythmic or emphatic functions [7]. The acoustic correlates of the primary accent are mainly duration, a change in f_0 and the potential use of pauses. An initial emphatic accent is rendered by a shorter duration but a change in f_0, potentially preceded by a pause [7].

Although Dutch is taught in most primary and secondary schools in Francophone Belgium, pronunciation and prosody are often neglected in DFL courses according to our surveyed students and teachers [also 8]. This means that most learners may not be familiar with Dutch prosody.

The production of Dutch word stress by Francophones has been addressed in studies on Dutch as a Second Language (DSL) [9] and as a DFL [10, 11]. Based on the results of these studies it seems clear that the DFL population has to be analysed separately from the DSL one, as the latter group, probably as a result of receiving another type of input (viz. native spoken Dutch), has been found to be more proficient in producing correctly located stress in simple and complex words. As for the DFL group, it was concluded that learners tend to stick to their final L1 pattern, but can also evolve to a penultimate stress (yet not always being the required stress position) with time. Research on DFL stress in nominal compounds (with the first compounding part bearing stress as a main rule) [12] has also shown that, in addition to a preference for the final syllable, DFL speakers do not necessarily show a consistent stress pattern across words.

The current research focuses on Dutch stress production by Belgian Francophone learners. Our analysis first concentrates on the realised position of the stress by DFL speakers. This analysis relies on a perceptual annotation by multiple annotators. The considerable amount of inter-rater disagreement also led us to investigate the acoustic realisation of prominences by both DFL and native speakers. Discrepancies between the groups might explain annotation confusion and annotators might rely on different acoustic correlates.

In this paper we also present some methodological aspects of our study. The tools we developed provide a detailed analysis of some aspects of DFL and native Dutch (DL1) prosody.

The paper is organized as follows. Section 2 presents the methodology and tools developed to phonetically align, annotate and analyse our corpus. The analysis of the prominences produced by DFL and native speakers is then presented and discussed in Section 3. Finally, Section 4 concludes the paper and discusses further works.

2. Method

2.1. Participants

20 DFL learners (age range 19-23, mean age 21.1, 14f, 6m) and 10 native speakers of Belgian Dutch (age range 20-51, mean age 28.6, 5f, 5m) took part in the experiment. French was the only mother tongue of the selected DFL speakers.

2.2. Materials

30 existing Dutch three-syllable words were used in the current study. They were selected and classified according to the stress rules for simplicia described in [13]. They were split into three canonical stress positions (SP): initial (*pagina*, 'page'), medial (*collega*, 'colleague') and final (*aoniem*, 'anonymous'). Each word X was randomly presented thrice in a carrier sentence (***X*** *heb ik gezegd* '**X** I said', *Ik heb* ***X*** *gezegd*

'I **X** said' and *Ik heb gezegd **X*** 'I said X'), leading to a 90-sentence reading task. Each target was presented in bold, italics and was underlined, showing focus marking.

2.3. Procedure

Speakers were recorded individually in a quiet room. Prior to the recording they filled in a form containing questions about their learner profile (duration of Dutch learning, age at start of learning, etc.). The trial phase started after an instruction and training session similar to the trial. A Tascam-07 MKII recorder and a Sennheiser PC131 head-set microphone were used.

2.4. Perceptual analysis

The data were perceptually labelled independently by two DFL-speaking native-French speakers and three native Dutch speakers, all of whom were phonetically trained. After listening to the stimuli as often as required, the annotators indicated which syllable they perceived as prominent (1-2-3). Cases of doubt could be expressed as "1?3?", etc. The annotators also gave a certainty score on a scale from 1 (very easy) to 5 (very difficult) representing the difficulty of making their decision as to which syllable was most prominent.

2.5. Linguistic analysis

An alignment between the speech signal and the phonetic transcription was necessary for the prosodic analysis of the syllables. Since manual alignment is a tedious and time-consuming task, several automatic alignment tools have been proposed in the literature [e.g. 14, 15]. They usually rely on pre-trained speaker-independent models to align new corpora. However, they cover a very limited number of languages and might not perform properly for different speaking styles. Most of these existing tools actually do not provide models for Dutch. To resolve this issue, we developed a new automatic phonetic alignment tool, *Train&Align* [16]. Its specificity is that it trains the models directly on the corpus to align, which makes it applicable to any language and speaking style. Previous experiments have shown that it provides results comparable to the other existing tools [17]. It also offers additional options like "bootstrap", allowing for a manually-aligned part of the corpus to be used to improve the model quality. While a basic alignment of our corpus with *Train&Align* achieved rather poor alignment rates, the use of 40 seconds of bootstrap led to significant improvement, reaching alignment rates of about 82% with a 20 ms tolerance threshold. The aligned files provided in TextGrid format were easily imported in *Praat* for further prosodic analyses. This alignment was then manually checked.

The corpus was then syllabified in *Praaline* [18] (cf. 2.6). A basic rule-based syllabifier relies on the sonority sequencing principle (sonority should increase from the first phoneme of an onset to the nucleus), and the maximal onset principle, which states that a syllable's onset should be extended at the expense of the preceding syllable's coda. Such a simple rule-based approach has been shown to achieve an accuracy of 93-95% [19]. The syllabifier was adapted to Dutch by providing a list of valid onsets. The syllabification was manually checked.

2.6. Corpus processing

In order to further process the data, we used *Praaline* [18], a toolkit for corpus management, annotation, querying and visualization. It interfaces with *Praat* and stores corpus data as a relational database, allowing the user to add external data sources. The annotator labels from the perceptual analysis were imported and linked to the corresponding corpus syllables. *Praaline* runs a cascade of scripts and/or external analysis tools, each of which may add features to an annotation level (e.g. syllables, words etc.). Using this interface, we applied *Prosogram* [20] for pitch stylisation on the entire corpus. *Prosogram*'s algorithm operates in two phases; for each syllable, vocalic nuclei are detected based on intensity and voicing. The f_0 curve on the nucleus is then stylised into a static or dynamic tone, based on a perceptual glissando approach. Several syllable features (duration, pitch, pitch movement etc.) were added to the database.

Subsequently, we constructed the datasets for further statistical analysis using a query editor. *Praaline* queries may include data from multiple levels of annotation, and the features of one level may be aggregated or normalised over another level. In this study, we correlated the perceptual annotations to the prosodic features of syllables. For each prosodic feature, we also calculated a z-score value normalised over each speaker. Queries may also include functions to calculate derived measures; we used this feature to obtain relative measures (cf. 3.4). The statistical analysis was performed using SPSS (v. 21) and R (v. 3.0.2).

3. Results

3.1. Perceived stress position

3.1.1. Inter-rater agreement

An inter-rater agreement analysis using the Fleiss' Kappa statistic [21] was performed on the DFL and DL1 data to determine consistency among annotators. "Low certainty" comprises all the cases for which the majority of annotators ($n \geq 3$) expressed a low confidence ordeal about the decision they made as to which syllable was bearing stress (certainty score > 1 on a scale from 1 to 5, see 2.4.). "High certainty" refers to high confidence in the annotators' decisions (score = 1, see 2.4.). As shown in Table 1, the κ-values are always lower for the DFL group than for the control. For all cases taken together the agreement is moderate ($\kappa = 0.570$) for the DFL group [22] and almost perfect for the DL1 group ($\kappa = 0.980$). Cases of high certainty comprise 75.88% of the cases for the DFL and 98.00% for the control group. While κ is substantial for the DFL group and almost perfect for the DL1 group for cases with high certainty, the κ-value drops to fair levels for low-certainty cases.

	DFL	DL1
General	0.570 (n=1799)	0.980 (n=900)
High certainty	0.681 (n=1365)	0.984 (n=882)
Low certainty	0.239 (n=434)	0.385 (n=18)

Table 1: *Inter-rater agreement (Fleiss' κ and counts) for overall annotations ("General"), and cases with high and low certainty per L1 group.*

3.1.2. Consensus

Based on the annotations of each labeller, a consensus variable was computed per word. Consensus is reached when *all* annotators marked the same syllable as prominent. Table 2 shows the consensus values per canonical stress position (SP)

for the DFL speakers. The shaded cells contain the cases where canonical and perceived stress concur, meaning that the stress was perceived and therefore probably produced in the expected position. The overall percentage of "correct" stress amounts to 26.7% (vs. 96.1% for the control group). Consensus over each syllable is not equally distributed over the three canonical SPs (χ^2 (4) = 188.69, p < .001). Canonical SP3 yields the best results (39.90% correct), followed by SP2 (25.8%) and SP1 (14.3%). On the whole DFL speakers tend to stress the 3rd syllable most often regardless of the canonical SP (25.80%), confirming our hypothesis. However this result is mainly due to the high percentage of 3rd-syllable stress in SP3.

Canonical SP	No consensus		Syll 1		Syll 2		Syll 3		Total	
1	53.20	(319)	14.30	(86)	12.80	(77)	19.70	(118)	100.00	(600)
2	52.00	(312)	4.30	(26)	25.80	(155)	17.80	(107)	100.00	(600)
3	57.70	(286)	6.50	(39)	5.80	(35)	39.90	(239)	100.00	(599)
	51.00	(917)	8.40	(171)	14.80	(267)	25.80	(464)	100.00	(1799)

Table 2: *Percentages (and counts) consensus between annotators for 1st, 2nd and 3rd canonical stress position broken down by perceived stress position.*

Strikingly, in 51.00% of the cases no consensus is reached. There are two possible reasons for this: either one or more annotators did not label the same syllable as being prominent due to a difference in perception or in acoustic correlates, or the speakers have produced multiple prominences within the same word, leading to 'doubt' cases (e.g. 1?3?). Table 3 shows all these cases where annotators labelled multiple syllables as prominent. "1?3?" doubt cases should be viewed separately from the other doubt cases as they probably signal double prominences within a word. The other cases might point to the use of ambiguous acoustic correlates.

Doubt cases (n = 623)	DFL	
1?2?	4.49%	(28)
1?3?	80.90%	(504)
2?3?	12.20%	(76)
1?2?3?	2.41%	(15)

Table 3: *Percentage (and counts) of perceived multiple prominences for the DFL group within all doubt cases.*

As previously mentioned (see 3.1.), the annotators also gave a high-certainty score in 75.88% of the cases. This is interesting as it shows that sometimes the annotators showed great confidence in their annotations whereas they did not perceive the same syllable as being the most prominent one.

3.1.3. Analysis of the correct vs. incorrect results

For this analysis all cases where consensus and canonical SP concur were labelled as "correct" and all others as "incorrect". A repeated measures ANOVA with Greenhouse-Geisser correction with canonical SP and word position in the sentence as within-subjects factors and L1 of the speakers as between-subjects factor was carried out on the percentage of (in)correct cases. As expected, an effect of L1 ((F(1, 28)) = 28.92, p < .001), but also of position of the word in sentence ((F(1.761, 49.308)) = 3.917, p < .005) was found. However, there is no effect of canonical SP for both speaker groups taken together ((F(1.80, 50.34)) = 3.04, n.s). The analysis also reveals an interaction between canonical SP and L1 ((F(1.80, 2.47)) = 3.38, p < .05), and position in sentence and L1 ((F(1.76, 68.96)) = 5.92, p < .001).

The same analysis was carried out per L1 group as an interaction between SP and L1 and word position in sentence and L1 had been found. For the DFL group an effect of SP ((F(1.70, 32.13)) = 10.33, p<.001) and position in sentence was found ((F(1.79, 33.95)) = 6.64, p< .05). Pairwise comparisons show that the effect of SP is caused by the difference between SP1 (14.3% correct, see Table 2) and SP3 (39.90%) but not with SP2 (25.8%). The same appears to be true for word position in sentence: 21.3% of cases in sentence-initial, 27.7% in sentence-medial, 31.1% in sentence-final position are correct. Figure 1 shows the percentage of correct results per canonical SP and position in sentence. There seems to be a trend towards more correct results when words are sentence-final, especially on the 3rd syllable (cf. our hypothesis). However, this result does not reach significance as the interaction between SP and position in sentence is not significant ((F(2.44, 45.39)) = 1.85, n.s.).

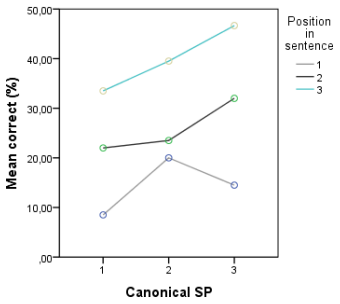


Figure 1: *Percentage correct cases per canonical SP and word position in sentence for the DFL data.*

For the control group, no effect of SP ((F(1.70, 15.31))=1.39, n.s.) or position in sentence ((F(1.68, 15.01))=0.08, n.s.) and no interaction between them ((F(4, 36))=0.41, n.s.) was found.

3.2. Acoustical realisation of prominent vs. non-prominent syllables

In order to study the prosodic correlates of perceived prominent syllables for the DL1 and DFL groups, we extracted several acoustic features of syllables. Inspired by the methodology presented in [23], four features were studied:

- Relative mean pitch: the difference of a syllable's mean pitch relative to the mean pitch of the word (in semitones);
- Pitch movement: intra-syllabic upwards or downwards movement (in semitones);
- Relative vowel duration: the ratio of the vocalic nucleus of a syllable, relative to the duration of the vocalic nuclei of the word;
- Relative peak intensity: the difference of a syllable's peak intensity in the vocalic nucleus relative to the mean intensity of the word (in dB).

These features are typically correlated with syllabic prominence. Syllables were included in the statistical analysis only when pitch could be detected and stylized by *Prosogram* in their corresponding word: in total 6891 syllables were analysed (2319 native, 4572 non-native; 1918 stressed, 4461 unstressed). The distributions of the four acoustic measures are

shown in Figure 2. Relative pitch and relative intensity follow a normal distribution. Relative vowel duration is positively skewed, while pitch movement follows a bimodal distribution, corresponding to falling and rising pitch.

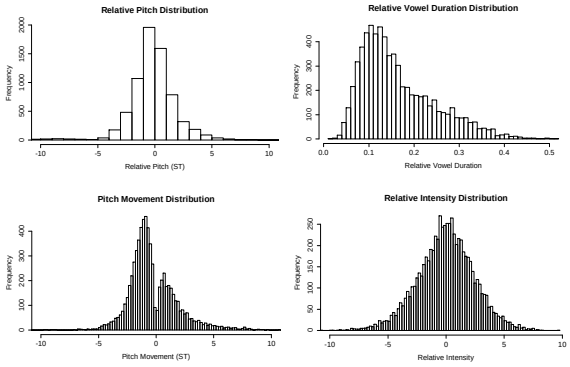


Figure 2: Distribution of the prosodic correlates of syllabic prominence under study (from left to right, top to bottom: relative pitch (ST), relative vowel duration (ratio), pitch movement (ST), relative intensity (dB).

A multivariate ANOVA was carried out on every acoustic measure; the fixed factors were L1 and presence/absence of prominence. Table 4 shows that the language groups make different use of duration ($p < 0.001$) and maximum pitch ($p < 0.05$), but use similar overall pitch and intensity strategies. There is a significant difference for every acoustic measure between prominent and non-prominent syllables ($p < 0.001$) as well as an interaction between both factors ($p < 0.001$, $p < 0.05$ for falling pitch movement).

Acoustic measure \ Factor	L1	Prominence	L1x Prominence
Relative syllable duration	**	**	**
Relative nucleus duration	**	**	**
Relative vowel duration	**	**	**
Relative pitch minimum	n.s.	**	**
Relative pitch maximum	*	**	**
Relative mean pitch	n.s.	**	**
Pitch movement, rising	n.s.	**	**
Pitch movement, falling	n.s.	**	*
Relative intensity	n.s.	**	**

Table 4: Main effects for L1, absence/presence of prominence and interaction between them on all acoustic measures (* $p<0.05$; ** $p<0.001$)

In order to assess the relative importance of each prosodic correlate for prominent and non-prominent syllables for both groups, we applied a binomial logistic regression model. A syllable was considered prominent or not (binary dependent variable) as long as there was a consensus (per syllable) between 4 or all 5 annotators. The acoustic measures were the model's predictors. In the DL1 group, relative mean pitch was found non-significant; all other predictors were significant with $p < 0.001$. Table 5 summarises the standardised beta coefficients and z-scores for each predictor for the two models. The results suggest that DL1 speakers signal prominence mainly through relative intensity, duration and rising pitch movement (in decreasing order of importance). DFL speakers on the other hand, use duration, then rising pitch movement and relative mean pitch. It is noteworthy that these were found to be the main prosodic correlates of syllabic prominence in French (along with succeeding pauses) [7].

Acoustic measure	DL1		DFL	
	β	z	β	z
Relative vowel duration	1.749	10.9	1.942	18.2
Pitch movement, rising	2.083	10.8	1.934	14.2
Pitch movement, falling	0.451	3.1	0.628	5.8
Relative mean pitch	-0.221	n.s.	-0.976	-8.4
Relative intensity	3.715	18.4	0.325	3.4

Table 5: Standardised β coefficients and z-scores for the DL1 and DFL logistic regression models predicting prominence.

4. Conclusion and perspectives

This paper investigated the realisation of Dutch stress by Belgian Francophone learners of Dutch. Our study showed low scores of correct stress position for the DFL group, pointing at their poor grasp of Dutch word stress position. On the whole, the DFL speakers relied on their final L1 pattern but mainly in canonical SP 3. This globally supports our hypothesis. There also seems to be a trend towards more correct 3rd-syllable stress in sentence-final position, but this result does not reach significance.

The manual annotation of perceived prominence in the corpus sometimes reached low agreement rates. In an attempt to explain this phenomenon, acoustical analyses were carried out. The analyses of variance seem to signal a different use of duration and maximum pitch by the language groups. The binomial logistic regression model points out that the DL1 group uses relative intensity, duration and rising pitch movement to signal prominence. The DFL group uses L1 accentuation strategies (duration, rising pitch movement and relative mean pitch). If annotators are more sensitive to different sets of acoustic correlates, this would explain the overall low consensus.

Further studies will focus on the comparison of the acoustic realisation of prominent syllables with high vs. low-certainty score cases. Speaker variability will also be investigated as it might also account for the lack of annotation agreement. Furthermore the analysis of the "doubt" cases (Table 3) should help us find evidence for multiple prominences within words.

While our study relied on a binary prominence label, it should be noted that research [e.g. 23, 24, 25] suggests that syllabic prominence is perceived as a gradual rather than a binary phenomenon. A manual annotation of relative prominence levels might give more insight into DFL stress production. Finally, it should be highlighted that the studied acoustic correlates are actually those of focus accent. While lexical stress should be studied in words out of focus [2], the main goal of our study was rather to compare stress position between DFL and DL1 speakers, stress and accent falling on the same syllable in Dutch. In an attempt to avoid total lack of prominence on the DFL stimuli, all words were put in focus. The acoustic correlates are used here as an attempt to explain perceptual differences between annotators and should not be considered as acoustic correlates of lexical stress.

5. Acknowledgments

The first two authors are supported by F.R.S.-FNRS grants. We would like to thank Dr. Thomas François for his advice on κ-coefficients.

6. References

[1] Rietveld, A. C. M., & Heuven, V. J., van. *Algemene Fonetiek*. Bussum: Coutinho, 2009.

[2] Sluijter, A. *Phonetic Correlates of Stress and Accent*. Leiden University, Den Haag: HILL, 1995.

[3] Lacheret-Dujour, A., & Beaugendre, F. *La prosodie du français*. Paris: CNRS Editions, 1999.

[4] Di Cristo, A. "Vers une modélisation de l'accentuation du français (seconde partie)", French Language Studies, 10, 27-44, 2000.

[5] Rasier, L. *Prosodie en vreemdetaal-verwerving. Accentdistributie in het Frans en Nederlands als vreemde taal*. Diss. Université catholique de Louvain, 2006.

[6] Vaissière, J. "Cross-linguistic prosodic transcription: French versus English", Problems and methods in experimental phonetics, In honour of the 70th anniversary of Prof. L.V. Bondarko. St.-Petersburg, St.-Petersburg State Univ, 147-164, 2002.

[7] Simon, A.C. *Phonologie et Prosodie*. Université catholique de Louvain, 2009.

[8] Hiligsmann, Ph. "Uitspraakvaardigheid van Franstalige leerders van het Nederlands: een ondergeschoven kind?", Handelingen, LII, 157-167, 1999.

[9] Caspers, J., & Santen, A., van. "Nederlands uit Franse en Chinese mond. Invloed van T1 op de plaatsing van klemtoon in Nederlands als tweede taal?", Nederlandse Taalkunde, 11 (4), 289-318, 2006.

[10] Heiderscheidt, S., & Hiligsmann, Ph. "De accentuering in de tussentaal van Franstalige leerders van het Nederlands", Leuvense Bijdragen, 89 (1/2), 17-131, 2000.

[11] Michaux, M.-C., Hiligsmann, Ph., & Rasier, L. "Het klemtoonpatroon in de tussentaal van Franstalige leerders van het Nederlands", XII. Internationaler Germanistenkongress Warschau 2010: Vielheit und Einheit der Germanistik weltweit, Warsaw, 321-332, 2012.

[12] Bui, A.V. *De woordklemtoon in de tussentaal van Franstalige leerders van het Nederlands: De beklemtoning van samenstellingen*, unpublished thesis, Université catholique de Louvain, 2012.

[13] Trommelen, M. & Zonneveld, W. *Klemtoon en metrische fonologie*. Muidergerg: Coutinho, 1989.

[14] Goldman, J.-Ph. "EasyAlign: an automatic phonetic alignment tool under Praat", Proceedings of Interspeech, 3233–3236, 2011.

[15] Bigi, B. & Hirst, D. "Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody", Proceedings of Speech Prosody '12, 19-22, 2012.

[16] Brognaux, S., Roekhaut, S., Drugman, T. & Beaufort, R. "Train&Align: A new online tool for automatic phonetic alignments", IEEE Workshop on Spoken Language Technologies, 410–415, 2012a,
Online: http://cental.fltr.ucl.ac.be/train_and_align/

[17] Brognaux, S., Roekhaut, S., Drugman, T. & Beaufort, R., "Automatic phone alignment. a comparison between speaker-independent models and models trained on the corpus to align", Lecture Notes in Computer Science, 7614, 300-311, 2012.

[18] Christodoulides, G. "Praaline: integrating tools for speech corpus research", Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 26-31 May 2014.

[19] Bartlett, S., Kondrak, G. & Cherry, C., "On the syllabification of phonemes", Proceedings of NAACL'09, 308-316, 2009.

[20] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model" in B. Bel & I. Marlien [Eds], Proceedings of Speech Prosody 2004, Nara, Japan, 23-26 March, 549-552, 2004.

[21] Geertzen, J. (2012). "Inter-Rater Agreement with multiple raters and variables", accessed on 23 December 2013, from https://mlnl.net/jg/software/ira/

[22] Landis, J. R., & Koch, G. G. "The measurement of observer agreement for categorical data", Biometrics, 33 (1), 159-174, 1977.

[23] Goldman, J.-Ph., Avanzi, M., Auchlin, A. & Simon, A.C., "A Continuous Prominence Score Based on Acoustic Features", Proceedings of Interspeech, Portland, Oregon, USA, 9-13 September, 2454-2457, 2012.

[24] Avanzi, M., Lacheret, A., Obin, N. & Victorri, B., "Vers une modélisation continue de la structure prosodique: le cas des proéminences", Journal of French Language Studies 21, 53–71, 2011.

[25] Arnold, D., Wagner, P. & Möbius, B., "Obtaining prominence judgments from naïve listeners. Influence of rating scales, linguistic levels and normalisation", Proceedings of Interspeech, Portland, Oregon, USA, 9-13 September, 2012.

Evaluation of bone-conducted ultrasonic hearing-aid regarding transmission of speaker gender and age information

Takayuki Kagomiya[1,2], *Seiji Nakangawa*[2]

[1]Center for Research Resources, National Institute for Japanese Language nad Linguistics, Japan
[2]Health Research Institute,
National Institute of Advanced Industrial Science and Technology (AIST), Japan

t-kagomiya@ninjal.ac.jp, s-nakagawa@aist.go.jp

Abstract

Human listeners can perceive speech signals in a voice-modulated ultrasonic carrier from a bone-conduction stimulator, even if the listeners are patients with sensorineural hearing loss. Considering this fact, we have been developing a bone-conducted ultrasonic hearing aid (BCUHA). The purpose of this study was to assess the usefulness of the BCUHA in transmission of speakers' physical attributes: gender and age. The evaluation used gender and age-identification experiments. The experiments were also conducted under air-conduction (AC) and cochlear implant simulator (CIsim) conditions. The results showed that: the BCUHA can well transmit speakers' gender information; the BCUHA can transmit speaker age information better than CIsim.

Index Terms: ultrasound, bone-conduction, hearing aid, paralinguistic information, speakers' attribute.

1. Introduction

We have developed a bone-conducted ultrasonic hearing aid (BCUHA) for sensorineural hearing-impaired patients [1]. A BCUHA consists of two components: an amplitude-modulated ultrasound processor and a bone-conduction vibrator.

Ultrasound is defined as sound with a frequency higher than the limitation of human perception (about 15 kHz). However, humans can perceive sound transmitted as ultrasound through a bone-conduction vibrator [bone-conducted ultrasound (BCU)] [2]. Moreover, if the ultrasound is amplitude modulated by speech sounds, the original speech sounds can be perceived in addition to the carrier sound by both normal-hearing (NH) and hearing-impaired listeners. We based the development of our BCUHA on these observations.

The cochlear implant (CI) was developed for and widely adopted by sensorineural hearing-impaired patients. A CI also consists of two components: speech signal processors and electrodes mounted in the cochlea. However, despite their widespread use, some problems with CIs have been reported. The biggest problem is that a CI requires surgical positioning, which causes irreversible damage to the cochlea. Another problem is that, because performance is limited by the number of electrodes, the CI transmits only partial or reduced information, not the entire sound. Nowadays, the number of CI electrodes is limited to between 12 and 24, and frequency resolution is limited according to the number of electrodes. This number may be sufficient for transmitting linguistic messages; however, it is reported that CI users have great difficulty perceiving music, speaker identity, emotional state, etc. [3, 4, 5, 6].

In contrast, the BCUHA does not require surgical fitting; users simply attach the bone-conduction vibrator with a hair band-like device (Figure 1). Furthermore, the BCUHA does not have the frequency limitations of CIs. However, in addition to speech signals, BCUHA listeners perceive high-frequency sound owing to the carrier signal [1]. Therefore, the carrier-originated sound may prevent clear perception of speech sounds. Consequently, the speech signal transmission performance of the BCUHA has been assessed using various techniques, such as using monosyllables [7] or word intelligibility scores [1]. These studies found that syllable articulation scores using BCU were over 60% [7] and that word intelligibility scores for high-familiarity words were over 85% [1]. The patterns of confusion in speech perception using BCU have many points in common with air conduction (AC) [7]. However, speech sounds convey not only linguistic messages but also indexical information about the speaker, such as gender, age, identity, and emotional state. As mentioned above, it is difficult for CI users to perceive such messages; thus, if the BCUHA performs better in this regard, it has a great advantage over the CI.

In this study, we focused on the ability to transmit speaker information. Previously, we compared the BCUHA with CI simulator (CIsim) based on speaker discrimination [8]. The results of this previous study showed that the speaker discrimination performance of the BCUHA is as good as that of the CI. However, this evaluation was performed using a discrimination task; the participants were simply requested to judge if the speaker

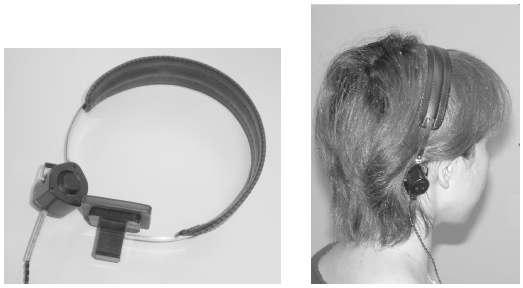


Figure 1: *Ceramic vibrator of the BCUHA attached to the mastoid with a hair band-like device*

of two sounds was "the same" or "different," and the result did not illustrate what types of speaker attributes were transmitted. Therefore, we assessed the ability of the BCUHA to transmit a speaker's physical attributes, including gender and age by conducting a series of listening experiments.

2. Methods

2.1. Stimuli

The experiments were designed as gender- and age-identification tasks, and stimuli were selected using the following procedures.

2.1.1. Speech material

To develop speaker gender- and age-identification tasks, we extracted speech material spoken by a large number of gender- and age-balanced speakers from "The Corpus of Spontaneous Japanese" (CSJ) [9]. From this corpus, we selected a phrase pronounced by various speakers: a short passage reading task "DNA." From the "DNA" task, the first phrase "di:-enu-e:" (DNA) was selected as the speech material. This phrase consists of voiced sounds and contains a nasal consonant. These types of sounds contribute to speaker identification [10]; thus, this phrase is expected to be suitable for the speaker attribute identification task.

Another reason for selecting this phrase was that it was spoken by 229 speakers in the CSJ. The speakers' genders were balanced, and the speakers' ages ranged from the low 20s to the high 60s. The numbers of speakers categorized by gender and age are listed in Table 1, in which "20L" represents low 20s, "20H" represents high 20s, etc.

Table 1: *Number of speakers categorized by gender and age*

	20L	20H	30L	30H	40L	40H	50L	50H	60L	60H	total
M	12	11	12	12	10	13	13	10	10	11	114
F	8	18	12	9	14	6	13	10	17	8	115
total	20	29	24	21	24	19	26	20	27	19	229

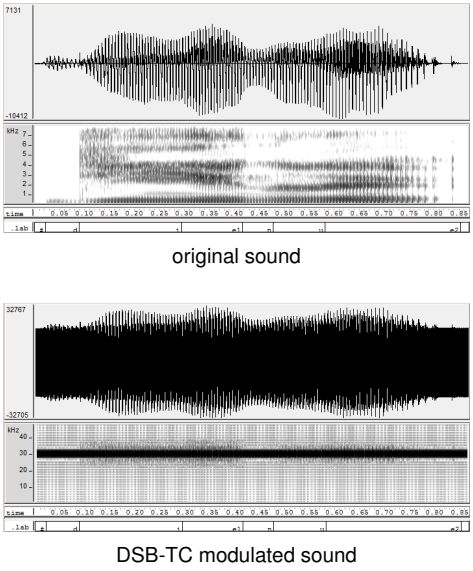


Figure 2: *DSB-TC amplitude-modulation*

2.2. Experiments

2.2.1. Bone-conducted ultrasound

The stimuli were converted to BCU stimuli in the form of amplitude-modulated 30 kHz sinusoid waves. A double-sideband transmitted-carrier (DSB-TC) amplitude-modulation method (Figure 2) was applied for this study because previous studies revealed it to be the best amplitude-modulation method for the BCUHA [1, 7]. With the DSB-TC method, the modulated speech signals $U(t)$ are represented by the following expression:

$$U(t) = [S(t) - S_{\min}] \times \sin(2\pi f_c t) \quad (1)$$

where $S(t)$ is the speech signal, $S_{\min}$ is the minimum amplitude of $S(t)$, and f_c is the carrier frequency (30 kHz).

The BCU stimuli were presented using a custom-made ceramic vibrator (Figure 1). Bone-conducted ultrasound can be perceived when it is applied to various parts of the body, and the mastoids are among the locations where such perception is high. Therefore, we applied the vibrator to the subject's left or right mastoid using a hair band-like device (Figure 1).

2.2.2. Cochlear implant simulator

To generate CI-simulated sounds, the Cochlear Implant Simulation (http://www.ugr.es/~atv/web_ci_SIM/en/ci_sim_en.htm) developed by the University of Granada was adopted in this study, and the software was config-

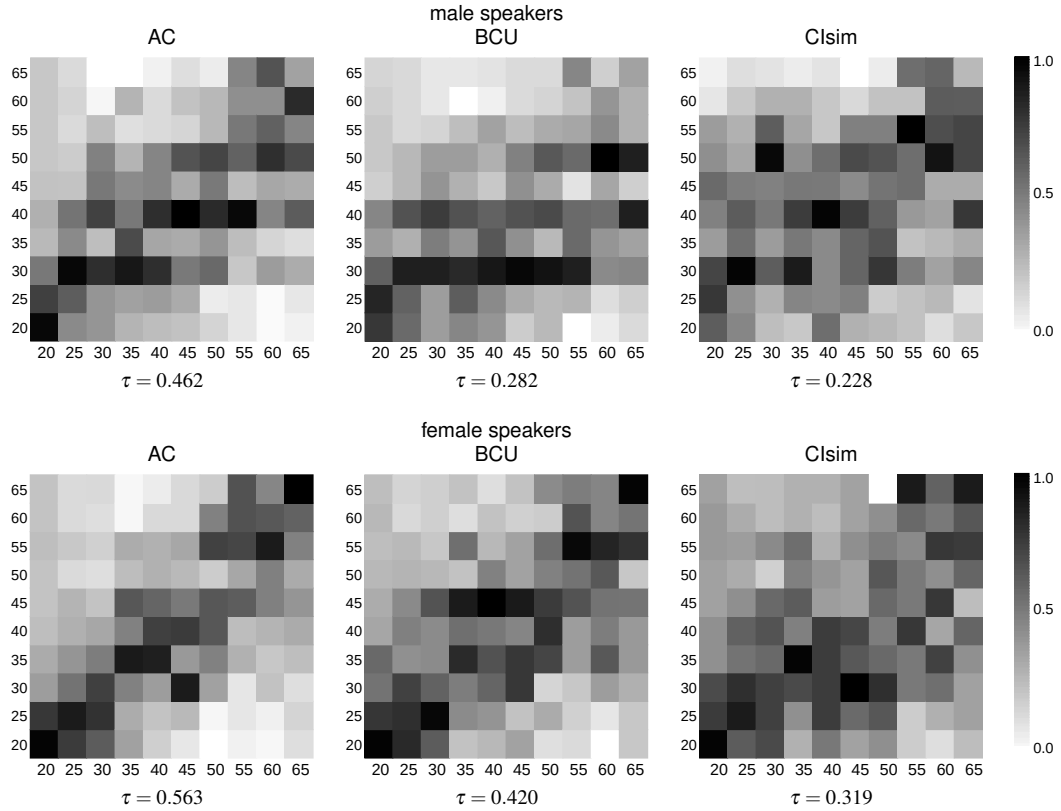


Figure 3: *Confusion ratios and Kendall's rank correlation coefficients (τ) obtained from the age-identification task*

Table 2: *Configuration of CI simulator*

length of implant	26.4 mm
number of channels	12
n-of-m	12 (CIS strategy)
interaction	2.4 mm
pulse rate	1515 pps/ch (18180 pps)

Table 3: *Confusion ratios from the gender-identification task*

	AC		BCU		CIsim	
	M	F	M	F	M	F
M	0.996	0.004	0.981	0.019	0.969	0.031
F	0.005	0.995	0.011	0.989	0.141	0.859

ured to simulate the MEDEL COMBI 40+ and TEMPO+ systems (see Table 2).

2.2.3. Participants

Seven native Japanese speakers (age: 19-41 years) with no reported hearing or speech defects participated in the experiments.

2.2.4. Procedures

All experiments were conducted in a soundproof chamber, and the sound levels of the stimuli were adjusted to the most comfortable level for each participant. For the AC and CIsim conditions, the sound stimuli were presented in a counterbalanced order through a set of headphones (Sennheiser HDA200). The participants were then requested to identify the speaker's gender (male or female) and age (10 levels listed in Table 1).

3. Results and Analysis

3.1. Speaker gender information

Table 3 lists the results of the gender identification task. The responses of all participants were pooled. The rows represent stimuli, and the columns show the responses. As shown in Table 3, speaker gender information was well transmitted in each condition. In the AC condition, the percentage of correct perceived ratios was higher than 99.5%, and in the BCU condition, it was higher than 98%. However, a small number of male-to-female errors (14%) were observed in CIsim condition.

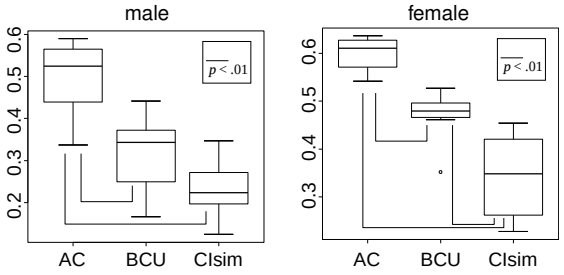


Figure 4: *Kendall's rank correlation coefficients by subjects*

3.2. Speaker age information

Figure 3 shows the confusion ratio and Kendall's rank correlation coefficients for speaker age. The responses of all participants were pooled. The columns represent stimuli, and the rows show the response. "20" represents low 20s, and "25" indicates the high 20s. Black-to-white contrast indicates the response ratio; darker cells indicate higher responses. The upper part shows the results of the male speaker condition, and the lower shows the female speaker condition.

From Figure 3 and the correlation coefficients, a large number of errors were observed. The rank correlation coefficients (τ) were lower than 0.57 (female voice) and 0.43 (male voice) even in the AC condition. In the BCUHA condition, the correlation coefficients were lower than 0.42 (female) and 0.29 (male) and 0.32 (female) and 0.23 (male) in the CIsim condition.

For each gender, the correlation coefficients were the best in the AC condition, then the BCUHA condition, and worst in the CIsim condition. To validate whether this tendency is statistically significant, correlation coefficients were calculated for each participant, and a series of ANOVA and post-hoc tests (multiple comparison with Holm's p adjustment method) was performed. The analyses were conducted for each speaker gender, and the results are shown in Figure 4. In the male speaker voice experiments, significant differences were observed between AC and BCU and between AC and CIsim ($p < 0.05$) and, in the female voice experiments, between all conditions ($p < 0.05$).

4. General Discussion

In each condition, speaker gender information was well transmitted. This result indicated that the BCUHA users can discriminate speaker gender as well as normal-hearing listeners can. This was also consistent with the results of the previous speaker discrimination tests [8], which showed that speaker errors were rarely observed in the AC, BCUHA, and CIsim conditions [8].

These high-accuracy gender identification results can be accounted for by good transmission of F0 information. It is well known that the F0 for an adult male voice is low and that for an adult female is medium. Moreover, BCUHA listeners can perceive the Japanese pitch accent [11] or prosodically salient paralinguistic information [12]. The results of these studies show that the BCUHA can transmit both local F0 modulations and global F0 range. Transmission of local F0 modulations allows pitch accents to be perceived, and global F0 range conveyance enables paralinguistic information, such as the speaker 's gender, to be obtained.

In contrast, BCUHA listeners have difficulty identifying the speaker's age. This result indicates that some voice timbre information is lost during BCU listening. However, this was also observed in the CIsim condition, and the correlation ratios between stimuli and perceived age by BCUHA listeners were superior to those of the CIsim listeners. Thus, the results of this study indicate that the BCUHA performs at least as well as the CI for speaker information transmission and outperforms the CI in some aspects.

5. Summary and Conclusions

The ability of the BCUHA to transmit speaker information was evaluated by conducting a series of speaker gender- and age-identification experiments. The results showed that the ability of BCUHA to make speaker gender judgments reaches the level of AC, whereas the BCUHA outperformed the CIsim for transmitting age information; however, the correlation coefficients between speaker age and perceived age using BCUHA were not sufficient. Further investigations are required to solve this problem.

6. Acknowledgements

This research was supported by a Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (24500675, 25280063) and the Funding Program for Next-Generation World-Leading Researchers provided by the Cabinet Office, Government of Japan.

7. References

[1] S. Nakagawa, Y. Okamoto, and Y. Fujisaka, "Development of a bone-conducted ultrasonic hearing aid for the profoundly sensorineural deaf," *Transactions of the Japanese Society for Medical and Biological Engineering : BME*, vol. 44, no. 1, pp. 184–189, 2006.

[2] M. L. Lenhardt, R. Skellett, P. Wang, and A. M. Clarke, "Human ultrasonic speech perception," *Science*, vol. 253, pp. 82–85, 1991.

[3] Q.-J. Fu, S. Chinchilla, and J. J. Galvin, "The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users," *Journal of the Association for Research in Otolaryngology*, vol. 5, no. 3, pp. 253–260, 2004.

[4] X. Luo, Q.-J. Fu, and J. J. Galvin, "Vocal emotion recognition with cochlear implants," in *Proceedings of Interspeech 2006*, 2006, pp. 1830–1833.

[5] J. Gonzalez and J. C. Oliver, "Gender and speaker identification as a function of the number of channels in spectrally reduced speech," *Journal of Acousitcal Society of America*, vol. 118, pp. 461–470, 2005.

[6] R. Müler, M. Ziese, and D. Rostalski, "Development of a speaker discrimination test for cochlear implant users based on the oldenburg logatome corpus," *Journal of Oto-Rhino-Laryngology, Head and Neck Surgery*, vol. 71, no. 1, pp. 14–20, 2009.

[7] Y. Okamoto, S. Nakagawa, K. Fujimoto, and M. Tonoike, "Intelligibility of bone-conducted ultrasonic speech," *Hearing Research*, vol. 208, pp. 107–113, 2005.

[8] T. Kagomiya and S. Nakagawa, "Evaluation of bone-conducted ultrasonic hearing-aid regarding transmission of speaker discrimination information," in *Proceedings of Interspeech 2011*, 2011, pp. 2209–2212.

[9] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, 2003, pp. 7–12.

[10] K. Amino and T. Osanai, "Speaker identification using japanese monosyllables and contributions of nasal consonants and vowels to identification accuracy," *Japanese Journal of Forensic Science and Technology*, vol. 18, no. 1, pp. 13–21, 2013.

[11] T. Kagomiya and S. Nakagawa, "Perception of Japanese prosodical phonemes through use of a bone-conducted ultrasonic hearing-aid," in *Proceedings of Speech Prosody 2012*, vol. 1, 2012, pp. 35–38.

[12] T. Kagomiya and S. Nakagawa, "An evaluation of bone-conducted ultrasonic hearing aid regarding perception of paralinguistic information," in *Proceedings of Speech Prosody 2010*, 2010, pp. 100 867:1–4.

Rhythm and Expression in *The Cat in the Hat*

Mara Breen[1], *Sarah Weidman*[1], *Katharine Guarino*[2]

[1] Department of Psychology and Education, Mount Holyoke College, South Hadley, MA, USA
[2] Department of Psychology, The University of Texas at Austin, Austin, TX, USA
mbreen@mtholyoke.edu, weidm22s@mtholyoke.edu, katieguarino22@gmail.com

Abstract

In recent years, there has been increasing interest in whether rhythmic interventions support young children's literacy development [1]. To begin to explore this connection, we assessed several aspects of rhythmicity and expressivity of productions of the notably rhythmic and rhyming children's book, *The Cat in The Hat* by Dr. Seuss. Participants subjectively rated either the rhythmicity or expressivity of speech taken from recordings of the book read aloud. These perceptual ratings were correlated with acoustic measures of rhythmicity and expressivity. Moreover, we observed a surprising lack of consistency between perceptual ratings of rhythmicity and expressivity. However, we observed a consistent relationship between the perceptual ratings of the first couplet of verses and the second, suggesting that readers showed self-entrainment in rhythmicity and expressivity between verse couplets. These findings can inform our investigation of the role of rhythm in literacy development and set the stage for further investigation into rhythmic entrainment across speakers.

Index Terms: rhythm, expression, production, child-directed speech

1. Introduction

1.1. Rhythm in speech perception

Speech researchers have long been interested in what factors contribute to isochronous speech in which stresses are perceived at regular intervals [2]. Although consistent acoustic measures of isochrony have been difficult to identify [3], there is consensus that perceptual isochrony is an important component of speech perception. For example, recent studies demonstrate that perceptual isochrony can help resolve ambiguity in speech perception [4,5]. Moreover, it is clear that babies attend to rhythm from a very early age [6], and can use rhythmic cues to learn words [7].

Many have observed that children's nursery rhymes are rhythmic and have consistent rhyming schemes. Moreover, exposure to rhyming texts has been shown to improve children's phonological awareness [8,9] and expressive vocabulary abilities [10]. However, little prior work has examined whether and how the rhythmic schemes of children's books improves children's literacy development.

There is reason to suspect that the rhythmicity of these texts supports reading acquisition, though this topic has received little direct investigation. For example, dyslexic children have been shown to be less sensitive to rhythm than normally reading children [11,12]. In addition, reading researchers have long advocated that rhythmic reading, including the use of an Interactive Metronome program, is an effective instruction tool [13]. One possibility, that we begin to explore here, is that rhythmic texts support self-entrainment [14], such that, within a verse, readers produce stresses at a consistent rate.

What can certainly not be denied is the fact that children like to encounter rhythm when reading, as evidenced by the popularity of books by Dr. Seuss and others. The goal of the current study is to explore the nature of the structure of rhythmic texts in order to begin to understand what aspects of their composition might support reading acquisition.

1.2. Rhythm in child-directed speech

A parallel line of inquiry in speech comprehension is determining the circumstances under which speech is most likely to be rhythmic. Recent investigations have demonstrated that multiple factors can contribute to the rhythmicity of speech, including the syntactic/semantic structure of the speech [15,16], the cognitive load of the speaker [17], and the audience to which the speech is directed.

It is this final factor that serves as the starting point for the current study. Recent work has demonstrated that the same speakers produce more rhythmic (even-timed) child-directed speech and singing than adult-directed speech [18,19] suggesting that readings of *The Cat in the Hat* will be good candidates for investigating speech rhythmicity.

1.3. Expression in child-directed speech

It is clear that expression plays a large role in child-directed speech. Acoustically, the greater perceived expressivity of child-directed speech is realized primarily by higher, and more variable, F0 than adult-directed speech [20]. Moreover, there is evidence that this expressivity facilitates speech comprehension. Examinations of child-directed speech indicate that exaggerated pitch peaks at the ends of utterances marking focused words may facilitate speech processing [21]. Moreover, previous research has demonstrated that infants are able to distinguish words from syllable sequences spanning word boundaries after exposure to infant-directed speech but not after exposure to adult-directed speech [22].

The current study further investigates the role of expression in children's literacy development, specifically examining how expressivity of child-directed speech may relate to the rhythmicity of child-directed speech.

1.4. Current Study

The objective of the current study was to investigate the rhythmic and expressive characteristics of child-directed speech; specifically, the book *The Cat in the Hat* by Dr. Seuss. The research questions we addressed were:

(1) Is there a consistent relationship between perceived rhythmicity and perceived expressivity in productions of *The Cat in the Hat?*

(2) Does the rhythmicity of the first couplet of a verse predict the rhythmicity of the second couplet?

(3) Does the expressivity of the first couplet of the verse predict the expressivity of the second couplet?

We could imagine two possible answers to question (1). On the one hand, if greater expressivity and greater rhythmicity are both predictive of better outcomes for audiences (of children), then it could be the case that these two factors will be correlated such that greater perceived rhythmicity corresponds to greater perceived expressivity. On the other hand, rhythmic variation is one marker of expressiveness in production. Therefore, we may observe a negative correlation between perceived rhythmicity and perceived expressivity of verses such that greater perceived rhythmicity corresponds to lower perceived expressivity.

Regarding (2) & (3), we were interested in whether there is rhythmic and expressive consistency within verses after accounting for within-speaker factors. That is, are there characteristics of the verses themselves that predict how expressive or rhythmic their productions will be?

2. Method

2.1. Participants

Participants were sixteen native speakers of American English (all female), ages 18-35. All received course credit for participating.

2.2. Materials

Participants all read "The Cat in the Hat" by Theodor Seuss Geisel (aka "Dr. Seuss") in its entirety. This text was written in an anapestic tetrameter such that every couplet (1A, 1B) consists of four anapestic (weak-weak-**strong**) feet. Two couplets form a verse (1). There are a total of 71 verses in the book.

(1) *A: "Put me **down**!" said the **fish**.*
*This is **no** fun at **all**!*
*B: Put me **down**!" said the **fish**.*
*"I do **not** wish to **fall**."*

2.3. Procedure

Participants were randomly assigned to read the hardcover book aloud in one of two environments: In a quiet room in the lab with no audience present; or to an audience of between three and five 5- and 6-yr-olds in a pre-kindergarten classroom. Both groups were instructed to read as naturally as possible. Their productions were recorded with Praat [23] through a Shure SM10A head-mounted microphone, connected to a Rolls Mini-Mic pre-amplifier. Productions were recorded with a sampling rate of 44100 Hz.

Initial analysis of acoustic measures demonstrated no significant differences between the productions recorded in the lab and those recorded in the classroom. This may be because our participants were all familiar with *The Cat in the Hat* and couldn't help but read it expressively and rhythmically regardless of whether there was an audience present. We will, therefore, report data from all productions together.

2.4. Acoustic Measurements

All productions were aligned with the book text in Praat using the Prosodylab-Aligner software [24] and then hand-corrected. Productions were segmented into verses (as in 1) and further divided into (A) and (B) couplets. All verses which contained a disfluency in either of the two couplets were excluded from further analysis. Of the 71 total verses in the book, we selected 18 for further analysis which met the following criteria: First, they were primarily anapestic tetrameter (though the first foot could be iambic); second, they appeared on four adjacent lines on one page of the book; third, they were fluently produced by at least 14 of the 16 speakers. These constraints resulted in a total of 252 verse productions (504 couplets).

These couplets were distributed across three lists which were balanced, as much as possible, across speakers and verses. The couplets that comprised each verse (e.g., (1A) & (1B)) did not appear on the same list, so that each couplet was rated by different listeners. The lists were comprised of 173, 166, and 165 couplets, respectively.

Using Praat, we extracted measures of duration, average F0, maximum F0, minimum F0, and average intensity from every word. For the current study, we looked only at measures of duration and mean F0.

2.5. Perceptual Evaluation

An additional 30 female participants were recruited to participate in a perceptual evaluation task. None had participated in the production experiment. They were randomly assigned to rate either the expressivity or the rhythmicity of the verse couplets.

Half of the participants (n= 15) rated their perception of the each couplet's rhythmicity on a scale of 1-4 where 1 was "very rhythmic" and 4 was "not at all rhythmic." Instructions were adapted from [5] such that raters were asked to determine how well they could find a steady beat for the couplet. Each of these 15 participants was assigned to one of the three lists, so that five participants rated each couplet. They were first presented with 8 practice couplets; four were categorized by the authors as "very rhythmic"; four were categorized as "not at all rhythmic." We observed agreement on these ratings which was significantly greater than chance, Fleiss' $\kappa = .09$, $z = 10.5$, $p = 0$ [25], indicating that raters were sensitive to couplet rhythm. Ratings from all five raters were averaged, resulting in an average Rhythmicity score for each couplet.

Half of the participants (n=15) rated their perception of the each couplet's expressivity on a scale of 1-4 where 1 was "very expressive" and 4 was "not at all expressive." Each of these 15 participants was assigned to one of the three lists, so that five participants rated each couplet. They were first presented with 8 practice couplets; four were categorized by the authors as "very expressive"; four were categorized as "not at all expressive." We observed agreement on these ratings which was significantly greater than chance, Fleiss' $\kappa = .29$, $z = 34.5$, $p = 0$, indicating that raters were sensitive to couplet expressivity. Ratings from all five raters were averaged, resulting in an Expressivity score for each couplet.

3. Results

3.1. Acoustic and perceptual ratings

Before investigating our research questions, we measured the consistency between acoustic and perceptual ratings of expressivity and rhythmicity.

3.1.1. Acoustic and perceptual measures of rhythmicity

In order to objectively quantify the rhythmicity of productions, we identified the timing of the onsets of strong syllables (**bolded** in (1)): For each couplet, we computed the standard deviation of the latencies (SD_Latency_Strong) between the onset of strong syllables 1 and 2 (***down*** and ***fish***), syllables 2 and 3 (***fish*** and ***no***), and syllables 3 and 4 (***no*** and ***all***). SD Latency ranged from 1 ms to 850 ms. This measure was highly correlated with the average Rhythmicity rating, R = .48, p < .0001, such that higher variability in the timing onsets of strong syllables was associated with lower Rhythmicity ratings (Figure 1). Recall that a lower number indicates a higher rating of Expressivity. This result suggests that timing variability was one factor that raters used in their Rhythmicity judgments.

3.1.2. Acoustic and perceptual measures of expressivity

As intonational variability has been identified as a strong marker of expressivity in speech [20], and child-directed speech exhibits greater pitch variability [21], we computed an acoustic measure of expressivity based on pitch. For each couplet (e.g., 1A, 1B), we computed the standard deviation of the mean F0 (SD_F0_Word) for each word. This measure was highly inversely correlated with the average Expressivity rating, R = -0.18, p < .001 (Figure 2). Recall that a lower number indicates a higher rating of Expressivity. This result demonstrates that F0 variability was one factor that raters used in their Expressivity judgments.

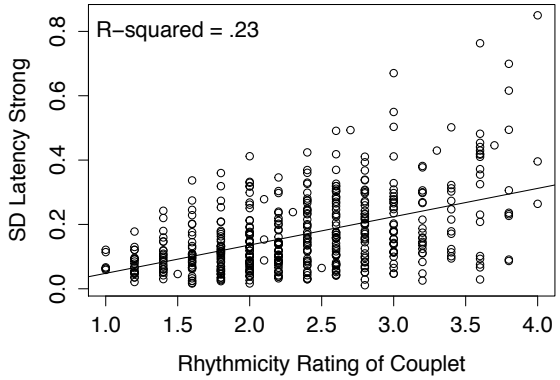


Figure 1: *Correlation of average perceived rhythmicity rating of a couplet and the standard deviation of the latency of onsets between the strong syllables of the couplet.*

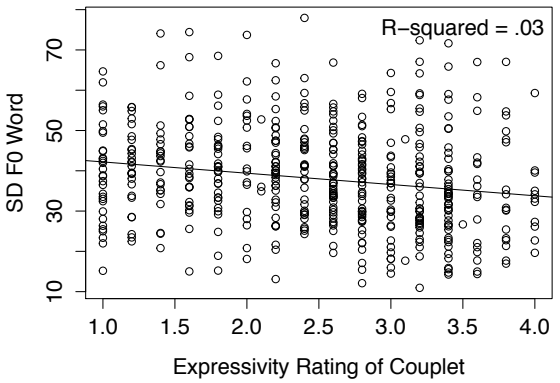


Figure 2: *Correlation of average perceived expressivity rating of a couplet and the standard deviation of the mean F0 of each word in the couplet.*

3.2. Relationship between Expressivity and Rhythmicity

In order to determine whether there is a consistent relation ship between rhythmicity ratings and expressivity ratings, we conducted a correlational analysis comparing the Expressivity and Rhythmicity scores for each couplet. Pearson's correlational coefficient was R = .03, which did not reach significance, p = .48, indicating no systematic relationship between perceived expressivity and perceived rhythmicity.

3.3. Intraverse Expressivity and Rhythmicity

In order to investigate the consistency between the Rhythmicity and Expressivity scores of the first and second couplets of a single verse, we analyzed these scores separately on a single-couplet basis using a linear mixed-effects model implemented in the languageR package [26] implemented in the R statistical programming language (R Core Development Team, 2012).

We computed models in which we predicted an A couplet's Rhythmicity/Expressivity score from one fixed effect (the B couplet's Rhythmicity/Expressivity score) and two random effects: The individual reader and the verse that the couplets comprised. In this way, we could see whether the Rhythmicity/Expressivity of one half of a couplet predicted the Rhythmicity/Expressivity of the second half after accounting for differences across speakers and items. We estimated the significance level of the fixed effect using Markov Chain Monte Carlo sampling [27].

Fixed Effects	Est.	SE	t	pMCMC
Intercept	2.045	0.17	12.06	<.001
Rhythmicity of couplet A	0.16	0.07	2.35	<.05

Table 1. *Parameter estimates of mixed-effects model predicting Rhythmicity of couplet B.*

Fixed Effects	Est.	SE	t	pMCMC
Intercept	1.45	0.18	8.08	<.001
Expressivity of couplet A	0.43	0.06	7.61	<.001

Table 2. *Parameter estimates of mixed-effects model predicting Expressivity of couplet B.*

Across speakers and verses, the Rhythmicity rating of the first couplet of a verse was predictive of the rating of the second couplet, $t = 2.35$, $p < .05$ (Table 1). The simple correlation between the Rhythmicity ratings of the first and second couplets are presented in Figure 3. This result demonstrates that, across speakers and items, the rhythmicity of the first couplet was predictive of that of the second couplet.

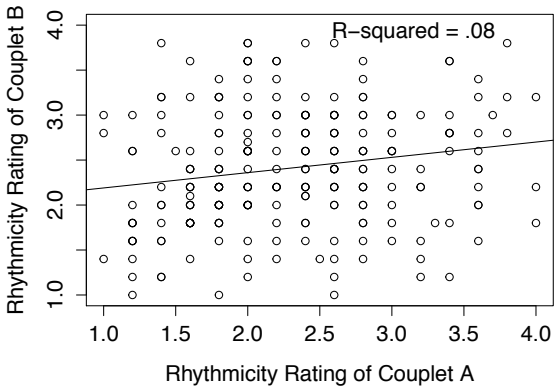


Figure 3: *Correlation of average perceived rhythmicity rating of the first and second couplet of verses.*

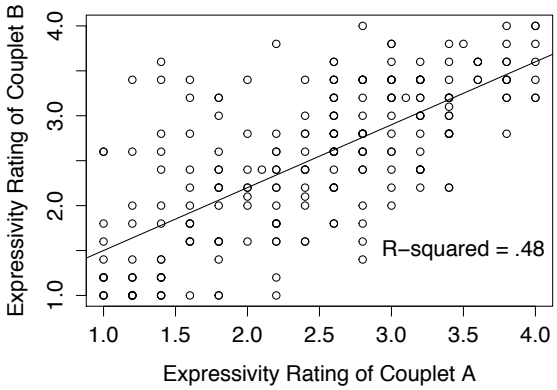


Figure 4: *Correlation of average perceived expressivity ratings of the first and second couplet of verses.*

Across speakers and verses, the Expressivity rating of the first couplet of a verse was predictive of the rating of the second couplet, $t = 7.61$, $p < .001$ (Table 2). The simple correlation between the Expressivity ratings of the first and second couplets is presented in Figure 4. This result demonstrates that, across speakers and items, the expressivity of the first couplet was predictive of that of the second couplet.

4. Discussion

In order to begin to explore the features of rhythmic children's books that support literacy development, we have constructed a spoken corpus of sixteen productions of *The Cat in the Hat* by Dr. Seuss. We gathered perceptual ratings of both the rhythmicity and expressivity of many of the verses from the text, and identified acoustic correlates of these perceptual ratings. Specifically, we found that rhythmicity ratings were correlated with variability in the timing of the onsets of the strong syllables of couplets and that expressivity ratings were correlated with F0 variability across words. These results demonstrate that naïve raters can successfully rate perceived expression and rhythm in speech [5].

We found no correlation between the rhythmicity and expressivity ratings of the individual couplets. This result demonstrates that expressive speech is not inherently any more or less rhythmic than non-expressive speech. Furthermore, it shows that participants are able to rate speech rhythmicity independently of speech expressivity, disproving an original concern of the study that participants would not be able to differentiate these two features.

Finally, we observed intraverse relationships for both the rhythmicity and expressivity ratings. That is, even after accounting for differences across individual speakers and individual verses, it is still the case that the rhythmicity or expressivity of the first couplet of a verse is significant predictor of the rhythmicity or expressivity of the second couplet. These results are particularly striking as the rhythmicity and expressivity of the first and second couplet of a verse were always rated by different sets of participants. This result suggests that readers of rhythmic texts are consistent in their productions of whole verses, suggesting that readers engage in self-entrainment in both their rhythmicity and expressivity.

It may be the case that, like the adults in the current study, children demonstrate self-entrainment when reading rhythmic prose, and that this process helps improve their rapid auditory processing, leading to improved reading fluency [13]. Follow-up studies will investigate (1) whether children reading *The Cat in the Hat* also show self-entrainment within verses and (2) whether reading overtly rhythmic prose improves children's overall fluency.

5. Conclusions

We have generated a corpus of child-directed speech that varies in its rhythmicity and expressivity. Moreover, we have identified acoustic correlates of these perceptual measures. We have demonstrated that there is no consistency between the rhythmicity and expressivity ratings of individual couplets. However, there is strong intraverse consistency for both rhythmicity and expressivity.

6. Acknowledgements

The authors would like to thank Suyin Taunton and Julia Karron for help with sound file alignment and Molly Morgan and Cam Vilain for assistance with running participants.

7. References

[1] Bhide, A., Power, A. & Goswami, U. (2013). A rhythmic musical intervention for poor readers: A comparison of efficacy with a letter-based intervention. *Mind, Brain, and Education, 7,* 113-123.

[2] Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5,* 253-263.

[3] Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm**.** *Journal of Phonetics, 40,* 351-373.

[4] Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica 66*, 95-112.

[5] Dilley, L. & McAuley, J.D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59, 294-311.*

[6] Nazzi, T. & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41,* 233-243.
[7] Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38,* 465-494.
[8] Dunst, C.J., Meter, D., Hamby, D.W. (2011). Relationship between young children's nursery rhyme experiences and knowledge and phonological and print-related abilities. *CELL reviews, 4(1),* 1-12.
[9] Hayes, D. S. (2001). Young children's phonological sensitivity after exposure to a rhyming or nonrhyming story. *Journal of Genetic Psychology, 162,* 253-259.
[10] Stadler, M., Watson, M., Skahan, S (2007). Rhyming and vocabulary: Effects of lexical restructuring. *Communication Disorders Quarterly, 28(4),* 197-205.
[11] Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., et al. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proc. Natl. Acad. Sci. USA, 6,* 10911-10916.
[12] Thomson, J. M., & Goswami, U. (2008). Rhythmic processing in children with developmental dyslexia: Auditory and motor rhythms link to reading and spelling. *Journal Of Physiology, 102(1-3),* 120-129.
[13] Ritter, M., Colson, K. A., & Park, J. (2013). Reading Intervention Using Interactive Metronome in Children With Language and Reading Impairment: A Preliminary Investigation. *Communication Disorders Quarterly, 34(2),* 106-119.
[14] Port, R. F., Tajima, K., & Cummins, F. (1996). Self-entrainment in animal behavior and human speech. *Online proceedings of the 1996 Midwest artificial intelligence and cognitive science conference*, Indiana University, Bloomington, IN.
[15] Cummins, F. (2003): Rhythmic grouping in word lists: competing roles of syllables, words and stress feet. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain,* 325-328.
[16] Cummins, F. (2005): Interval timing in spoken lists of words. *Music Perception, 22,* 497-508.
[17] Dilley, L., Wallace, J., & Heffner, C. (2012). Perceptual isochrony and fluency in speech by normal talkers under varying task demands. *Prosodies: Context, Function, and Communication,* O. Niebuhr and H. Pfitzinger (Eds.), *Language, Context, and Cognition series,* Berlin/New York: Walter deGruyter, pp. 237-258.
[18] Payne, E., Post, B., Astruc, L., Prieto, P. & Vanrell, M. (2009). Rhythmic modification in child directed speech. *Oxford University Working Papers in Linguistics, Philology & Phonetics, 12,* 123–144.
[19] Nakata, T., & Trehub, S. E. (2011). Expressive timing and dynamics in infant-directed and non-infant-directed singing. *Psychomusicology: Music, Mind & Brain, 21,* 45-53.
[20] Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40,* 227–256.
[21] Fernald, A & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology, 27,* 209-221.
[22] Thiessen, E., Hill, E., & Saffran, J. (2005) Infant-directed speech facilitates word segmentation, *Infancy, 7,* 5-71.
[23] Boersma, P. & Weenink, D. (2002). Praat, a system for doing phonetics by computer. Software and manual available online at: <http://www.praat.org>.
[24] Gorman, K., Howell, J. & Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech, Proceedings of Acoustics Week in Canada, Quebec City.
[25] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, Vol. 76, No. 5 pp. 378–382.
[26] Baayen, R. H. (2008). Analyzing linguistic data: a practical introduction to statistics using R. Cambridge: Cambridge University Press.
[27] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

Lengthened Consonants are Interpreted as Word-Initial

Laurence White[1], Sven Mattys[2], Linda Steffansdottir, Victoria Jones

[1]School of Psychology, Plymouth University, Plymouth, UK
[2]Department of Psychology, University of York, York, UK

laurence.white@plymouth.ac.uk, sven.mattys@york.ac.uk

Abstract

Prosody facilitates listeners' segmentation of the speech stream into a sequence of words and phrases. With regard to speech timing, vowel lengthening is interpreted as a cue to an upcoming boundary, in accordance with the iambic-trochaic law. However, the impact of consonant lengthening on segmentation, in the absence of other boundary cues, has not been tested.

In a series of artificial language learning experiments, we examined how durational variation affects listeners' extraction of novel trisyllables defined by transition probabilities. In line with previous research, syllables containing lengthened vowels were interpreted by listeners as word-final. However, syllables with lengthened onset consonants were interpreted as word-initial. Thus, the structural interpretation of durational variation depends upon localization: longer vowels cue a following boundary; longer consonants cue a preceding boundary.

Index terms: speech timing, speech segmentation

1. Introduction

Variations in suprasegmental dimensions – pitch, duration, loudness – are consistently associated with speech structure at prosodic heads and edges [1]. Firstly, the heads of prosodic domains – stressed syllables and accented words – are more prominent through a combination of higher pitch, greater duration and greater loudness, although the relative contribution of these dimensions is language-specific. Secondly, boundaries between words and between phrases are associated with intonational and durational variation. Boundary-adjacent intonational contours vary between languages [2], but upcoming boundaries may be universally associated with segmental lengthening [1]. Indeed, the slowing of articulation as boundaries approach has been associated with a non-linguistic principle of deceleration at the end of motor sequences [3,4].

It is well established that listeners use suprasegmental variations to segment speech into words and phrases [5,6,7]. Considering specifically speech timing, lengthened vowels are interpreted as word-final in artificial language streams [8]. Similarly, with natural language stimuli, long stressed syllables were more likely to be interpreted as monosyllabic words rather than the start of disyllables (e.g., *ham* vs *hamster* [9]). Additionally, greater magnitude of syllable lengthening is associated by listeners with higher-level phrase or utterance boundaries [7].

Such results are in line with the iambic-trochaic law [10], which proposes that the interpretation of prosodic salience depends on its phonetic realization: in particular, sounds made salient through greater loudness are recognized, other things being equal, as sequence-initial, whilst sounds made salient through lengthening are perceived as final.

Despite English and French differing markedly in the distribution and realization of phrasal prominence, support for the iambic-trochaic law was found with both English and French listeners, and with speech and non-speech sounds [11]. Expanding salience cues to consider also pitch, high-low disyllables were better recalled than low-high; thus, pitch salience, like loudness, is interpreted as sequence-initial [12]. However, considering durational contrast in the same study, recall of disyllabic sequences was better when second syllables had longer vowels compared to when both vowels had similar durations or when first syllables had the longer vowel [12].

For native English-speaking listeners, lengthened syllables promote segmentation of artificial language streams when word-final but not when word-initial [8,13]. In particular, lengthening of vowels in word-initial syllables does not facilitate segmentation for English listeners, despite English having predominantly word-initial stress, associated with segmental lengthening within the syllable, together with pitch excursion, increased loudness and other cues [14]. This raises the question of whether lengthening can only ever serve as a cue to an upcoming boundary rather than a preceding boundary.

Listeners' relative weighting of durational *vs* other cues to prosodic structure (e.g., lexical, segmental, suprasegmental) may be modulated according to language-specific patterns of occurrence [15]. Compared to Dutch, for example, vowel duration in English may be less important for signaling stress than is vowel quality [16, 17]. Additionally, the distribution of lengthening effects within words may serve to disambiguate their different structural interpretations [18, 19]. For example, analysis of a corpus of English speech showed that the distinct patterns of lengthening within words due to lexical stress and word-/phrase-finality appear sufficient to allow listeners to distinguish the two structural interpretations [20]. If so, lengthening could be disambiguated and reliably used as a pre-boundary or stress cue by listeners, even in the absence of additional segmental and suprasegmental cues.

Of course, a full account of the distribution of durational effects associated with prosodic structure must also include consonantal lengthening. In particular, lengthening of consonants in word-initial position is consistently observed in several studied languages. For English, syllable onset consonants are substantially longer when uttered in word-initial position than word-

medially [21], an effect subsequently observed in French, Korean, and Taiwanese [22]. Whilst multiple consonants within the onset may be lengthened, the durational effect does not extend to the vowel nucleus of that syllable [18, 21]. As with lengthening of domain-final vowels, the magnitude of the consonant-lengthening effect increases at higher prosodic boundaries [23, 24], although consonants may be as short in absolute utterance-initial position as when word-medial, perhaps in part because the termination of silence serves as an unambiguous cue to prosodic structure [18].

Several studies have investigated the impact of consonant length, in conjunction with other durational and segmental cues, on word -level segmentation [5, 25, 26, 27]. Typically, these studies have manipulated a within-task contrast between consonants in initial and medial/final position. For example, in segmentally ambiguous phrases like *two lips* vs *tulips* (near-homophonous in American English), a longer consonant in word-initial position (/l/) encouraged cross-modal priming of *lips* [25]. Also for English, word-initial consonant lengthening, together with word-final vowel lengthening and other naturally occurring cues to prosodic boundaries, affected the interpretation of ambiguous sequences such as *paper* vs *pay per* in adults and infants as young as ten months [5, 6].

In Dutch, listeners' interpretation of segmentally ambiguous sequences like *diep in* vs *die pin* was affected by duration of the pivotal consonant, which tended to be interpreted as word-initial when relatively long [26]. Consonant duration affected Italian listeners' judgements regarding both lexical segmentation and identification of geminates *vs* singletons [27], whilst French listeners interpreted longer consonants as more likely to be word-initial than in liaison context (e.g., *dernier oignon* vs *dernier rognon*, [28]). There is also evidence that phrase-initial lengthening and articulatory strengthening affect listeners' interpretation of the structure of ambiguous phrases [29].

All of the foregoing studies suggest that lengthened consonants tend to be interpreted as word-initial by listeners. However, all used natural speech – sometimes resynthesized to manipulate segment durations – with multiple potential cues to word boundaries. In particular, segmental cues, such as boundary-related allophonic variations, and other suprasegmental cues, including lengthening of word-final vowels, were also available to listeners [26]. Furthermore, participants' awareness of the implicit contrast between the two interpretations of near-homophonous sequences (e.g., *two lips* vs *tulips*) might modulate their use of segmentation cues relative to when there is only one lexical solution available.

To address these confounds, we used an artificial language learning paradigm to focus on lengthening of consonants and lengthening of vowels, both separately and together. Listeners have consistently been shown to be able to learn and subsequently recall novel words from a nonsense speech stream when the syllable-to-syllable transition probabilities within words are higher than those between words [8]. Using such a paradigm, with artificial speech streams created through diphone synthesis, we obviate the need to use near-homophonous sequences from natural languages, and eliminate the presence of other potential cues to word boundaries. This allows us to focus precisely on the question: does longer duration increase the tendency for consonants to be interpreted as word-initial?

2. Experiment 1

2.1. Method

Three durational manipulations of two artificial languages were used to determine the impact of consonantal lengthening on segmentation, and thereby on subsequent recall, of statistically-defined words. We predicted that words should be better recalled when word-initial consonants are lengthened during language exposure, relative to when all consonants had the same duration or when word-medial consonants were lengthened.

2.1.1. Participants

We tested 120 native British English speakers, with no reported speech or hearing problems. They were randomly allocated to the three duration conditions (40 in each condition). Within each duration condition, 20 participants were allocated to Stream 1 and 20 to Stream 2. All participants received a small honorarium or course credit for their participation.

2.1.2. Materials

We prepared two different artificial languages, similar to those used in earlier studies [30], each comprising four trisyllabic words (C1V1-C2V2-C3V3).

Stream 1 words: *pabiku, golatu, tinudo, daropi*

Stream 2 words: *tudaro, bikuti, golatu, nudopa*

Six-minute streams containing these words in random sequence were generated using the *en1* male British English voice in the diphone synthesizer MBROLA [31]. Fundamental frequency was a constant 120Hz. To eliminate the strong segmentation cue of hearing silence at the beginning and end of the stream, the streams were faded in and faded out with five-second ramps.

As each word could be followed by any of the other three words, but not by itself, the transition probability between words was always 1/3. As each syllable only occurred once within the language, the transition probability between within-word syllables was always 1.

Total trisyllabic word duration was kept constant at 720ms, whilst the duration of individual segments was manipulated to generate three "lengthening" conditions.

Flat: All segments – vowels and consonants – were 120ms.

C1: The onset consonant of the first syllable of each word (*pabiku* etc.) was 170ms *vs* 110ms for all other segments.

C2: The onset consonant of the second syllable of each word (*pabiku* etc.) was 170ms *vs* 110ms for all other segments.

In the test phase, following exposure to the six-minute stream, isolated words and foils were played to participants. Test-phase foils were constituted of the syllables of the language, either part-words derived from the end of one word and start of another (e.g., Stream 1: *bikuti* from *pabiku tinudo*) or non-words, syllable strings that never occurred in the language (e.g., *tipala*). The words in Stream 1 were part-words in Stream 2 and *vice*

versa. Words and foils for the test phase were synthesized with flat durational profiles (all segments 120ms) in all three conditions.

2.1.3. Procedure

Participants were told they would hear an artificial language through headphones for six minutes, and that their task was to listen and try to discover the words in the language. After the exposure phase, they were given test phase instructions. In the test phase, they heard 24 pairs of trisyllabic strings, where one string was a word in the language stream just heard and the other string was a part-word or non-word. The two trisyllabic strings were separated by 500ms. For each pair, participants had to press the left shift key on a computer keyboard if the artificial language word was the first string of the pair, and the right shift key if it was the second string.

2.1.4. Statistical analysis

All analyses were carried out on the raw response data – "correct" or "incorrect" – using mixed-effects logistic regression models, including the random factors of subjects and trials (*lmer* package in *R*, [32]). Models were compared using log-likelihood χ^2 tests.

2.2. Results and discussion

Mean correct responses by lengthening condition are shown in Figure 1. In the Flat condition, mean correct was 67%, $z = 5.07$, $p < .0001$, replicating previous findings that listeners can recognize words defined by transition probabilities in novel streams of syllables [8]. Above chance performance was also found in the other lengthening conditions: C1 – 73%, $z = 6.12$, $p < .0001$; C2 – 63%, $z = 4.29$, $p < .0001$.

A logistic regression including fixed factors of Lengthening (*Flat vs C1 vs C2*) and Stream (1 *vs* 2) found a main effect of Lengthening, $\chi^2(2) = 11.59$, $p < .005$. There was also a main effect of Stream, $\chi^2(1) = 4.51$, $p < .05$, with words from Stream 1 recalled better than those from Stream 2. There was no interaction between Lengthening and Stream, $\chi^2(5) = 0.11$, $p = .95$. Thus, the advantage of Stream 1 over Stream 2 was consistent across the three lengthening conditions, and so further pairwise analyses were collapsed across streams.

Lengthening of the consonant in the first syllable (C1) improved performance compared to lengthening of the consonant in the second syllable (C2), $\chi^2(1) = 9.93$, $p < .005$, and compared to the Flat condition, $\chi^2(1) = 4.20$, $p < .05$. There was no difference between *C2* vs *Flat*, $\chi^2(1) = 2.15$, $p = .14$. These results indicate that segmentation of the artificial language was promoted by localized lengthening of the word-initial consonant. Thus, consonantal lengthening appeared to cue listeners to the presence of an immediately preceding boundary.

Lengthening of a vowel in a similar artificial language stream has been shown to act as a cue to a following boundary [8]. This suggests a functional difference in listeners' interpretation of vowel and consonant lengthening. An alternative hypothesis is that longer syllables – whether through greater vowel or consonant duration – tend to be perceived as word-edges, which could be either initial or final. This view is not encouraged by findings that vowel lengthening in word-initial syllables failed to facilitate segmentation relative to no lengthening [8]. However, in that experiment, initial syllable vowels were only lengthened in half of the six artificial words. In order to confidently assert our interpretation – that consonant lengthening, in contrast with vowel lengthening, is a cue to a preceding boundary – we attempted a more direct replication of [8] with our artificial language materials, testing the effect on segmentation of lengthening the first syllable vowel in every trisyllabic word.

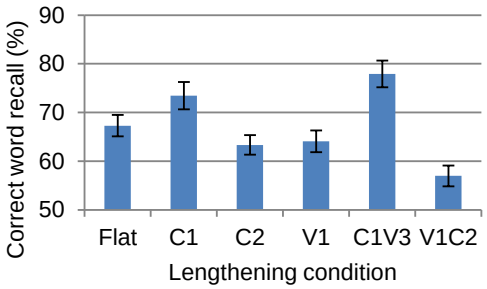


Figure 1: *Mean correct responses and standard errors: Exp1 – Flat, C1, C2; Exp 2 – V1; Exp 3 – C1V3, V1C2.*

3. Experiment 2

3.1. Method

The participants and experimental procedure were as for Experiment 1. However, in Experiment 2, participants heard the artificial language streams with the first vowel of each word lengthened: thus, the underlined vowel in *pabiku* etc. was 170ms vs 110ms for all other segments. This *V1* condition was implemented for both artificial language streams, with 20 participants hearing each one. As before, after exposure to the streams, participants heard 24 pairs of trisyllabic strings – words and foils – with all segments having 120ms duration, and had to choose which string within a pair belonged to the artificial language.

3.2. Results and discussion

Overall mean correct word recognition in the *V1* condition was 64%, reliably above chance, $z = 4.66$, $p < .0001$. To test the hypothesis regarding the localization of durational segmentation cues, the important comparisons are with the *Flat* and *C1* conditions in Experiment 1. Figure 1 shows the mean correct responses for the three critical conditions.

Collapsing, as before, across the two artificial language streams, there was no difference in recognition between the *Flat* and *V1* conditions, $\chi^2(1) = 1.15$, $p = .28$, replicating previous findings that lengthening of the vowel in a word-initial syllable does not serve as a cue to a preceding boundary for English listeners, despite the prevalence of word-initial stress in English [8, 13].

Performance on the *C1* condition was reliably better than the *V1* condition, $\chi^2(1) = 7.57$, $p < .01$. This supports the hypothesis that localization of lengthening is important for segmentation: a lengthened consonant cues a preceding boundary; a lengthened vowel cues a following boundary. In Experiment 3, to explore the power of such cues further, we tested the efficacy of vowel and consonant lengthening in combination. In particular, we examined whether a lengthened vowel

immediately followed by a lengthened consonant was a strong cue to an intervening boundary. We contrasted two cases, one where the lengthened consonant-vowel sequences coincided with the location of boundaries indicated by syllable transition probabilities and the other where the durational and statistical cues conflicted.

4. Experiment 3

4.1. Method

The participants and experimental procedure were as for Experiment 1, but here the duration of the artificial language streams was manipulated in two new conditions. In condition *C1V3*, the first consonant and the final vowel of each word (e.g., *pabiku*) were 160ms, *vs* 100ms for all other segments. In condition *V1C2*, the vowel of the first syllable and the consonant of the second syllable (e.g., *pabiku*) were 160ms, *vs* 100ms for all other segments. This was effectively a composite of the V1 and C2 conditions. Note that the lengthened segments were 160ms and the others 100ms, in contrast with 170ms and 110ms in the other experiments: this was to preserve total word duration at 720ms in all conditions across the three experiments.

4.2. Results and discussion

As shown in Figure 1, performance was reliably above chance in the *C1V3* condition, 78%, $z = 8.15$, $p < .0001$, and the *V1C2* condition, 57%, $z = 2.29$, $p < .05$. However, performance was significantly better in the *C1V3* condition, where the lengthened vowel and consonant straddled a statistically-defined word boundary, $\chi^2(1) = 31.69$, $p < .001$.

Comparison with the earlier experiments showed that performance on *C1V3* was no better than on *C1*, $\chi^2(1) = 1.10$, $p = 0.29$. However, performance on *C1V3* was better than on all other conditions ($p < .001$ in all cases). This may be due to intrinsic performance limitations on this type of language learning task, given the memory component combined with the repeated exposure to words and foils during the 24 two-alternative forced-choice test trials.

Performance in the *V1C2* condition was worse than in all other conditions ($p < .05$ for all comparisons). In this case, the word boundary implied between the lengthened vowel and the subsequent lengthened consonant was incongruent with that defined by transition probabilities. This accords with previous findings that statistically-defined trisyllables that straddled intonationally-defined boundaries in artificial language streams were not well recognized [33].

5. Conclusion

The three experiments demonstrate that segmental lengthening can serve as a cue to both preceding and following prosodic boundaries, depending on its distribution. As shown in Figure 1, performance was best in the two conditions (*C1* and *C1V3*) where the onset consonant of the first syllable in each word was lengthened. The worst performance was in condition *V1C2*, where a lengthened vowel was followed by a lengthened consonant within the same word.

Thus, even in the absence of other segmental and prosodic cues, listeners interpret lengthened consonants to indicate the start of a new word. This suggests that a modification is required to the iambic-trochaic law for spoken language to reflect the perceptual importance of the *locus* of prosodic lengthening effects [18, 19]: lengthened vowels cue a following boundary; lengthened consonants cue a preceding boundary.

The relative importance of timing compared to other boundary cues is not examined here. Natural speech typically provides multiple congruent sources of information about segmentation: when higher-level cues, such as lexicality and syntactic structure, offer an unambiguous guide to structure, acoustic-phonetic cues appear to be minimally exploited by listeners [15].

Our proposal for a functional division between vowels and consonants is congruent with claims that they carry distinct informational loads in speech processing, even for neonates [34, 35]. Language experience is probably required, however, before the development of differential sensitivity to localized durational effects in vowels and consonants. A similar preference for initial pitch-salience to that established in adults has been shown with 7-month-old infants, but no distinction between initial and final length-salience was found at the same age [12], indicating that more linguistic exposure is required before vowel lengthening is associated with a following boundary. The same probably applies for the interpretation of consonant lengthening as a cue to a preceding boundary. It remains to be seen whether this functional distinction holds in languages other than English.

6. Acknowledgements

This work was supported by a British Academy grant to the first author. We thank Elizabeth Gabe-Thomas, Laura König and Jean Roper for help with running experiments.

7. References

[1] Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tohkura, E. Vatikiotis-Bateson & Y. Sagisaka (eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457-463). Oxford: IOS Press.

[2] Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

[3] Fowler, C.A. (1990). Lengthenings and the nature of prosodic constituency: Comments on Beckman and Edwards's paper. In J. Kingston & M.E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. (pp. 201-207). Cambridge: Cambridge University Press.

[4] Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America*, *126*, 367-376.

[5] Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, *51*, 523-547.

[6] Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access II. Infant data. *Journal of Memory and Language*, *51*, 548-567.

[7] Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*, 2956-2970.

[8] Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.
[9] Salverda, A.P., Dahan, D., & McQueen, J.M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51-89.
[10] Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press.
[11] Hay, J.S., & Diehl, R.L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics*, *69*, 113-122.
[12] Bion, R. A., Benavides-Varela, S., & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and Speech*, 54, 123-140.
[13] Toro, J. M., Sebastian-Galles, N., & Mattys, S. L. (2009). The role of perceptual salience during the segmentation of connected speech. *European Journal of Cognitive Psychology*, *21*, 786-800.
[14] Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133-142.
[15] Mattys, S.L., White, L., & Melhorn, J.F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*, 477-500.
[16] Cutler, A., Wales, R., Cooper, N., & Janssen, J. (2007). Dutch listeners' use of suprasegmental cues to English stress. In *Proceedings of the XVIth International Congress of Phonetic Sciences* (pp. 1913-1916).
[17] Cutler, A., & Pasveer, D. (2006). Explaining cross-linguistic differences in effects of lexical stress on spoken-word recognition. *Proceedings of Speech Prosody 2006*, *Dresden* (pp. 237-240).
[18] White, L. (2002). *English Speech Timing: A Domain and Locus Approach*. University of Edinburgh PhD dissertation.
[19] White, L. (under revision). Communicative function and prosodic form in speech timing: Structure is signalled by localised lengthening effects
[20] Monaghan, P., White, L., & Merkx, M.M. (2013). Disambiguating durational cues for speech segmentation. *Journal of the Acoustical Society of America*, *134*, EL45-EL51.
[21] Oller, D.K. (1973). The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America*, *54*, 1235-1247.
[22] Keating, P. A., Cho, T., Fougeron, C., & Hsu, C. (2003). Domain-initial strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.). *Papers in Laboratory Phonology 6* (pp. 145-163). Cambridge: Cambridge University Press.
[23] Fougeron, C. & Keating, P.A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728-3740.
[24] Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America*, *118*, 3860-3873.
[25] Gow, D.W. & Gordon, P.C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 344-359.
[26] Quené, H. (1992). Durational cues for word segmentation in Dutch. *Journal of Phonetics*, *20*, 331-350.
[27] Tagliapietra, L., & McQueen, J. M. (2010). What and where in speech recognition: Geminates and singletons in spoken Italian. *Journal of Memory and Language*, *63*, 306-323.
[28] Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, *48*, 233-254.
[29] Cho, T., McQueen, J., & Cox, E. (2007). Prosodically driven detail in speech processing: the case of domain-initial strengthening in English. *Journal of Phonetics*, *35*, 210-243.
[30] Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month old infants. *Science*, 274, 1926-1928.
[31] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van Der Vreken, O. (1996). The MBROLA project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the International Conference on Spoken Language Processing, Philadelphia* (pp. 1393-1396).
[32] Baayen, R.H., Davidson, D.J., & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
[33] Shukla, M., Nespor, M. & Mehler, J. (2007) Interaction between prosody and statistics in the segmentation of fluent speech. *Cognitive Psychology*, *54*, 1-32.
[34] Benavides-Varela, S., Hochmann, J.R., Macagno, F., Nespor, M., & Mehler, J. (2012). Newborn's brain activity signals the origin of word memories. *Proceedings of the National Academy of Sciences*, *109*, 17908-17913.
[35] Bonatti, L. L., Pena, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations. The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*, 451-459.

Prosody patterns of feedback expressions in Hungarian spontaneous speech

Alexandra Markó[1], *Mária Gósy*[1,2], *Tilda Neuberger*[2]

[1] Department of Phonetics, Eötvös Loránd University, Budapest, Hungary
[2] Department of Phonetics, Research Institute for Linguistics of HAS, Budapest, Hungary
marko.alexandra@btk.elte.hu, gosy.maria@nytud.mta.hu, neuberger.tilda@nytud.mta.hu

Abstract

Speech communication incorporates non-verbal signals and semi-lexical vocal phenomena as well as words used as the listener's responses to the speaker's message. They are most common in conversation with various functions regardless of language. A specific subcategory is feedback expressions (FEs) that can be found in the listener's production as well as in the current speaker's speech production when reacting to the former speaker's message. This paper reports on the temporal and intonational characteristics of four types of FEs identified in 20 interviews and conversations from the BEA Hungarian database. Altogether 262 occurrences were categorized into four discourse functions signaling 'attention', 'comprehension', 'agreement' and 'other attitude'. Durations showed statistically significant differences across discourse functions. They were significantly longer in females than in males in all functions. The pitch range data revealed a statistically significant difference depending on discourse function and gender only in the case of the 'attention' function. The dominant frequency contour was a rise in the functions of 'attention' and 'agreement' (90%). The same contour was observed only in 75.5% of the 'comprehension' function. An integrated approach is proposed to analyze these phenomena in spontaneous speech.

Index Terms: discourse functions, prosody patterns, speaker–listener interaction

1. Introduction

Verbal communication incorporates verbal and non-verbal signals that interact both in speakers' speech production and in listeners' speech comprehension [1], [2], [3]. Semi-lexical non-verbal phenomena, short sound sequences like *ah, eh, ehm, er, erm, hmm, huh, mm, mmhm, oh, ooh, oops, phew, uh, uh-huh, um,* and even words like *yes, I see, right, okay* occur in conversations underlying cooperation between the participants of the dialogue (e.g., [4], [5]). A specific subcategory can be interpreted as 'feedback expressions' (FEs) that can be found in the listener's production as well as in the speaker's speech production when reacting to the former speaker's message. There are various terms referring to the nature of the listener's reactions like 'listener responses', 'accompaniment signals', 'continuers', 'assessments', 'acknowledgments', 'reactive tokens' (cf. [4], [6]) with meanings similar to that of the most commonly used term 'backchannel'. Backchannels have diverse definitions and descriptions (e.g., [4], [7], [8]). They are produced by one participant (the listener) in a conversation while the other one is talking. They do not cause the other speaker to cede the floor, fail to signal any intention to interrupt the speaker, and are generally non-information seeking phenomena (see [8]). The most widely identified and accepted function of backchannels is to signal attention, i.e. that the listener is attending to the speaker, reassuring the latter about his/her continuous attention [7], [9], [10]. The discourse function of a phenomenon like *uh-huh* notifying the speaker that the listener is listening was defined as early as in 1961 by Trager [11]. In a broader interpretation, however, backchannels may have various other discourse functions (such as signaling recognition, comprehension, emotional state, agreement, disagreement, attitude, support, etc., see [7], [8], [12], [13], [14], [15]), give feedbacks that make verbal communication more accurate or more continuous, and they may either support the mutual agreement between the participants or signal that some problem has arisen between them.

Backchannels signaling attention are reported to be prosodically well-defined in American English dialogues as opposed to affirmative words expressing other pragmatic functions, and the L-H% pattern was found to be characteristic of the analyzed phenomena (e.g., [7]). In addition, these backchannels were longer than affirmative words in other functions and similar to those expressing agreements.

In this paper we use the term 'feedback expression' in order to emphasize that both the listener's and the speaker's multifunctional feedback phenomena are considered in our analysis (see also [16]). In addition, FEs can occur both within turns and at turn-taking points and all of them were nonverbal phenomena.

The purpose of our study was to characterize the temporal and intonation patterns of most frequent types of FEs by measured data in Hungarian spontaneous speech. Our main question was how speakers indicate and listeners interpret the functional variations of FEs in their feedback expressions. We identified FEs as expressions that (i) responded directly to other participants' messages (irrespective of the turn of the participants), and (ii) did not require acknowledgement by the speaker. The FEs could be characterized by the following articulation gestures in most of the cases: (i) voicing emerges from the nasal cavity, potentially accompanied by an [h]-type noise component while the oral cavity is inactive, (ii) an [h]-type consonant-like (mostly voiced) sound is inserted between two low vowels. (In cases where the consonant is replaced by a very short pause, the sequence indicates negation.) In general, both versions sound as disyllabic sequences and are more or less stable in their articulation (although shorter forms might occur). The first version will be indicated by the sequence *mhm* while the second one by *uh-huh*. The articulation gestures described seem to be similar to FEs found in other languages (see 'phonetic components' in non-lexical conversational sounds in [6], [15], [17]). Former studies about some types of FEs in Hungarian supported the claim that the affirmative and interrogative functions can be correctly identified in FE extracted from spontaneous speech [18], [19], and shed light on some of their acoustic properties [20]. The prosodic structures of FEs with various functions were analyzed using experimental settings [18]. The data obtained demonstrated that the three basic types (meaning 'yes', 'no', 'question') differed in their temporal complexity and their melodic patterns. The high proportions of the listeners' correct identifications of their semantic contents confirmed the mutual

interaction between speaker(s) and listener(s). In this study, we address the acoustic patterns that disambiguate the interpretation of the three analyzed types of FEs.

Our main hypothesis was that both the durations and melody patterns of FEs are dependent on their discourse functions (see [7]). We supposed that there would be large gender differences in the various functions of FEs in all measured data.

2. Subjects, material, method

Conversations and interviews of twenty subjects (10 females and 10 males, aged between 20 and 76, mean age: 39 years) from the BEA Hungarian Spontaneous Speech Database [21] were used. All of the participants were native speakers of Hungarian from Budapest. The interviewer was always the same young female speaker (her FEs were not considered in the analysis). The total duration of the recordings was 15.2 hours; 45.7 minutes per recording, on average. All instances of FE produced by the 20 participants (irrespective of their being a listener or a speaker) were marked and labeled together with turn properties in Praat [22] by two of the authors independently of each other. Discourse function was identified by analyzing the semantic context and the speaker–listener interaction. In case of rare disagreement between the authors (below 10%), the third author was consulted. Durations, mean, minimum and maximum values of F0, pitch range (based on voice reports that were corrected manually if it was necessary) as well as the intonational structures were measured. Since a great number of FEs occurred as overlap phenomena, the melody structures could be analyzed only in 68.3% of all instances (the categorization was not problematic even in these cases). The data were subjected to various statistical analyses (one-way ANOVA, Tukey's post hoc test, Mann–Whitney U test, Kruskal–Wallis test as appropriate) using SPSS 15.0.

3. Results

3.1. Discourse functions of FEs

The majority of instances of FEs were categorized according to three main discourse functions while the fourth one contained various other attitudes that occurred in our corpus. (i) The term 'attention' will be used here when the listener signals that s/he is aware of what is being said. This is the function of notifying the speaker that the listener is listening [7]. (ii) The term 'comprehension' will be used when the listener's intention is to reassure the speaker that s/he has understood the message. (iii) The term 'agreement' will be used when the listener obviously agrees with the speaker, supports their ideas. These FEs were frequently accompanied by words like *yes*, *I see*. (iv) The term 'other attitude' serves as an umbrella term referring to phenomena that express attitudes or semantic content other than the former three types (such as surprise, disagreement, etc.). The examples demonstrate how FEs provide information about some hidden cognitive processes of the listener.

(i) Discourse function 'attention':
Interviewer: *én ugye még a régi rendszerben érettségiztem* ('well, I graduated in the old system')
Listening partner: *ühüm* ('mhm' meaning attentiveness)

(ii) Discourse function 'comprehension':
Interviewer: *mesélj egy kicsit arról hogy milyen szakos vagy, illetve hogy mivel akarsz majd későbbiekben foglalkozni* ('tell me about your university subject and about your plans for the future')
Listening partner: *ühüm öö hát én ugye magyar szakra járok* ('mhm /meaning I understand the task/ [*öö* = filled pause] well my main subject is Hungarian')

(iii) Discourse function 'agreement':
Interviewer: *az olvasáshoz hozzátartozik a színház is nyilván és az is ilyen ellenkezéseket szokott kiváltani a diákokból* ('reading is connected with theatre, obviously, and the latter often provokes disagreement from the students')
Listening partner: *aha igen persze ühüm* ('uh-huh, yes, sure, mhm')

(iv) Discourse function 'other attitude':
Interviewer: *az expressz járattal közlekedtem hát öt percenként indult* ('I used buses operating as express routes well that started in every 5 minutes')
Listening partner: *hm* ('hm' expressing surprise)

3.2. Occurrences of FEs

Altogether 262 instances of FEs with various discourse functions were found in our corpus (135 in the conversations, and 127 in the interviews). The mean occurrence of these phenomena was 13.1 per speaker (SD 11.74). Although the majority of the speakers used FEs less than 10 times in their spontaneous utterances (see the histogram in Fig. 1), great individual differences were found among them (from a single instance to 68). 142 of FEs were produced by female speakers while 120 by male speakers. No statistically significant difference was found depending on gender (mean occurrence for females is 14.2 per speaker, SD 11.44 and for males 12.0 per speaker, SD 11.6).

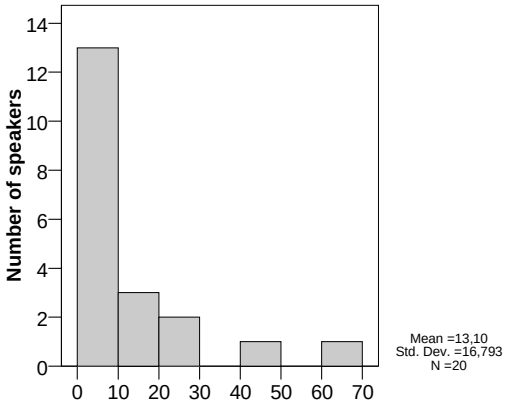


Figure 1: *The instances of tokens (FEs) across speakers (x axis = Frequency of tokens).*

The instances of FEs were heavily dependent on discourse function (Fig. 2). Females preferred FEs in the function of attention and used them more frequently than males did while males produced more instances in the discourse function of 'comprehension' than females did (Fig. 3). Since FEs occurred in the function of 'other attitude' rarely they were not included in our statistical analyses.

As expected, the majority of FEs (n = 130; 49.6% of all) occurred signaling **attention** and were produced both as in-between or overlap phenomena (see [23]), meaning that they were inserted (by the listener) during the speaker's turn in a pause period or they were uttered while the speaker was continuously speaking. The dominant form was *mhm* (90.8%; n = 118); *uh-huh* occurred in 8.5% (n = 11) of the cases and

one instance of *uh-hum* (1.7%) was also found. The latter is supposed to be the blending of the sequences *mhm* and *uh-huh.*

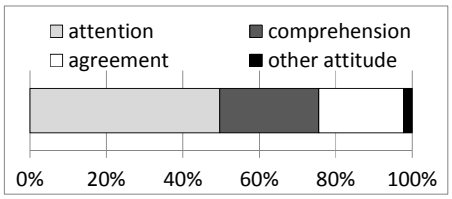


Figure 2: *The proportion of discourse functions of FEs in the corpus.*

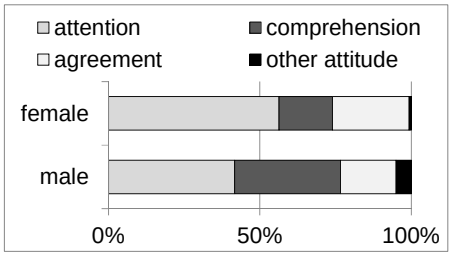


Figure 3: *The proportion of discourse functions of FEs depending on gender in the corpus.*

The discourse function signaling **comprehension** accounted for 25.6% of all instances (n = 67), and a greater variety of forms was used than in the 'attention' function. The type *mhm* was found in 70.1% (n = 47) of all instances in this function, while *uh-huh* occurred in 28.4% (n = 19), and one blended *u-um* form (1.5%) was found here, too. The instances of this function were produced during the speaker's turn in 77.6% of all cases, and 17.9% of them occurred at turn boundaries (Fig. 4).

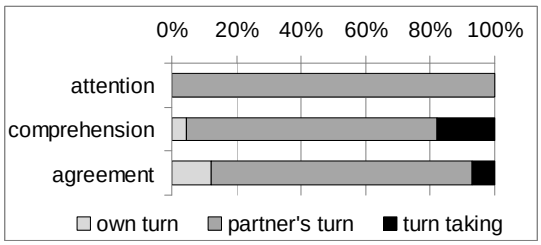


Figure 4: *The proportion of the placements of the three types of FEs.*

The discourse function **'agreement'** occurred in 22.1% of all cases (n = 58). 75.9% of them (n = 44) was identified as *mhm* and 24.1% of them (n = 14) as *uh-huh*. The great majority of the instances in this function (81.0%) were found during the speaker's turn and 6.9% of them occurred at turn boundaries, while some (12.1%) were found in the former listener's own turn.

3.3. Durations of FEs

The interrelations of the discourse functions and their durations are demonstrated in Figure 5. The forms *mhm* are longer (mean 250 ms, SD 56 ms) than those of *uh-huh* (mean 247 ms, SD 55 ms) in the function of 'attention'. The mean duration of *mhm* in the function of 'comprehension' is 289 ms (SD 70 ms) while that of *uh-huh* in the same function is 253 ms (SD 41 ms). The mean duration of *mhm* in the function of 'agreement' is 258 ms (SD 71 ms) while that of *uh-huh* in the same function is 232 ms (SD 63 ms). Instances in the function of 'comprehension' had the longest durations while no large differences were found in the cases of the other two functions. Statistical analysis confirmed a significant difference in the durations of instances depending on discourse function (one-way ANOVA: $F(5, 239) = 3.443$, $p = 0.005$). The Tukey's post hoc test shows, however, that the difference was significant only in the duration of *mhm* forms, and only between the functions of 'attention' and 'comprehension' ($p = 0.004$).

Durations were also analyzed in terms of gender. In this analysis only the *mhm* forms were considered due to the low number occurrences of the other forms. The mean duration of *mhm* in the function of 'attention' was 267 ms in the females (SD 40) and 200 ms in the males (SD 49). In the function 'comprehension' it was 335 ms (SD 42) in females and 262 ms (SD 69) in males, and in the function of 'agreement' it was 291 ms (SD 38) in females and 198 ms (SD 44) in males (Fig. 6). Statistical analysis including Tukey's post hoc test revealed a significant difference in durations of FEs between females and males (one-way ANOVA: $F(5, 196) = 21.011$, $p < 0.001$).

3.4. Pitch ranges and melody contours of FEs

Both the fundamental frequency values and the intonation patterns of all measurable instances of FEs were analyzed. In the discourse function of 'attention' 84 tokens were eligible for analysis in terms of pitch, as well as 49 instances of signals of comprehension, and 40 occurrences of the discourse function of 'agreement'. In other functions altogether 6 tokens were analyzed. The pitch range data are shown in Table 1.

Table 1. *Pitch range values depending on discourse function and gender.*

Discourse functions	Pitch range (semitone)			
	Females		Males	
	Mean	SD	Mean	SD
Attention	4.15	1.48	2.74	1.06
Comprehension	3.54	1.55	3.26	1.16
Agreement	3.48	1.35	3.80	1.67

Statistical analysis confirmed a significant difference in the pitch ranges (Kruskal–Wallis test: $\chi^2(5) = 15.813$, $p = 0.007$) while the Mann–Whitney U test revealed that there was significant difference in pitch ranges depending on gender only in the function of 'attention' ($Z = -4.004$, $p = 0.001$).

The prototypical element of the melody patterns of FEs was, in general, a final rise which was frequently preceded by a fall, a descent or monotonous contour as a preparatory one. Sometimes glottalized syllables occurred preceding the rise (e.g., [4]). 94.0% (n = 79) of all instances in the function of 'attention' ended with a rise (Fig. 7).

A similar rising contour or preparatory contour + rising was characteristic of the function of 'comprehension' in 75.5% (n = 37) of the cases. However, the descent contour was also produced in this function, in 24.5% (n = 12) of all cases. Acoustically, the steepness of these contours was diverse and various complex structures like descent + step up/rise + descent, fall + monotonous, rise + descent were also found.

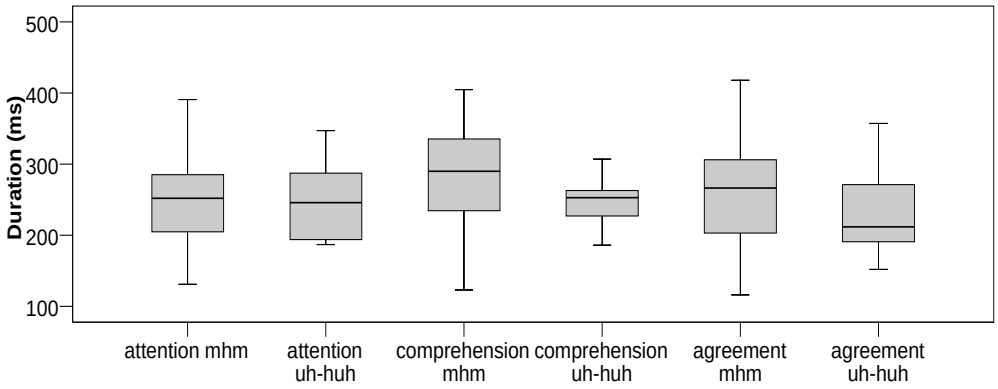


Figure 5: *Durations (medians and ranges) of various forms of FEs in the various discourse functions.*

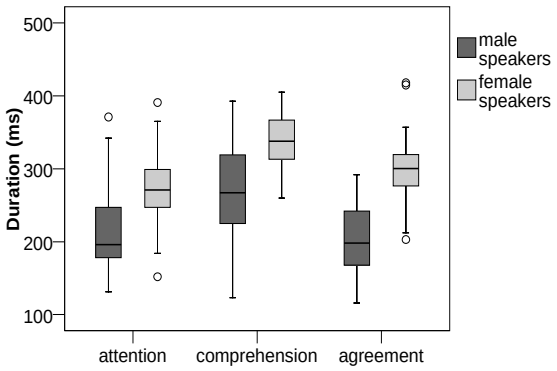


Figure 6: *Durations (medians and ranges) of* mhm *depending on the discourse function and as a function of gender.*

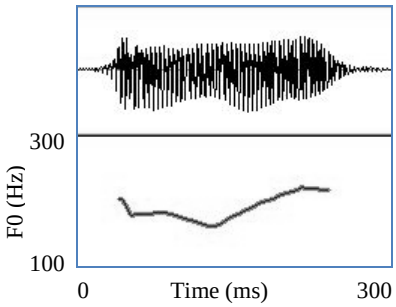


Figure 7: *Prototypical melody pattern of a* mhm *signaling attention.*

In the function of 'agreement', the dominant melody contour was also the rise (92.5%; n = 37), while a small portion (2.5.0%; n = 3) showed intonation structures with no rising contours.

Four instances of the mixed type function expressing surprise and disagreement showed descent melody contours. One token expressing a question had a rising contour while another one expressing 'no' had a monotonous contour.

4. Discussion

In this study we provided measured data on instances of FEs concerning their occurrences, temporal and melody patterns in view of the three main discourse functions they reflect. As expected, listeners felt it to be the most important to notify the speaker about their attentiveness. However, females and males behaved differently: the former used FEs in the function 'attention' more frequently than males did while the latter preferred to signal their comprehension during conversation. Males' frequent signaling of comprehension might be explained by the fact that the interviewer was a female speaker.

The main hypothesis of the research, however, was only partly confirmed. The closest interrelations were found between the durations of the various forms of FEs and their discourse functions, suggesting that speakers seem to differentiate the discourse functions by articulating them differently along the durational scale. In addition, the different forms (*mhm* and *uh-huh*) with different durations depending on function support the listeners' intention to inform the speaker about their attitudes. Females seem to express certainty and reassurance concerning the speaker's message by longer durations of FEs than males do. The non-neutral rising intonation contours of the majority of FEs reinforce the information that the discourse functions convey, while the relatively frequent fall/descent contours signal that the utterance is self-contained and finished (see [24]).

5. Conclusions

Our research findings evidenced speaker and listener interactions in conversations by measured acoustic-phonetic data of FEs with three main discourse functions. The temporal and melody patterns that the speakers produced do not seem to be incidental; however, differences in the actual acoustic patterns could be shown across languages like English, Italian, Japanese, Swedish, and Hungarian [4, 7, 10, 13, 17, 25]. However, identification of a discourse function must take into account various other factors (like gaze direction). We can conclude that analyzing feedback expression phenomena using an integrated approach considering communication situation, participants, contexts, various types of feedbacks, function, and acoustic patterns [e.g., 2, 3, 10] is crucial to fully understand spontaneous conversations.

6. Acknowledgements

We wish to thank Louise Mycock for her help concerning an earlier version of this paper.

This research was supported by OTKA project No. 108762.

7. References

[1] Schmidt, J. E., “Neue Wege der Intonationsforschung”, Georg Olms Verlag, Hildesheim, Zürich, New York, 2001.

[2] Jones, S. E. and LeBaron, C. D., “Research on the Relationship between Verbal and Nonverbal Communication: Emerging Integrations”, Journal of Communication, 52(3):499-521, 2002.

[3] Allwood, J., and Cerrato, L. “A study of gestural feedback expressions”, In First Nordic Symposium on Multimodal Communication, Copenhagen, 7-22. 2003.

[4] Ward, N. and Tsukahara, W., “Prosodic features which cue back-channel responses in English and Japanese”, Journal of Pragmatics, 32(8):1177-1207, 2000.

[5] TEI: Text Encoding Initiative: http://www.tei-c.org/index.xml, download: 17 Nov 2013.

[6] Miller, L. Verbal listening behavior in conversations between Japanese and Americans. The Pragmatics of International and Intercultural Communication. Amsterdam: John Benjamins Publishing Company, 111-130. 1991.

[7] Benus, S., Gravano, A., and Hirschberg, J. The prosody of backchannels in American English. In Proceedings of ICPhS 2007, 1065-1068. 2007.

[8] Lai, C. “Prosodic Cues for Backchannels and Short Questions: Really?” Speech Prosody Conference. 2008.

[9] Yngve, V. “On getting a word in edgewise”, Papers from the Sixth Regional Meeting of the Chicago Linguistic Society: 567-577. 1970.

[10] Gravano, A., Benus, S., Chavez, H., Hirschberg, J., and Wilcox, L. On the role of context and prosody in the interpretation of okay. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages 800-807. 2007.

[11] Trager, G.L., “The typology of paralanguage”, Anthropological Linguistics, 3:17-21, 1961.

[12] Ward, N., "Pragmatic functions of prosodic features in non-lexical utterances." Speech Prosody 2004, International Conference, 325-328, Japan, 2004.

[13] White, S., “Backchannels across cultures: a study of Americans and Japanese”, Language in Society, 18:59-76, 1989.

[14] Gardner, R., “When Listeners Talk: Response Tokens and Listener Stance”, Amsterdam, J. Benjamins Publishing, 2001.

[15] Ward, N., “Non-Lexical Conversational Sounds in American English”, Pragmatics and Cognition, 14:113-184, 2006.

[16] Allwood, J., Nivre, J., & Ahlsén, E. On the semantics and pragmatics of linguistic feedback. Journal of semantics, 9(1), 1-26. 1992.

[17] Cerrato, L. Some characteristics of feedback expressions in Swedish. In Proc. of Fonetik. Vol. 44, pp. 41-44. 2002.

[18] Markó, A., “Szavak nélkül. Nonverbális vokális közlések fonetikai elemzése” [Without words. A phonetic analysis of nonverbal vocal communication], in Hungarian, Magyar Nyelvőr, 129:88-104, 2005.

[19] Markó, A., “A special conversational device: humming in Hungarian”, The Phonetician, 95:28-31.

[20] Neuberger, T. and Beke, A. “Automatic Laughter Detection in Spontaneous Speech Using GMM-SVM Method”, in I. Habernal and V. Matousek [eds.], Text, Speech and Dialogue, 16th International Conference, TSD 2012, Pilsen, Czech Republic, Springer, 113-120, 2012.

[21] Gósy, M., “BEA – A multifunctional Hungarian spoken language database”, The Phonetician, 105/106:50-61, 2012.

[22] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer”, Computer program, Version 5.2 retrieved 10 Sept 2010 from http://www.praat.org/

[23] Feke, M. S., “Effects of Native-language and Sex on Back-channel Behavior”, in L. Sayahi [ed.], Selected Proceedings of the First Workshop on Spanish Sociolinguistics, 96-106, Somerville, MA: Cascadilla Proceedings Project, 2003.

[24] Varga, L. “Intonation and stress. Evidence from Hungarian”. Houndmills, Basingstoke, Palgrave Macmillan. 2002.

[25] Cerrato, L. “Investigating communicative feedback phenomena across languages and modalities”, Doctoral Thesis, Stockholm, Sweden. 2007.

Intonation Patterns of Morelos Nahuatl

Eduardo Patricio Velázquez Patiño[1]

[1] Facultad de Lenguas y Letras, Universidad Autónoma de Querétaro, Mexico
utka@yahoo.com

Abstract

There are still relatively few studies on the phonetics and phonology of the indigenous languages of Mexico, and just a minority of them deals with less explored areas like prosody or, specifically, intonation. This study reports a preliminary analysis of Nahuatl intonation, taking into account its phonological characteristics: a) trochaic binary rhythm; b) generation of secondary stress inside rhythmic structures; c) generation of rhythmic groups according to clause structures; d) phonetic syllable lengthening at the end of sentences; e) laryngealization or voicelessness at the end of utterances, and f) vowel lengthening. Data collected by means of different methods, developed in order to obtain authentic and spontaneous utterances, show that different sentence types tend to have specific intonation patterns with many typologically common features and some original characteristics.
Index Terms: prosody, intonation, acoustic phonetics, indigenous languages, language documentation, Nahuatl

1. Introduction

Nahuatl is part of the Uto-Aztecan family and its ten dialects are spoken in vast regions of Mexico and Central America. Dialectal differences cause varying levels of intelligibility among speakers. Moreover, there are few monolingual speakers, since the language is rapidly losing ground to Spanish [1]. Here we focus on Morelos Nahuatl. In 2000, Morelos Nahuatl was spoken by 18,700 persons [2]. In the village of Cuentepec, the most dynamic community of all Morelos Nahuatl-speaking villages, there were 3,052 Nahuatl speakers who were older than five years, but only 69 of them were monolinguals [3].

Although classical Nahuatl or Aztec is one of the most documented and studied indigenous languages in the Americas, especially with respect to its polysynthetic morphology, but the phonology of its current dialects, and specifically its prosody, which is usually regarded by non-specialists as 'very simple', has not been studied from the approach of experimental and acoustic phonetics [cfr. 4, 5, 6, 7]. The model presented in this paper takes into account specific features of Nahuatl phonology and how they influence intonation.

2. Previous studies

2.1. Phonological inventory of Morelos Nahuatl

The phonological system of Morelos Nahuatl, specifically the variant spoken in the village of Cuentepec, is shown in Table 1. Phonemes borrowed from Spanish, used in non-assimilated loanwords, are not included. The system has a characteristic absence of voiced stops affricates and fricatives, as well as a squared vowel system. Regarding phonotactics, Table 1 shows consonantal phonemes appearing only in the onset (/h^{w}-/) or coda position (/-ʔ/). Phonetically, short and long vowels differ not only in their length but also in their quality and tension, with short vowels being lower and lax, especially the high front vowel, /i/ → [ɪ], and the back vowel, /o/ → [ɔ ~ ʊ]. However, in spite of the existence of minimal pairs, the vowel length contrast is not robust in the variant of Cuentepec, Morelos, since it tends to be neutralized and subordinated to rhythmic structure and emphatic speech. Moreover, syllables with a glottal stop /ʔ/ or fricative /h/ in coda position, which tend to neutralize, produce a shortening of the vowel's modal portion followed by a glottalization, which is produced as creaky voice or voicelessness. This phenomenon occurs only at the end of an utterance [6]. Other allophonic phenomena not shown in Table 1 are velar stop fronting before a front vowel, /k/ → [c], nasal velarization at the end of a word, /n/ → [ŋ], and devoicing and fricatization of the lateral consonant at the end of a word, /l/ → [ɬ ~ tɬ].

Table 1. Phonological Inventory of Morelos Nahuatl.

	CONSONANTS						
	bilabial	dental	alveolar	palatal	velar	velar-labial	glottal
Stops	p	t			k	k^{w}	(-ʔ)
Affricates		t͡s	t͡ɬ	t͡ʃ			
Fricatives		s		ʃ		(h^{w}-)	h
Nasals	m	n					
Lateral			l				
Approximants				j		w	
	VOWELS						
				iː i		o oː	
				eː e		a aː	

2.2. Syllable structure

Nahuatl syllable structure has been traditionally (e.g. [8]) said to allow four possible syllable patterns, which are derived from the generic pattern (C)V(C), i.e. V, CV, VC, and CVC. Allegedly, there are no consonant or vowel clusters inside a single syllable (dipthongs are regarded as CV or VC structures), and a syllable is composed at most of three phonemes. However, in Cuentepec Nahuatl there is evidence of CGVC (e.g. [ˈtjaŋ.kɪs] 'market') and CVGC (e.g. [k^{w}ejtɬ]'skirt') structures. Last but not least, lexical stress functions as a culminative, hierarchical, delimitative and rhythmic unit [9], falling regularly on the penultimate syllable.

Typologically, Nahuatl has a polysynthetic structure with the possibility of the object and even some modifiers being incorporated to the verb root. Structures above five or six syllables are not the most frequent, but they are not unusual either. In such words secondary stress is found. According to Kager's classification [9], these synthetic structures in Morelos Nahuatl are typical of a bound or limited system,

which is also non-sensitive to quantity, and it assigns stress from right to left, producing left-oriented, i.e. trochaic feet:

[nik.ˌkʷah.kʷal.ˌt͡ɬaʃ.kal.ˈma.ka] (1)
/ni-k-kʷaːhkʷal-t͡ɬaʃkal-maka/
1SG.SUJ-3SG.OBJ-beautiful-tortilla-give.PRES
'I give him/her beautiful tortillas.'

All three acoustic correlates (intensity, duration, and fundamental frequency or F_0) signal, to a greater or lesser extent, the location and prominence level of primary stress, secondary stress, and lexical stress throughout the sentence. As such, the only constant acoustic correlate of stress is F_0 [6].

2.3. Influence of stress and length on vowel quality

In a previous study [10] a young male speaker of Morelos Nahuatl was asked to read and record a series of sentences. From this recording, we obtained 542 vowel tokens (including 246 short vowels, and 296 long vowels, in both stressed and unstressed syllables), after eliminating every observation in unclear contexts, adjacent to a semivowel, or before any sonorant consonant. Values for the three first formants were calculated using a Praat script [11]. The values were obtained from the most stable portion of every vowel. F2' [12] was calculated from the F2 and F3 values (see Table 2, Figure 1).

Table 2. Formant values of Nahuatl vowels.

Stress	Length	#	$\overline{F_1}$	$Q_{F1}(0.1)$	$Q_{F1}(0.9)$	$\overline{F_2'}$	$Q_{F2'}(0.1)$	$Q_{F2'}(0.9)$
-	a	17	594	501	658	1640	1434	1774
	aː	24	702	651	758	1703	1529	1790
+	a	76	669	566	734	1739	1592	1893
	aː	60	721	669	788	1733	1573	1846
-	e	24	503	428	582	1918	1782	2049
	eː	22	535	495	573	2095	1957	2174
+	e	30	520	452	586	2072	1916	2243
	eː	33	536	474	586	2068	1844	2207
-	i	29	376	259	455	2283	2018	2539
	iː	70	313	279	358	2462	2253	2641
+	i	21	316	286	350	2392	2315	2500
	iː	20	321	311	331	2553	2416	2707
-	o	26	473	383	540	1476	1341	1627
	oː	15	515	493	546	1398	1335	1479
+	o	23	559	484	629	1459	1340	1604
	oː	52	582	521	635	1409	1308	1511

In Table 2, the resulting data are presented according to their timbre, presence or absence of stress, and length. The next column indicates the number of occurrences (#) for each vowel category, their mean values, as well as the 0.1 and 0.9 quantiles inside every category for F_1 and F_2'. These latter values were used to plot ellipses with the most stable 80% of the occorrences of all four vowel extreme values.

In Figure 1, we observe that long vowels (depicted with dashed lines) are better delimited, since they occupy a smaller acoustic space; their spaces between stressed (black line) and unstressed (gray line) vowels intersect to a lesser or greater extent, and the different vowel timbres are clearly differentiated. Meanwhile, short vowels (depicted with solid lines) occupy larger acoustic spaces, where short unstressed /i/ covers a vast space towards /e/, whereas short unstressed /o/ and /a/ tend to occupy higher spaces, causing /a/ to intersect large areas occupied by phonological /o/. The Nahuatl vowel system is thus square, not triangular.

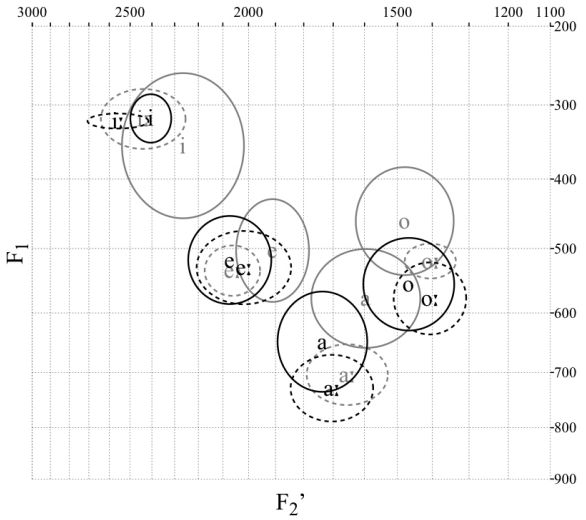


Figure 1: Distribution of Nahuatl vowels.
(Black line = stressed vowel, gray line= unstressed vowel, solid line= short vowel, dashed line= long vowel)

2.4. Tones and break indices (ToBI)

For this study, we use the ToBI (Tones and Break Indices) prosodic transcription system, based on the autosegmental-metrical model or AM [13, 14, 15, 16, 17, among others]. In this transcription sstem, F_0 contours are analyzed as a sequence of local pitch variations based on the metrical structure of sentences. As such, there are two types of phonological elements: *pitch accents*, associated with lexically stressed syllables coinciding with nuclei of syntactic clauses, and *boundary tones*, associated with the borders of prosodic domains. There are just two basic tones: *L* (low tone) and *H* (high tone), which are combined thoughout the utterance.

3. Data Elicitation

For the macro-project on the prosodic analysis of Nahuatl, within which the present investigation is placed, we have taken into account the elicitation techniques used in other projects, such as the Atlas interactivo de la entonación del español [18] the HCRC Map Task Corpus [19], and the CHILDES Project [20]. We have developed a whole series of methods and materials in order to elicit close-to-natural intonation. These materials include interviews, map tasks, narratives, and communicative situations. We have used these materials to obtain data from a number of Nahuatl varieties, including those spoken in Morelos, Guerrero, Huasteca Veracruzana and Sierra de Zongolica. This paper focuses exclusively on the results obtained for Morelos Nahuatl.

3.1. Interviews

The topics presented during interviews are previously prepared, but also improvised in order to help the native speaker talk about his/her life experiences and interests, such as life stories, descriptions of his/her work or trade, daily life, family, community, as well as folk stories, tales or legends. Ideally, the recordings should be more like a conversation than an interview, and performed by another native speaker. They yield mostly declarative sentences.

3.2. Map Task

Another elicitation technique that we have employed is the one known as 'Map Task' [19]. This is a cooperative task between two native speakers performing different roles: the instruction giver and the instruction follower. The instruction giver gives instructions about a path to be followed through a village map in order to arrive somewhere, and the other participant follows those instructions, trying to trace the route. However, the maps are slightly different, and both participants have to figure out how to navigate through them successfully [19, 21]. This method yields declaratives, interrogatives, and imperatives.

3.3. Narrative

In order to elicit narratives, booklets were created with image sequences depicting a story. However, this technique was not always successful for this purpose, as illiterate participants described individual images in detail instead of producing a narrative. This situation yielded mostly declaratives, but also other types of sentences.

3.4. Communicative situations

The researcher read aloud 75 common situations, intended to elicit different types of sentences as a response: declaratives, (absolute, partial, and reiterated) interrogatives, imperatives, and vocatives. For each main type of sentence, neutral and non-neutral variants were elicited. Native speaker participants were asked to act out a sentence, or to interact in a close-to-natural way, according to the given situation (see Figure 2). The questionnaire was not just translated [18, 22, 23], but also culturally and thematically adapted in order to create topics and characters, which would appear familiar to native speakers of Nahuatl. Sometimes, a Spanish version was used and adapted on the fly in order to obtain different answers.

Figure 2: *Sample from the 'Communicative Situations' task.* (4a. Look at this image and tell me what they are doing. [Possible answer:] –A woman is giving water to her son.)

4. Methodology

For this study, three native speakers of Morelos Nahuatl, one man and two women from the community of Cuentepec, were selected for the *Communicative Situations Questionnaire*. They were asked to answer every question three times. Four different speakers were also selected for the Map Task, the narratives and the interviews, in order to provide complementary data.

All linguistic productions were recorded in audio as stereo WAV files with an Olympus LS-11 recorder at a 44.1MHz sample rate, and in video as a backup. The questionnaire was recorded in a soundproof room at the Universidad Autónoma de Querétaro, Mexico, while the complementary data were recorded in quiet spaces in Cuentepec. The resulting audio files were edited in Audacity, and then segmented, transcribed and prepared in Praat [24, 25].

5. Results

Figure 3 shows a sample image obtained by means of superposing all three utterances of the male speaker, corresponding to three superposed emissions of a partial interrogative sentence, with a H% [H*+H] [H*] [h* H*] LH% pattern, where brackets show the domain of phrasal structures.

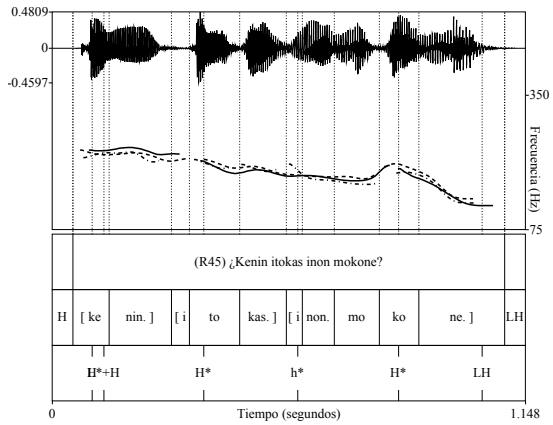


Figure 3: *Three emissions of the same partial interrogative sentence.* ('How will you name your child?')

The results obtained from the quantitative and qualitative analysis of all sentences in a given category, their pitch accents and their simplified global contours were mapped on to the stylized contours shown in Figure 4. For the sake of simplicity, secondary stress pitch accents (notated with lower case h* in Figure 3) are not depicted here, since they show the same inventory as is used for primary stress, and the difference between primary and secondary accents is just relative and contextual. On the other hand, H* differs from * in that H* is a perceptually clear F_0 peak (over +1.5 semitones), while * is not so easily perceivable, because its F_0 is under 1.5 st, relative to its surrounding syllables. Parentheses and slashes show second-best alternative pitch accents (also depicted with a thinner line), which correspond mostly to local emphasis.

5.1. Declaratives

Neutral declaratives (A.1. in Figure 4) have by far the simplest contours: low initial and final boundary tones (L%) and unperceivable nuclear and prenuclear stresses (*) or, alternatively, peaks within the stressed syllable domains (H*). Non-neutral declaratives (A.2.) differ from neutral ones in that the initial and final accents are early rises (L+H*). Their final boundary might be a fall or a steep fall (L% or LL%).

5.2. Absolute interrogatives

Both neutral (B.1. in Figure 4) and non-neutral (B.2.) absolute interrogatives have a low nuclear accent followed by a rise to a level tone (L* HL%). However, neutral ones have a high initial boundary tone followed by a low/falling tone, while non-neutral ones have a similar contour with respect to non-neutral declaratives, except for the nuclear pitch accent and final boundary tones (L+H* L% vs. L* HL%, which is a low tone followed by a steep rise that ends at a mid/neutral level).

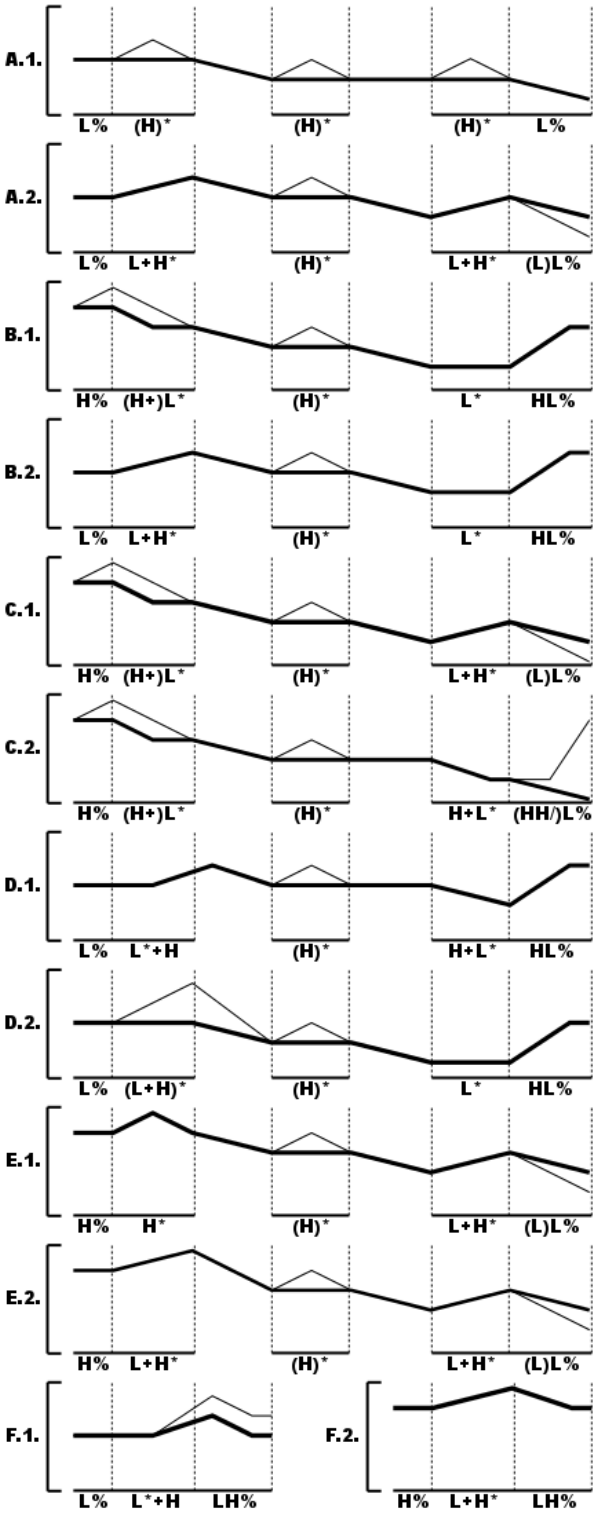


Figure 4: *Intonation Patterns of Morelos Nahuatl.*

5.3. Partial interrogatives

Partial interrogatives, both neutral (C.1. in Figure 4) and non-neutral (C.2.), have the same contours as neutral absolute interrogatives (H% L* or H% H+L*), but they have different nuclear pitch accents: neutral ones resemble non-neutral declaratives (A.2.) and non-neutral ones have a falling tone (H+L*) followed by totally different contours: a fall or a steep rise, L% vs. HH% (maybe due to pragmatic factors).

5.4. Reiterated interrogatives

Neutral (D.1.) and non-neutral (D.2.) reiterated interrogatives have a characteristic low initial boundary tone, and a rising to level final boundary tone. They differ only in the first prenuclear tone (L*+H vs. * or L+H*), and the nuclear tone (H+L* vs. L*).

5.5. Imperatives

Imperatives, both commands (E.1.) and pleas (E.2.), begin with high tones, and end with nuclear early rises followed by low/extra low final tones (L+H* L% / LL%). Their prenuclear accents differ in peak alignment (H* vs. L+H*).

5.6. Vocatives

Vocatives have falling to level final tones (LH%), but they have different initial tones, L% vs. H% for calls (F.1. in Figure 4) and phrase-edge vocatives (F.2.), respectively, and nuclear F_0 peak alignments (L*+H vs. L+H*), where calls have a deeper pitch range.

6. Conclusions

We propose a combinatorial model made up of isolated tones, which also takes into account the trochaic binary rhythm of Nahuatl, and its automatic generation of secondary stress inside rhythmic structures according to phrasal (clause) structures, as shown in Figure 5. So far, we have identified slightly different intonation contours for each of the main pragmatically neutral sentence types, where neutrality is mainly expressed by means of simpler or less salient tones. The combinatorial model in Figure 5 must be taken as provisional and subject to revision based on the results of further data collection and analysis.

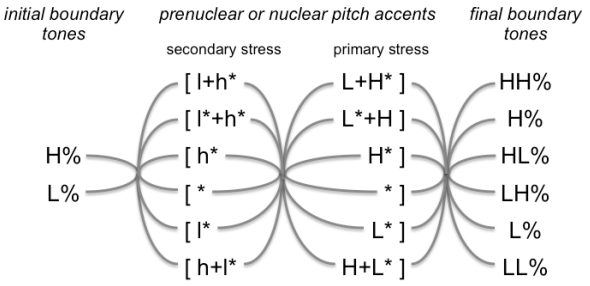


Figure 5: *Tone combinations in Morelos Nahuatl.*

7. Acknowledgements

The author would like to thank Victorino Torres Nava and his relatives for providing us with access to their language and their culture. A special acknowledgment goes to the Fondo de Fortalecimiento a la Investigación, the group of research assistants at Universidad Autónoma de Querétaro, José Ignacio Hualde (University of Illinois at Urbana-Champaign), and Carolyn O'Meara (Univ. Nacional Autónoma de México).

8. References

[1] Canger, U., "Nahuatl", in K. Brown and S. Ogilvie [Eds], Concise Encyclopedia of Languages of the World, Elsevier, 745, 2006.

[2] Lewis, M. P. [Ed], Ethnologue: Languages of the World. 16th ed., SIL International, 2009.

[3] INEGI (Instituto Nacional de Estadística y Geografía), "II Conteo de Población y Vivienda 2005", Mexico City: INEGI, 2005.

[4] Robinson, D. F. "Puebla (Sierra) Nahuat Prosodies", in D. F. Robinson [Ed], Aztec Studies I. Phonological and Grammatical Studies in Modern Nahuatl Dialects, Mexico City: SIL, 17-32, 1969.

[5] Canger, U. "(Changing) Word Prosody in Nahuatl" in C. Chamoreau, Z. Estrada Fernández and Y. Lastra [Ed], A New Look at language Contact in Amerindian Languages, Munich: Lincom Europa, 49-69, 2010.

[6] Velázquez Patiño, E. P. "Estructura acentual del náhuatl de Cuentepec, Morelos", in E. Herrera Zendejas [Ed], Tercera Mesa de Trabajo del Seminario Phonologica: Tono, Acento y Estructuras Métricas en Lenguas Mexicanas, Mexico City: El Colegio de México, in press.

[7] Guion, S. G., Amith, J. D., Doty, Ch. S. and Shport, I. A. "Word-level prosody in Balsas Nahuatl: The origin, development, and acoustic correlates of tone in a stress accent language", in Journal of Phonetics, 38-2, 137-166, 2010.

[8] Guzmán Betancourt, I., Gramática del Náhuatl de Santa Catarina, Morelos, Mexico City: INAH, 1979.

[9] Kager, R. "The Metrical Theory of Word Stress", in J. A. Goldsmith [Ed], The Handbook of Phonological Theory, Blackwell, 367-402, 1995.

[10] Velázquez Patiño, E. P. "Duración vocálica y acento en el náhuatl de Cuentepec, Morelos" in E. P. Velázquez Patiño and I. Rodríguez Sánchez [Eds], Estudios de Lingüística Funcional, Universidad Autónoma de Querétaro, in press.

[11] Boersma, P. and Weenink, D. 2012. "Praat: doing phonetics by computer (Computer program, Version 5.3.60)". Online: http://www.praat.org/, accesses on 29 Dec 2013.

[12] Fant, G., "The Acoustics of Speech", in G. Fant [Ed], Speech Sounds and Features, MIT, 3-16, 1973.

[13] Pierrehumbert, J. "The phonetics and phonology of English intonation", PhD dissertation, MIT, 1980 [IULC edition, 1987].

[14] Pierrehumbert, J. and Beckman M., Japanese Tone Structure, MIT, 1988.

[15] Ladd, R., Intonational phonology, CUP, 1996.

[16] Beckman, M. E. and Ayers, E. G., "Guidelines for ToBI Labelling. Version 3 (1997)". Online: http://ling.ohio-state.edu/~tobi/ Guidelines for ToBI Labelling.htm, accessed on 29 Dec 2013.

[17] Beckman, M. E., Díaz-Campos, M., Tevis McGory, J. and Morgan, T. A., "Intonation across Spanish, in the Tones and Break Indices framework", in Probus, 14: 9-36, 2002.

[18] Prieto, P. and Roseano, P. [Coords], "Atlas interactivo de la entonación del español, 2009-2010". Online: http://prosodia.upf.edu/atlasentonacion/, accessed on 29 Dec 2013.

[19] Anderson, A. H., Clark, A. and Mullin, J., "Introducing information in dialogues: How young speakers refer and how young listeners respond", in Journal of Child Language, 18: 663-687, 1991

[20] McWhinney, B., The CHILDES Project: Tools for Analyzing Talk, Volume 2, Routledge, 2000.

[21] Brown, G., "Investigating listening comprehension in context", in Applied Linguistics, 7, 284-302, 1986.

[22] De la Mota, C., Martín Butragueño, P. and Prieto, P., "Mexican Spanish Intonation", in P. Prieto and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, Munich: Lincom Europa, 319-350, 2010.

[23] Estebas-Vilaplana, E. and Prieto, P., "Castilian Spanish Intonation", in P. Prieto and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, Munich: Lincom Europa, 17-48, 2010.

[24] Cantero Serena, F. J. and Font Rotchés, D., "Entonación del español peninsular en habla espontánea: patrones melódicos y márgenes de dispersión", in Moenia, 13, 69-92, 2007.

[25] Cantero Serena, F. J. and Font Rotchés, D., "Protocolo para el análisis melódico del habla", in Estudios de Fonética Experimental, XVIII, 17-32, 2009.

Modelling interlanguage intonation: the case of questions

Sophie Herment[1], Nicolas Ballier[2], Elisabeth Delais-Roussarie[3], Anne Tortel[4]

[1] Aix-Marseille Université, UMR 7309 - Laboratoire Parole et Langage, France
[2] Université Paris-Diderot, CLILLAC – ARP (EA 3967), France
[3] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France
[4] Université Nice-Sophia Antipolis, UMR 7320 - Bases, Corpus, Langage, France

sophie.herment@univ-amu.fr, nicolas.ballier@univ-paris-diderot.fr,
elisabeth.roussarie@wanadoo.fr, atortel@unice.fr

Abstract

In this paper, we study the intonational patterns observed in learners' productions in order to evaluate what motivates the deviations observed: systemic differences between the learners' L1 and the L2, differences in phonetic implementation, etc. The analysis consists of a cross-comparison of the intonation of yes-no questions in French, English and English as an L2. It is based on five information-seeking yes-no questions that were extracted from the AixOx corpus, which contains a set of 40 texts that were read by 10 native French speakers, 10 Native English speakers and 20 French learners of English. The analysis of the data showed that the differences between native and non-native speakers do not affect the form of the nuclear contour. It mostly shows that French speakers of English have a tendency to assign a rising pitch movement at the end of any prosodic words, which leads to a clear difference in rhythm.

Index Terms: acquisition of prosody in L2, intonation, prosodic phrasing, rhythm, prosodic modelling, learner corpora, interphonologies.

1. Introduction

Research on interlanguage intonation has shown that the intonational patterns observed in learners' productions are often influenced by their L1s (see, among others, [1], [2] and [3]). As a consequence, the notion of L1 transfer is often invoked to account for the observed patterns. As pointed by [4], however, transfer may apply at the phonological as well as at the phonetic level. Transfers at the phonological level result from differences in the metrical structure or the tonal inventory. In a study on the intonation of tag questions in English, [5] have shown for instance that Spanish speakers of English use rises at the end of the question tag for confirmation request, whereas native English speakers will use falls, these patterns being thus analyzed as resulting from a phonological transfer. By contrast, transfers at the phonetic level occur when an identical phonological form differs in the way it is phonetically implemented in both languages. Differences in the temporal alignment of pitch accents may be for instance a case of phonetic transfer (see [1] for concrete examples). The distinction between different types of transfers or deviations is of great help to study interlanguage intonation, as pointed by [6].

In this paper, we will show however that classifying the observed deviations is not an easy task since the deviation type may change over time, and a mere cross-comparison of the surface forms may not always be sufficient. This will be done through the analysis of information-seeking *yes-no* questions realized by French learners of English and extracted from the AixOx corpus ([7] and [8]).

The paper is organized as follows: in section 2, the prosodic characteristics of information-seeking *yes-no* questions in French and English are described; section 3 presents the data and methodology used for the prosodic analysis; the results obtained in the cross-comparison are given in section 4; section 5 discusses the results and offers perspectives and concluding remarks.

2. The intonation of *yes-no* questions

2.1. Yes-no questions in French

From a morpho-syntactic point of view, three distinct constructions may be used in French to build up a *yes-no question*: (i) declarative structures similar to the ones observed in assertive sentences (1); (ii) subject-object inversion, be the subject nominal or pronominal (2); and (iii) an interrogative particle *est-ce que* can be inserted in sentence initial position, the rest of the sentence having the same syntactic structure as in assertions (3).

(1) *Vous avez appris des langues étrangères ?* ('Did you learn any foreign languages?')
(2) *Pierre est-il venu ?* ('Did Pierre come?')
(3) *Est-ce que c'est vrai?* ('Is that true?')

As far as intonation is concerned, rising tones are seen as the canonical form associated with declarative questions (see, among others, [9] and [10]). By contrast, in *yes-no* questions in which the modality of the utterance is indicated by a morpho-syntactic or a lexical marker (subject-verb inversion or *est-ce que* particle respectively), non-rising tunes may be used on a par with rising ones (see, amongst others, [9], [11] and [12]). Note, however, that the rising tune is by far the most frequently used in information-seeking *yes-no* questions, regardless of the construction (see, amongst others, [13]).

2.2. *Yes-no* questions in English

Contrary to French, in English information-seeking *yes-no* questions, the modality of the utterance is always indicated by morpho-syntactic means: either by subject/auxiliary inversion (4), or by the use of auxiliary *do* (5).

(4) *Is Peter coming?* (*vs. Peter is coming.*)
(5) *Does he live in Paris?* (*vs. He lives in Paris.*)

Declarative questions can be found in English, but they are usually echo-questions, which are not present in our corpus.

As opposed to *wh*-questions, for which the default tone is a fall in English, *yes-no* questions are uttered with a rising tone, even if a falling tone can also be heard, but less frequently (see amongst others [14], [15] and [16]).

3. Corpus and method

3.1. Corpus

3.1.1. The AixOx corpus

The AixOx corpus ([7], [8]) is a multilingual learner corpus which consists of recordings of 40 1-minute passages in English and French read by native speakers and L2 learners (all aged 20-35, see [8] for details about the informants). The passages were extracted from the Eurom 1 corpus ([17]). Non-native speakers were divided into two groups, B and C, according to their level of proficiency in the Common European Framework of Reference for Languages (CEFR). Learners of group B are independent users, B1/B2 in the CEFR and learners of group C are proficient users, C1/C2. Hence the corpus is composed of 6 groups of speakers, as shown in table 1 below. For each group, 10 speakers were recorded (5 females and 5 males), the corpus thus amounting to 60 speakers and about 30 hours of speech.

Table 1. *Speaker groups in AixOx.*

Language of recording	Native speakers	L2 learners B1/B2	L2 learners C1/C2
English	ENEN	FRENB	FRENC
French	FRFR	ENFRB	ENFRC

3.1.2. The extracted questions

Before extracting the questions on which the study is based, we had to take a few facts into account:

- The English and French corpora are not word-to-word translations, even if they are very close, since the pragmatic contexts in which the texts are uttered are similar; hence the English corpus contains 23 questions when the French one contains 22.
- *Yes-no* and *wh*-questions are present in the corpus.
- For the total questions, the French corpus includes different types of constructions, whereas yes-no questions in the English corpus all display subject/auxiliary inversion. In order to compare the data, only questions with an interrogative marker have been taken into account in the present study.
- Some yes-no questions in both languages are not neutral questions since they may be rhetorical questions addressed to oneself (‘*What will 1992 really mean to the person in the street?’*), disguised orders (‘*Can you give me a firm date now?*’) or have the meaning of partial questions (‘*Can you tell me what’s on television tonight?’*).

In order to make a cross-comparison between *yes-no* questions, we took into account the same types of questions, both on the syntactic and pragmatic levels. We therefore concentrated on 5 information seeking *yes-no* questions, in which the modality of the utterance is indicated by morpho-syntactic means. They are listed below (the English and French questions are not necessarily translations, since some of the French questions have been disregarded because of their declarative structure):

- French questions

F1 : Est-ce que vous pourriez me donner leur nouveau numéro de téléphone ?
F2 : Est-ce que c'est vrai ?
F3: Est-ce que vous pourriez me donner la liste des restaurants de mon quartier ?
F4 : Est-ce que vous avez des tarifs spéciaux pour les collectivités ?
F5 : Est-ce qu'un organisme universitaire peut en bénéficier ?

- English questions

E1 Can you give me their new number please?
E2: Could you please tell me the best connections to Sheffield from East Greenstead?
E3: Do you take reservations by telephone?
E4: Can you give me a list of the restaurants in the neighbourhood?
E5: Do you have special corporate academic institutions?

We therefore studied 5 questions uttered by 30 speakers (10 in the three groups – natives and learners from groups B and C - for the 2 languages) in the two languages, the corpus thus amounting to 300 sentences.

3.2. Method for the prosodic analysis

In order to compare the intonation of the extracted questions uttered by the native speakers and by the B and C learners, we used two distinct approaches to encode the tones observed in the various utterances: a perceptual approach, and a semi-automatic approach.

3.2.1. The perceptual approach

Following [18] and the British tradition, the perceptual analysis relies for English on tonality (the division into intonation phrases), tonicity (the place of nuclear syllables) and tones (the distinctive pitch movements).

We consider, following [19], that there is only one level of boundary, associated with the intonation phrase (IP). In short questions, one IP will be realized but longer questions like *Could you please tell me the best connections to Sheffield from East Greenstead?* can be divided into 2 IPs by the natives, and even more by the learners. The IP boundaries are marked by a slash in the paper.

The place of the tonic syllable, the nucleus, is looked at. The principle that there is only one nuclear syllable in an IP is adopted. It is the most prominent one, that bearing the tone (the distinctive pitch movement) of the IP (see for example [20] or [21]). The nuclear syllable is underlined in the given examples.

Finally, the tones are encoded. The tone is the distinctive pitch-movement, that bearing on or starting on the nucleus and extending on the post nuclear syllables (if any). The symbols used are F for a simple fall, R for a simple rise, HF for high fall and FR for fall-rise, again according to the British tradition but with a limited tone inventory, following [22] and [23].

The questions are therefore encoded as follows in (6), (7) and (8), respectively a native speaker, a learner from group B and a learner from group C:
(6) *Can you give me their new <u>num</u>ber, please F /*
(7) *Can you give <u>me</u> F / their <u>new</u> number, please FR /*
(8) *Can you give me their new num<u>ber</u>, please R /*

So as to allow cross-comparison, the French questions were encoded in the same way, as shown in examples (9), (10) and (11), respectively a native speaker and learners from groups B and C. The global contour of the IP was taken into consideration in our study, and not the contours of the accentual phrases.
(9) *Est-ce que vous pourriez me donner leur nouveau numéro de télé<u>phone</u> R /*

(10) *Est-ce que vous pourriez me donner F / leur nouveau numéro de téléphone R /*
(11) *Est-ce que vous pourriez me donner leur nouveau numéro de téléphone R /*

3.2.2. The semi-automatic approach

A semi-automatic approach was also performed. The questions studied were extracted under PRAAT [24], automatically aligned into words with SPPAS [25] and manually labelled into syllables. The SAMPA phonological representation of our syllabification is: k@n / ju / gIv / mi / @ / lIst / @v / D@ / rEst / rQnts / In / D@ / neIb / @ / hUd. The syllable tier of the annotation was used for automatic extractions of acoustic parameters using ProsodyPro [26].

3.3. Revisiting the intonation of English yes-no questions

The first result we obtain concerns the intonation produced by native speakers. Table 2 below shows the results of the perceptual analysis: only the global contours (on the final IP), *i.e.* Falling (F and HF) or Rising (R and FR) of the questions studied are reflected in the table:

Table 2. *Contours for ENEN and FRFR questions*

Questions	English natives	French natives
E1/F 1	F 80%	R 100%
E2/F 2	F 100%	R 90%
E3/F 3	R 70%	F60%
E4/F 4	F 60%	R 90%
E5/F 5	F 80%	R 80%
Total	**F 70%**	**R 80%**

The table clearly shows, contrary to what the literature claims, that the rising tone is not the most frequent for information-seeking *yes-no* questions in English: 70% of the 50 English questions under scrutiny are uttered with a falling tone. For French, our data confirm that the rising tone is the default tone for *yes-no* questions.

3.4. The intonation of learners: global contour and phrasing

Table 3 gives the results for the French learners of English:

Table 3. *Contours for FRENB and FRENC questions*

Questions	FRENB	FRENC
E1	R 90%	R 60%
E2	R 80%	R 70%
E3	R 90%	R 60%
E4	R 80%	F 60%
E5	R 90%	R 80%
Total	**R 86%**	**R 66%**

The learners of both groups massively pronounce the questions on a rising tone, as in their mother tongue, but the tendency is less strong for the proficient group: 44% of the questions are uttered with a fall, which is quite a lot compared to the tendency for the native English speakers.

But the global contour is not the most relevant feature for proficiency. The phrasing is probably the most salient difference between the independent (group B) and the proficient (group C) learners. The independent learners tend to paste the French phrasing to the English sentence: they produce a pitch movement at the end of word groups which would correspond to accentual phrases in French, as exemplified in (12) and (13):

(12) *Can you give me a list R / of the restaurants R / in the neighbourhood R /*
(13) *Could you please R / tell me R / the best connections R / to Sheffield R / from East Greenstead R /*

A clear evolution can be noted for the proficient learners, who no longer group words as in French, but, as a few native speakers do so too, divide E4 in 2 IPs for 8 speakers out of 10 (3 IPs for 1 speaker and 1 IP for 1 speaker) and E2 in 2 IPs for 100% of the speakers, as in the occurrences below:
(14) *Can you give me a list of the restaurants F / in the neighbourhood F /*
(15) *Could you please tell me the best connections to Sheffield F / from East Greenstead R /*

3.5. Multi-speaker modelling

With the semi-automatic approach, it is possible to visualize and compare the curves for the natives and the learners using R [27], as is shown in figure 1below for question E4:

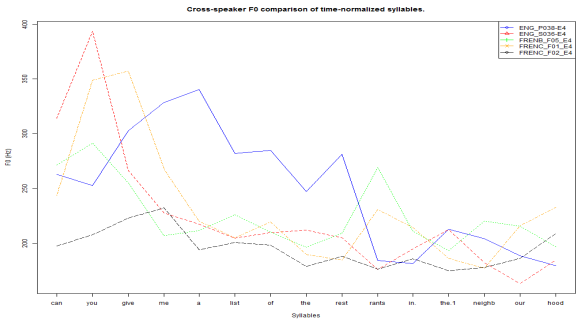


Figure 1. *Comparison between two natives, two C learners and one B learner*

For clarity's sake, the number of speakers represented has been limited and only female speakers have been taken into account in Figure 1. In addition, one should be made particularly aware that the curvature does not represent F0 variation *per se*, but the statistical software interpolations between the points representing the means estimated for each syllable. ProsodyPro offers a finer-grained representation, which is based on 10 successive measures of pitch over the same syllable. This corresponds to a kind of time normalization: the duration of each syllable might be different from one speaker to another, but each tenth of a syllable duration can be compared across speakers.

The nuclear syllables are easily identifiable on the curves. Figure 1 shows that for one native a pitch movement occurs on *RESTaurants* (the blue curve). For one FRENC learner, a peak is also clearly visible on *RESTaurants* (blackline), but the prenuclear part of the curves is quite different from those of the natives. As for the other C learner (yellow curve) and the B learner (green curve), a pitch movement also occurs on the word *restaurant*, but we see that the peaks have moved right, on the final syllable of *restauRANTS*. It is to be noted that the natives and the learners of group C pronounce *restaurants* with 2 syllables while many learners of group B pronounce 3 syllables as in French. The non compression of the median

syllable and the stress shift on the right are typical of a French learner's pronunciation. A closer look at the curves also confirms the perceptual impression that pitch movements take place on words preceding *restaurants*. The pitch contour in Figure 2 is representative of that tendency: peaks on *me* and *list* appear.

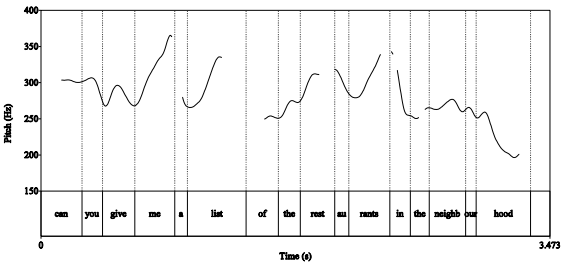


Figure 2: *Pitch track of E4 uttered by a B learner*

4. Discussion and conclusion

With the automatic procedure, cross-sectional comparisons are established on the basis of normalized duration, so that performances are equally considered on the basis of the prosodic target (here, pitch) for each syllable. In that sense, the semi-automatic procedure consists in discourse alignment, not in time-alignment, allowing prosodic realizations to be compared. This simplifies speech time variability and enables the design of "confidence intervals" for native prosodic realizations and potential non-native mismatches. A tentative representation of this modelling lies in the graphing and statistical representation, using R once more. Due to space limitation, we do not include the corresponding boxplot, but annotation at the word level is not so telling and syllable-based analysis is much more convincing. This semi-automatic approach paves the way for the characterization of learner profiles and their interaction with criterial features [28]. We give a rough outline of some of the interlanguage stages that can be detected with this kind of approach. Figures 3, 4 and 5 represent the boxplots synthesizing inter-speaker variation of mean fundamental frequency (computed for each syllable) for E4 realized by respectively the female natives, some of the female learners from group B and from group C. In the figures, the small rectangles show that the dispersion is limited and that there is a consensus on the prominence of the syllable. The large rectangles on the contrary point to a larger variation, *i.e.* some sort of non-consensus. The median (central line in bold) allows the visualization of the melody. If we first compare the median on the three figures, we clearly see that the natives (figure 3) divide the sentence into 2 IPs and favour a falling pattern on the first IP (on *RESTaurants*) and a rising one on the second (*NEIGHBourhood*); learners from group C (figure 4) also divide the sentence into 2 IPs, with similar nuclei as the natives, but realize a rise on both IPs; the less advanced B learners (figure 5) divide the sentence into what corresponds to the French phrasing with rising movements on *me*, *list, restauRANTS* and *neighbourHOOD,* followed by an F0 resetting after the first three words. If we now look at the rectangles, a large dispersion noticeably appears in figures 3 and 4 on tool words like *you*, *me*, *a* and *of*, which are far less important prosodically, as opposed to the stressed syllables of *restaurants* or *neighbourhood* for example, which display small rectangles (in figure 4 even more strikingly): the natives and the C learners therefore show strong agreement as to which syllable they should make prominent. The prenuclear contour, however, is still somewhat hesitant for the C learners (the dispersion is very high on *you* and *give*). Finally, the rectangles for B learners (figure 5) show much less difference in the dispersion, this reflecting isosyllabic realizations, typical of French rhythm.

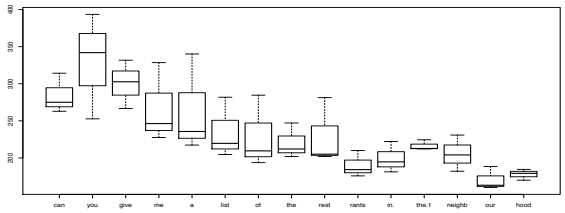


Figure 3. *Native female speakers*

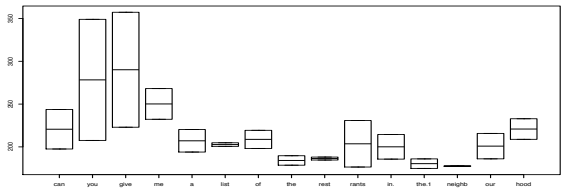


Figure 4. *Female learners from group C*

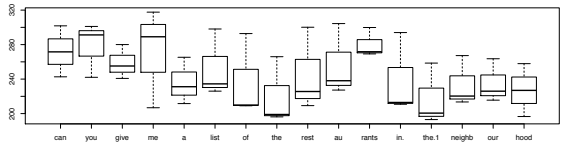


Figure 5. *Female learners from group B*

As explained in [29], there are limitations to this methodology (resyllabification issues, neutralization of rhythmic variation, pitch detection errors), but one consequence is for sure: modelling interlanguage intonation like this means that syllable division is high on the agenda for this kind of spoken learner corpus research.

The comparison of the productions of information-seeking *yes-no* questions by natives and learners showed that the form of the nuclear contour is not so much affected. The most important differences concern rhythm, and the prenuclear syllables. French speakers of English, in particular at B level, have a tendency to assign a rising pitch movement at the end of prosodic words, which leads to a clear difference in rhythm. The study of the intonation of the questions by English learners of French (which could not be developed in this paper for lack of space) also shows that the nuclear contour is somehow well realized too and it is to be noted that it seems easier for English learners of French to utter questions than for French learners of English, probably because the prenuclear syllables are better realized. Most studies on L2 intonation focus on the phonological nuclear form. The present paper encourages further study on the phonetic implementation of the prenuclear contour.

5. References

[1] Mennen, I., "Bi-directional interference in the intonation of Dutch speakers of Greek", Journal of Phonetics 32, 543–563, 2004.

[2] Jilka, M., The Contribution of Intonation to the Perception of Foreign Accent, Doctoral dissertation, University of Stuttgart, 2000.

[3] Rasier, L., & Hiligsmann, P., "Prosodic transfer from L1 to L2. Theoretical and methodological issues", Nouveaux cahiers de linguistique française 28, 41-66, 2007.

[4] Mennen, I., "Phonological and phonetic influences in non-native intonation", in J. Trouvain & U. Gut [eds.], Non-native Prosody: Phonetic Descriptions and Teaching Practice, pp. 53–76, Mouton De Gruyter, 2007.

[5] Ramírez, D. & J. Romero, "The pragmatic function of intonation in L2 discourse: English tag questions used by Spanish speakers", Intercultural Pragmatics 2 (2), 151-168, 2005.

[6] Mennen, I., "Beyond segments: towards an L2 intonation learning theory (LILT)", in Delais-Roussarie, E., Avanzi, M. & S. Herment [eds.], Prosody and languages in contact: L2 acquisition, attrition, languages in multilingual situations, Springer Verlag, accepted, to appear.

[7] Herment, S., Loukina, A. & A. Tortel, "The AixOx corpus", SLDR, 2012. http://sldr.org/sldr000784/fr

[8] Herment, S., Tortel, A., Bigi, B., Hirst, D. & A. Loukina, "AixOx, a multi-layered learners' corpus: automatic annotation", in Díaz Pérez, J. and A. Díaz Negrillo, [eds.], Specialisation and variation in language corpora, Bern: Peter Lang, to appear.

[9] Delattre, P., "Les Dix Intonations de base du français", The French Review 40 (1), 1-14, 1966.

[10] Di Cristo, D., "Intonation in French", in Hirst, D. & A. Di Cristo [eds.], Intonation systems: A survey of twenty languages, 195-218, Cambridge: Cambridge University Press, 1998.

[11] Martin, P., "Une grammaire de l'intonation de la phrase française 2", Rapport d'Activité de l'institut de phonétique 9/2, pp. 77-96, Institut de Phonétique de l'Université Libre de Bruxelles, 1975.

[12] Martin, P., Intonation du français, Paris: Armand Colin, 2009.

[13] Santiago-Vargas F. & Delais-Roussarie, E., "Acquiring phrasing and intonation in French as a second Language: the case of Yes-No questions produced by Mexican Spanish Learners", Proceedings of Speech Prosody, Shanghai, China, 2012.

[14] Cruttenden, A., Intonation, Cambridge: Cambridge University Press, 2nd ed. 1997 (1st ed. 1986).

[15] Wells, JC., English Intonation, an Introduction, Cambridge: Cambridge University Press, 2006.

[16] Roach, P., English Phonetics and Phonology, A practical course, Cambridge: Cambridge University Press, 4th ed. 2009 (1st ed. 1983).

[17] Chan, D., Fourcin, A., Gibbon, D., Grandström, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Linderg, B., Moreno, A., Mouropoulos, J., Senia, F., Transcoso, I., Velt, C. & J. Zeiliger, "EUROM – A Spoken Language Resource for the EU", Proceedings of Eurospeech '95, (Madrid) 1, 867-880, 1995.

[18] Halliday, M.A.K., Intonation and Grammar in British English, The Hague-Paris: Mouton, 1967.

[19] Grabe, E., Post, B. & F. Nolan, "Modelling intonational variation in English. The IViE system", in Puppel, S. & G. Demenko [eds.], Proceedings of Prosody 2000, Adam Mickiewicz University, Poznan, Poland, 2001.

[20] Tench, P., The Intonation Systems of English, London: Cassell, 1996.

[21] Halliday, M.A.K. & W.S. Greaves, Intonation in the Grammar of English, London-Oakville: Equinox, 2008.

[22] Hirst, D., "Form and function in the representation of speech prosody", in Hirose, K., Hirst, D. & Y. Sagisaka [eds.] Quantitative prosody modeling for natural speech description and generation (Speech Communication 46 (3-4)), 334-347, 2005.

[23] Gussenhoven, C., On the grammar and semantics of sentence accent, Dordrecht: Foris, 1984.

[24] Boersma, P. & D. Weenink, "Praat, a system for doing phonetics by computer", Glot International, 5:9/10:341-345, 2001.

[25] Bigi, B., "SPPAS: a tool for the phonetic segmentation of speech", Proceedings of the Language Resource and Evaluation Conference, Istambul, Turkey, 1748-1755, 2012.

[26] Xu, Y., "ProsodyPro, A Tool for Large-scale Systematic Prosody Analysis", Proceedings of the TRASP conference, Aix-en-Provence, France, 2013.

[27] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org

[28] Cauvin, E. "Intonational phrasing as a potential indicator for establishing prosodic learner profiles", In Granger, S., Gilquin, G. & F. Meunier [eds.], Twenty Years of Learner Corpus Research: Looking back, Moving ahead. Proceedings of Corpora and Language in Use, Louvain-la-Neuve: Presses universitaires de Louvain, 75-88, 2013.

[29] Ballier, N. & P. Martin, "Developing corpus interoperability for phonetic investigation of learner corpora" in Díaz-Negrillo, A., Ballier, N. & P. Thompson [eds.], Automatic Treatment and Analysis of Learner Corpus Data, Amsterdam: Benjamins, 33–64, 2013.

Avoidance of Stress Clash in Perception of Conversational American English

Amelia E. Kimball, Jennifer Cole
Department of Linguistics and Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign, Urbana, Illinois, United States

akimbal2@illinois.edu; jscole@illinois.edu

Abstract

We examine evidence for a regularity bias in the perception of sentence-level stress patterns, asking to what degree listeners perceive speech as *metrically regular*, with few or no occurrences of stress clash. We assess regularity through a stress perception task carried out by untrained listeners annotating transcripts of recorded speech, with sentences designed to have regular stress, and sentences drawn from a corpus of spontaneous conversational speech. Results show listeners report perceiving fewer stress clashes than predicted by random placement of stresses or by concatenating the citation form stress patterns of each individual word in a given sentence, though some incidence of stress clash is reported for both the regular and irregular speech materials. These findings suggest that listeners perceive English speech in accordance with a weak regularity bias. Inter-transcriber agreement rates also reveal substantial disagreement in perceived stress patterns at the sentence level, for regular and irregular sentences alike, suggesting variability in the perception of acoustic cues to stress at these levels.

Index Terms: stress clash, meter, stress perception, rhythm, metrical regularity

1. Introduction

Metrical patterns in speech arise from the sequencing of stressed and unstressed syllables across words and phrases [1]. In English, regular patterns occur when words are sequenced such that stressed syllables occur at regular intervals. For instance, the sentence 'HEIdi SOMEtimes SAW the JUry LEAving' has the stress pattern [SW SW S W SW SW], with a recurring pattern of stressed (strong) and unstressed (weak) syllables. However, due to the frequency of mono-syllabic words and the variety of patterns of word-level stress in English, metrically irregular phrases and sentences are also very common, e.g., 'JILL LIKES to SKI CAREfully' with the stress pattern [S S W S SWW]. To avoid confusion with notions of isochrony based on acoustic duration (discussed below), here we use the terms regular and irregular to refer to *metrical patterns of phonological stress* in sentences.

Despite the fact that sentences and phrases in English are not necessarily—or even typically—regular in their stress patterning, there is evidence that listeners are biased to perceive stress in terms of such regular patterns, and that more generally, regular stress patterns are privileged in speech processing. Early evidence for a regularity bias comes from studies of English phrasal stress. In phrases where stress clash results from the sequencing of word-level stresses, (e.g. *thirTEEN MEN*), speakers have the option of resolving the clash in favor of an alternating stress pattern (*THIRteen MEN),* or a pattern with only a single stress (*thirteen MEN*). When asked to identify the stressed syllables in such instances, listeners report hearing the alternating, clash-free pattern (THIRteen MEN) when listening to the entire intact sentence [2,3], but do not reliably perceive the resolved stress (THIRteen or thirteen) in the first word in the sequence when it is extracted and presented by itself [3]. These findings suggest that English speakers are biased to perceive stress patterns in phrases or sentences as regular, even when the acoustic evidence is not particularly strong.

The bias for listeners to perceive phrasal stress as regular may reflect a more general bias for regular stress patterns in speech processing. Studies on the perceptual processing of speech show that sentences with regular stress patterns yield faster and more accurate phoneme and word recognition [4-6]. In addition, ERP studies have shown that regularity modulates the amplitude of the n400 response [7], suggesting that speech perception and semantic integration are made easier by predictable, alternating stress patterns. In speech production, strings of non-words with regular stress patterns are easier to produce than irregularly patterned strings of the same non-words [8]. These experimental results point to a basis for a regularity bias in the mechanisms of speech processing, e.g., in neural oscillatory processes, as claimed by [9].

The studies cited above investigated stress regularity in speech production and perception with experimenter-controlled, read speech materials. This leaves us to wonder about the production and perception of speech that is produced under more natural conditions, e.g., conversational speech. This paper represents an initial step in the investigation of the effects of stress regularity in everyday speech, with an experimental approach combining experimenter-designed sentences read aloud by a model speaker with speech samples from a corpus of conversational speech that are re-enacted by the same speaker. Our focus in this paper is on **perceived** stress patterns. The goal is to compare the observed patterning of listeners' reported stress perception to random placement of stresses and to word stresses as reported in the dictionary. If there is a bias towards regularity, we expect observed patterns will be more regular than predicted patterns.

Note that in this paper metrical regularity is defined as an alternating pattern of strong and weak syllables. This is distinct from temporal measures of regularity as defined over acoustic intervals [e.g., 10,11]. In other words, in this paper we are interested in whether listeners report adjacent stressed syllables, regardless of when these syllables happen in time. Acoustic measures of our sample are not within the scope of the present paper, but are the subject of our ongoing investigation.

2. Experiment

This experiment tests the hypothesis that listeners' perception of stress patterns in naturally occurring sentences of English will be biased towards regular patterns. Specifically, we predict that listeners will report fewer instances of stress clash (stress on adjacent syllables) than are expected based on the location of stress within each content word, and also fewer than expected by three other calculations of chance occurrence, described further below. Listeners' perception of stress is assessed through a beat annotation task presented in a web survey format.

2.1. Stimuli

The test materials consisted of twenty sentences of conversational speech from the Buckeye Corpus of Conversational American English [12]. Also, twenty

sentences designed to have regular stress patterns, selected from previous experiments by the first author and from published studies, were included in this experiment as a control condition where no stress clash is expected to be perceived. To minimize the effects of speaker-dependent variability in speech rate, in patterns of phonetic reduction, or in other aspects of the phonetic realization of stress, all speech materials used in this study were re-enacted by a model speaker who is a native speaker of American English trained in linguistics, but who had no knowledge of the research goals of this study. A text transcript of the speech excerpts was presented to the model speaker, and the speaker was instructed to repeat the utterances in a natural and conversational style.

The form of the Buckeye corpus is informal sociolinguistic interviews with 40 speakers in the Columbus, Ohio area. Buckeye sentences were taken from interviews with four different speakers, chosen randomly from the larger corpus. Sentences were selected by the first author and chosen for the absence of disfluencies or major internal prosodic phrase breaks. For each of the four interviews the first five prosodically-demarcated utterances of the target duration and with no significant internal prosodic breaks were taken. Excerpts were approximately 1-2 seconds in duration. The 20 Buckeye excerpts taken together consisted of 159 words with a total of 205 syllables. The sentences with regular stress patterning (prepared by the experimenter) consisted of 140 words with a total of 198 syllables.

The regular sentences conformed to three metrical patterns: trochaic (SWSW…), iambic (WSWS…), or dactylic (SWWSWW…). The model speaker read all sentences of one stress pattern together. Examples of each sentence type are listed in table 1 below

Buckeye	My grandmother's from Ireland I go to Northland high school right now
Trochaic	Read a bedtime story. Heidi sometimes saw the jury leaving.
Iambic	Michelle foresees mistakes. My shoes are beige and black.
Dactylic	Sally is hoping to travel to Canada. Thomas has already taken geography.

Table 1. *Stress patterns for sample sentences*

2.2. Participants

55 participants total (31=Female) were recruited on Amazon Mechanical Turk, an online marketplace for human intelligence tasks. Participants' age ranged from 19 to 55 (mean =32.5, s.d.=8.8). Built-in Mechanical Turk screening tools ensured that the posting was displayed only to those workers who reported their location as in the United States. Only data from the 48 native English speakers with no reported hearing problems were eligible for inclusion in this study.

2.3. Procedure

Participants were presented with a display via an online survey built with Qualtrics survey tools [13]. Instructions stated that they would listen to a series of sentences and should "mark the beats" in a sentence by checking boxes. They listened to an example sentence ("I like to run and jump") and wrote the last word of that sentence in an answer box, in order to confirm that their audio was working. All participants correctly identified the last word in the example sentence. Participants were then shown an example of checked boxes for that sentence, and told "this person thinks the beats are on *like, run*, and *jump*." An example of the user interface is below in Figure 1.

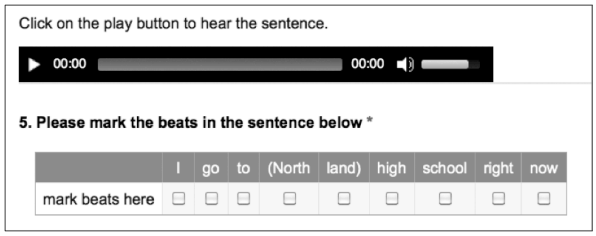


Figure 1: *User interface for the experiment.*

Participants were explicitly told to mark beats as the speaker said them in the audio recording, and not how the participant would say them. They were told that if they were unsure, they should "make their best guess", and that experimenters expected that listeners "may differ in where they mark beats." The listeners could play the audio as many times as they wanted, and were required to check at least one box. They were told that if they could not hear the audio they should check all the boxes. Two catch questions were presented in which the speech was purposefully obscured. Data was analyzed only from the 46 participants who (in addition to reporting as native speakers with no hearing problems) followed directions and identified the catch questions by marking all boxes.

3. Results

This experiment tested the hypothesis that listeners have a regularity bias in the perception of stress in conversational speech. If this hypothesis is true, we expect that when asked to mark stressed syllables in a corpus, listeners will report fewer clashes than predicted by chance or than are predicted by the concatenated stresses from the dictionary entry for each individual word. Results showed 7 out of 8 predictions of clash frequency were greater than the observed rate of clash, meaning that listeners hear fewer clashes than predicted.

3.1. Clash measures

For both the regular and Buckeye sentences five different frequency counts of the number of clashes were conducted: the observed number of clashes, plus four measures of the expected number of clashes.

Observed clashes were calculated based on participant responses. For every n sequential checkmarks, n-1 clashes were counted, such that four beats in a row would be marked as three clashes. This same counting method was used for the expected clashes.

The first rate of expected clashes was calculated based on the rate at which an individual participant marked syllables as beats for each sentence. For example, if a participant marked 2 beats out of every 4 syllables in a given sentence, their rate for that sentence was 2/4. This rate was squared to get the probability that two adjacent syllables would be marked as beats (a clash) and then the rate was multiplied by the number of adjacent syllable pairs in a given sentence, which represents the number of clashes possible for that sentence. Equation 1 below shows the method. Where C is the number of checks and N is the number of syllables:

$$\left(\frac{C}{N}\right)^2 (N-1) = \text{Number of expected clashes} \quad (1)$$

The second measure of expected clashes was created through a random sampling simulation using the R statistical computing language [14]. Separate simulations were run for dummy sentences with the same number of syllables as the test sentences, and for all possible numbers of beat marks within those sentences. 10,000 trails were run for each sentence length. Each trial looped through the total number of syllables, randomly selecting either a check or no check for each syllable, without replacement. Clashes were counted for each trial, and a mean clash occurrence was calculated across trials and compared with observed values.

The third measure of expected clashes was the number of clashes predicted by the dictionary entry of the citation form of the word. This was done by marking the primary stress of each word as reported in the Oxford English Dictionary and then counting occurrences of adjacent stressed syllables. Though this method was expected to overestimate because of the preponderance of monosyllabic words, it was included nonetheless because all words, including function words, may be variably stressed in conversational speech.

The last measure of expected clashes was also calculated based on the stresses marked in the dictionary for each word, but this time marking only the stressed syllables of content words as beats, with no beats marked on function words.

Figure 2 below compares the total clashes counted by each measure across sentences. Clashes are reported in clashes per syllable, to normalize for the differing number of syllables in the two samples.

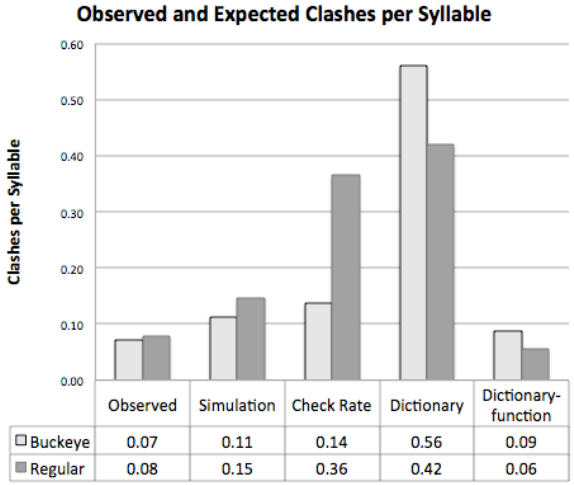


Figure 2: *Observed vs. expected clashes per syllable in both regular and conversational sentences*

Paired sample t-tests comparing the observed number of clashes vs. expected number of clashes under each calculation showed that for both samples the observed number of clashes was statistically significantly smaller than the expected number of clashes based on participant check rate, simulation, or the dictionary with and without stressing function words. The dictionary prediction that includes function words proved to be an especially poor model, grossly overestimating the amount of observed clashes in both regular and conversational sentences. However, the expected values based on dictionary stress markings *not including function words* provided the closest approximation of participants' responses, though it still differed significantly from the observed clash rate—in the case of the *regular* sentences, dictionary stress without function words predicted *fewer* clashes than observed, in the case of *Buckeye* sentences dictionary stress without function words predicted *more* clashes than observed. Table 2 below lists the *t* statistic, degrees of freedom, mean of the differences, and *p* value for each measure of expected clash frequency as compared to observed clash frequency.

REGULAR	t	df	mean of the differences	p
Check rate	-54.4861	919	-2.827624	<.001
Simulation	-24.1348	919	-0.6640634	<.001
Dictionary	-28.7383	919	-3.365217	<.001
Dictionary- function	3.9725	919	0.2347826	<.001

BUCKEYE	t	df	mean of the differences	p
Check rate	-28.0077	919	-0.6749942	<.001
Simulation	-18.1389	919	-0.4359521	<.001
Dictionary	-48.3472	919	-5.026087	<.001
Dictionary- function	-3.532	919	-0.176087	<.001

Table 2. *Results of paired sample t-tests comparing observed and expected clash rates. Negative mean differences indicate that expected measures were higher than observed.*

3.2. Agreement

In addition to a comparison of observed clashes to expected clashes, we also calculated the inter-transcriber agreement for each syllable. Each syllable received a rating based on the number of listeners that marked it as a beat. Agreement scores are proportions ranging from 0 to 1. If listeners were in total agreement, the distribution of these scores would be bimodal, with peaks at 0 (meaning many syllables were marked by no listeners) and 1(meaning many syllables were marked by all listeners). Instead, as is clear from figure 3 below, the distribution was spread within the conversational Buckeye sentences. Relatively few syllables were marked by a majority of listeners, though many syllables were left unchecked by all listeners.

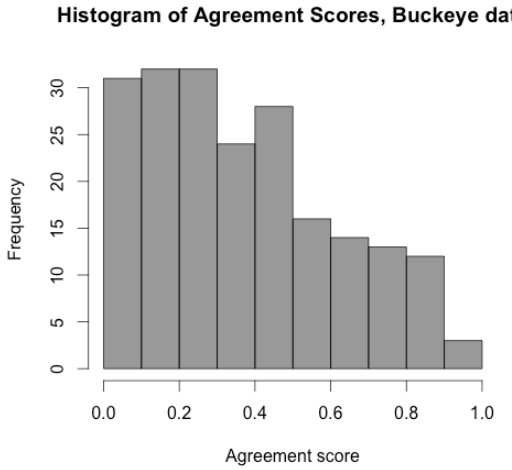


Figure 3: *A histogram of the distribution of agreement scores in the Buckeye sentences.*

It was expected that stress patterning would be more salient in the regular sentences, making them easier for listeners to annotate, and yielding higher levels of inter-transcriber agreement. Contrary to this expectation, we find that in both samples listeners fail to agree on the transcription for a majority of syllables.

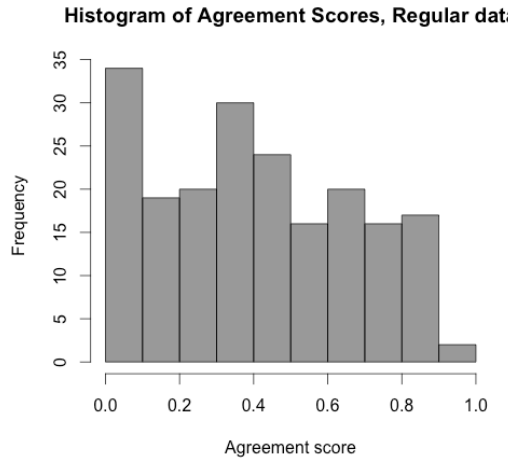


Figure 4: *A histogram of the distribution of agreement scores in the Regular sentences.*

4. Discussion

Given that metrically regular speech is advantaged in production and perception, we predicted that listeners would hear speech as regular and report few clashes. Our results bear this out—participants mark significantly fewer clashes than would be expected if syllables were randomly marked as beats, and also fewer than are expected based on the dictionary stress of the citation form of the words. This finding is consistent with the experimental hypothesis, suggesting the influence of a regularity bias in the perception of sentence-level stress. However, despite this apparent bias participants *do* report stress clashes, even in the regular sentences which were designed to have no clashes. This finding was surprising and points to the difficulty English speakers have in identifying and reporting lexical stress. Overall, these results support a 'soft' regularity bias [15,16], which favors regular alternating stress patterns at the sentence level, but which also allows for deviation from the preferred regular pattern in speech perception (and, we hypothesize, in speech production).

An important finding of this study is that listeners do not achieve a high level of agreement in their responses. Phonological theory which projects stress from accented syllables to the phrase level [1] would lead us to assume that patterns of word stresses are straightforwardly determined by the words of the utterance. However, our results show that individual listeners report different stress placement after being exposed to the same acoustic stimuli.

Some of this variation my be due to listeners defining "beat" differently, or interpreting the task somewhat differently. Then too the current results do not address individual differences in listeners, who varied in age (and no doubt in various cognitive measures). However, we believe that despite these potential sources of noise, the variability in the reported results point to variability in the perception of the acoustic cues to stress at the sentence-level. If listeners were uniform in their stress perception, we would expect high levels of agreement when a group of listeners were presented with the same acoustic stimuli, designed to be metrically regular. Our results are not consistent with this prediction.

The tendency to minimize clash in the perceived pattern of stresses in a sentence is shown to be all the more robust when inter-transcriber agreement in the marking of stress beats is taken into consideration. Though listeners did not agree in their placement of "beats," they nonetheless as a group avoided marking clashes. The low incidence of reported stress clash, together with the low rates of agreement in the marking of stress beats suggests a regularity bias in perception. These results motivate a thorough investigation of acoustic correlates of stress in these stimuli as predictors of listener responses, and individual differences in stress perception. This investigation is currently underway.

5. Conclusions

There are three main conclusions of this study. First, the comparison of observed rates of perceived stress clash with expected rates suggests a regularity bias in the perception of sentence-level stress, in that listeners report perceiving fewer clashes than would be expected. Second, despite the fact that listeners as a group report fewer stress clashes than expected, individual listeners disagree on stress placement for a given sentence, suggesting that the perception of stress may vary from listener to listener. Lastly, this study shows that concatenating citation form stress as marked in a dictionary provides a poor model of listeners' perception of sentence-level stress. Though citation stress without function words is a better model of perceived stress patterns, it still differs significantly from listeners' reported perception.

Further research is called for to determine which measure of stress is most accurate as a representation of stress as produced by a speaker, or as perceived by a listener, and whether the two measures converge on a common stress pattern. Finally, we note that inter-transcriber variability in the perception of stress beats, as reported here, is similar to the variability in pitch accent perception reported in studies of prosodic transcription [e.g., 17]. This parallel is expected if stress beats at the sentence level are equated with prominence-lending pitch accents in prosodic transcription systems.

6. Acknowledgements

Thanks are due to Cody T. Johnson for data collection and analysis, and Evangeline Reynolds for simulation coding. This project is supported by a Cognitive Science and Artificial Intelligence Award to the first author from the Beckman Center for Advanced Science and Technology at the University of Illinois at Urbana-Champaign. The second author's contribution was supported by NSF BCS 12-51343.

7. References

[1] Selkirk, E. 0. *Phonology and syntax: the relation between sound and structure*. Cambridge, Mass.: MIT Press. 1984.

[2] Vogel, I., Bunnell, T, and Hoskins, S. "The Phonology and Phonetics of the Rhythm Rule." In B. Connell and A. Arvaniti [Ed.], *Papers in Laboratory Phonology IV*, Cambridge: University of Cambridge Press. 1995.

[3] Grabe, E. and Warren, P. "Stress Shift: Do Speakers Do It or Do Listeners Hear It?", In B. Connell and A. Arvaniti [Ed.], *Papers in Laboratory Phonology IV*, Cambridge: University of Cambridge Press. 1995.

[4] Zheng, X., and Pierrehumbert, J, "The Effects of Prosodic Prominence and Serial Position on Duration Perception." *Journal of the Acoustical Society of America* 128 (2): 851. 2011 doi:10.1121/1.3455796.

[5] Quené, H., and Port, R.F. "Effects of Timing Regularity and Metrical Expectancy on Spoken-word Perception.", *Phonetica,* 62 (1): 1–13. 2005.

[6] Brown, M., Salverda, A. P., Dilley, L. C., Tanenhaus, M. K "Metrical expectations from preceding prosody influence spoken word recognition." Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX. 2012.

[7] Rothermich, K., Schmidt-Kassow, M.,and Kotz, S. A. "Rhythm's gonna get you: Regular meter facilitates semantic sentence processing." *Neuropsychologia*, *50*(2), 232–244. 2012. doi:10.1016/j.neuropsychologia.2011.10.025

[8] Tilsen, S. "Metrical Regularity Facilitates Speech Planning and Production." *Laboratory Phonology* 2 (1) Jan. 2011. doi:10.1515/labphon.2011.006. http://www.degruyter.com/view/j/labphon.2011.2.issue-1/labphon.2011.006/labphon.2011.006.xml.

[9] Peelle, J. E., and Davis, M.H. "Neural Oscillations Carry Speech Rhythm through to Comprehension." *Frontiers in Psychology* 3. 2012. doi:10.3389/fpsyg.2012.00320.

[10] Dauer, R.M. "Stress-timing and Syllable-timing Reanalyzed." *Journal of Phonetics* 11 (1): 51–62. 1983.

[11] Low, L.E., Grabe,E. and Nolan,F. "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English." *Language and Speech.* Dec. 2000

[12] Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler- Lussier, E. (2007) Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

[13] Qualtrics survey software. www.Qualtrics.com

[14] R Core Team, "R: A Language and Environment for Statistical Computing" Vienna, Austria, 2013. http://www.R-project.org

[15] Beckman, M.E. "Evidence for Speech Rhythms Across Languages." In Y. Tohura, E. Vatikiotis-Bateson, and Y. Sagisaka [Eds.] *Speech Perception, Production and Linguistic Structure*, 457–63. Tokyo: IOS Press. 1992.

[16] Laver, J. *Principles of Phonetics*. Cambridge: Cambridge University Press. 1994.

[17] Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. "Evaluation of prosodic transcription labeling reliability in the ToBI framework." In Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 123-126, 1994.

Transitions, pauses and overlaps: Temporal characteristics of turn-taking in Czech

Lenka Weingartová, Eliška Churaňová, Pavel Šturm

Institute of Phonetics, Charles University in Prague, Czech Republic

lenka.weingartova@ff.cuni.cz

Abstract

This study aims to describe temporal characteristics of pausing and turn-taking phenomena in conversation. The material comes from the VASST corpus of contemporary Czech and uses four spontaneous dialogues in the form of an informal interview. We describe both general and idiosyncratic effects found in our data and compare them with results from other languages. In our material, transitions with a silent gap, overlaps and back-channels all display notably similar temporal distributions with the median around 360 ms and a marked skewing. The four dialogues did not differ in the proportion of turns belonging to the interviewer (58 %) vs. interviewee (42 %), which is hypothesized to characterize the experimental task. Despite a number of general tendencies, individual differences in pausing and turn-taking behaviour of the speakers were found as well. For instance, the ratio of pauses and gap transitions proved to be highly dialogue-specific. We also gathered evidence for a substantial change in the speech behaviour of the interviewer resulting from a change of her communication partner.

Index Terms: conversation, turn-taking, transition, overlap, back-channelling, pause

1. Introduction

The structure of conversation and turn-taking has been systematically investigated since the 1970s. A pioneering study in this regard is that of Harvey Sacks and his colleagues [1], who introduced an early model attempting to describe and explain the organization of turn-taking in natural conversation. A key assumption is that the structure of syntactic and semantic units should allow the listener to anticipate the end of the speaker's turn and to take over at a *transition-relevance place* (TRP). In addition to these syntactic cues, however, prosodic features (intonation contour, final lengthening, loudness) also play an important role [2], [3], [4].

The model in [1] takes into account several further assumptions that are supposed to describe any modifications in the organization of conversation. One of these is the temporal principle of "minimizing gap and overlap", i.e. of making the transition of speakers as smooth as possible. Several studies (e.g. [5], [6]) challenged this principle, providing a new set of data that yielded the most frequent pause duration at speaker transitions approximating 200 milliseconds, while gaps or overlaps of less than 10 ms were scarce. Thus the close succession of turns remains to be further investigated [6].

The distribution of *gaps* (pauses at turn transitions) and *overlaps* (intervals with overlapping speech at transitions) has been widely investigated in English and a few other languages. In the material of [6] overlaps appeared in approximately 40 % of all turn transitions. The importance of overlaps was also emphasized by ten Bosch et al. [7] and Shriberg et al. [8]. The latter, investigating multi-participant dialogues, found overlaps in 17 % of all words, and in 54 % of intonation units without a pause. Similarly, in an analysis of 26 meetings featuring several participants, overlaps took up 12 % of all speaking time [9]. These findings imply that overlaps represent a relevant feature in the organization of conversation, and it is thus necessary to include them in any model of turn-taking.

Naturally, pauses (not only) at turn transitions have been the focus of much research. Discontinuities in speech may be classified in several ways based, for instance, on their function in conversation (pause, gap, lapse; [1], [6]), their form (silent vs. filled pauses, different forms of hesitations; [10], [11], [12]), their relation to syntax (boundary vs. hesitation pauses; [13]), their duration [14], [15] or on the speaker's intention to continue or pass the word [11], [16]. This also influences the used terminology which is not unified. The issue of what constitutes a pause is also debated. Usually, the authors determine a minimum cut-off boundary in the range of 100-150 milliseconds [4], [17], [18], [19], [20], but other decisions are possible, e.g. [13], [14], [21]. The detection threshold for a pause in turn transitions was determined to be 120 milliseconds [22]. Nevertheless, it is important to keep in mind that decisions on both pause type and minimum pause duration always depend on the purpose of the analysis.

One of the central concerns of this study is the relation between turn transitions and the duration of pauses. This was investigated for instance by Wennerstrom and Siegel [4]. Their results suggest that the probability of a turn transition decreases within the first 500 milliseconds of the pause, and then increases for longer pause durations (approximately 1500 milliseconds). The probability of giving the floor to another speaker is thus highest for short and long pauses, while lowest for pauses of middle duration – which, in other words, tend to occur within a single turn.

Both duration and occurrence of pauses demonstrate a great amount of variability in different languages, speech styles and speech tempos. Studies also dealt with a strong influence of individual habits on the duration of pauses [23], [24]. However, it was ascertained that pauses are connected to major syntactic breaks and coincide with prosodic boundaries of intonational phrases [18], [25]. The duration of the pause has been shown to correlate with the strength of prosodic boundaries [26] and the length of the phrase (e.g. [27]).

We may hypothesize that the organization of conversation is to a certain degree universal and independent of the particular language. To our knowledge, none of the above mentioned experiments have been replicated in Czech. The present paper therefore aims to compare the findings from well-investigated languages with our data on Czech, but also to enlarge the scope of interest and explore idiosyncratic patterns of speakers and accommodation towards the dialogue partner.

2. Method

2.1. VASST corpus

The speech material was taken from the VASST corpus (the acronym meaning *group and style variation* in Czech, see [28]), which is currently being built at the Institute of Phonetics in Prague. The aim of the corpus is to capture the variability of contemporary spoken Czech in different styles and sociolinguistic groups. To this day, the corpus comprises 168 speakers from eight different regions and three specific social groups. Each subject provided five different speaking styles (ordered by formality): a read list of sentences, read continuous text, picture description, controlled interview and spontaneous dialogue.

For this study, the last part of the corpus – spontaneous dialogue – was used. It should be noted that the level of spontaneity of the dialogues is subject to discussion and depends on several factors (familiarity with the interviewer, character of the subject, position within the dialogue, etc.). The fact of being recorded and interviewed presented a rather unnatural situation for the subjects, however it was done in a familiar environment (at the participants' homes) and the experimenters were instructed to make the subjects as comfortable as possible. It could be stated that the recordings we selected from the corpus show a high degree of spontaneity.

2.2. Speech material

Four dialogues were analyzed with a total duration of 85 minutes (on average 21 minutes per speaker pair). The speakers were female, aged 75 to 85, who all came from the same region in Northern Bohemia and spoke colloquial Czech. The experimenter (i.e. their dialogue partner) was in all cases the same, a female student of Phonetics.

The utterances were recorded on a portable professional device Edirol HR-09, with a sampling frequency of 48 kHz and a 16-bit quantization. Afterwards, the recordings were downsampled to 32 kHz and manually post-processed and labelled in Praat [29] by experienced phoneticians (including two of the authors). The boundaries of breath-groups were marked for both dialogue partners, as well as pauses and turn transitions. A breath-group was defined as a stretch of speech of one speaker between his two breath intakes. A turn is the stretch of speech (consisting of one or more breath-groups) of one speaker uninterrupted by the second (with the exclusion of back-channels, see below). Another solution was adopted by [30] and [31] whose main analysis unit was the interpausal stretch. However, this information is obtainable from our annotation as well.

Pauses within speakers' turns were categorized as follows:

- *silent pauses* (Ps): unfilled pause
- *hesitations* (Ph): pause containing a hesitation sound
- *breath pauses* (Pb): pause containing breath intake

The minimum duration of a pause was set to 120 milliseconds; shorter pauses were treated as part of segmental articulation. Pauses that partake in speaker change (i.e. gaps) were not included in this category, see below.

For labelling the turn-taking phenomena, a classification based on [5] was used:

- *gap transitions* (Tg): speakers switch their turns following a gap
- *overlaps* (To): speakers switch by overlapping each other's speech
- *back-channels* (Tbc): intervals of short overlap not resulting in speaker change; often consisting of a single sound or word

It was shown that this classification can cover as much as 96 % of all turn transition phenomena [5].

In the next step, the duration of turns, pauses and turn transitions was measured. Since the duration of transitions or pauses cannot be expected to be normally distributed (the values are only positive and the number of short and long durations is bound to be extremely different), medians instead of arithmetic means are reported. To assess the significance of the results, non-parametrical statistic tests, i.e. Mann-Whitney U test and Kruskal-Wallis one-way ANOVA, were used.

3. Results

3.1. Transitions and overlaps

Gap transitions (Tg) represent the most frequent type of turn-taking in our material. They were over two times more

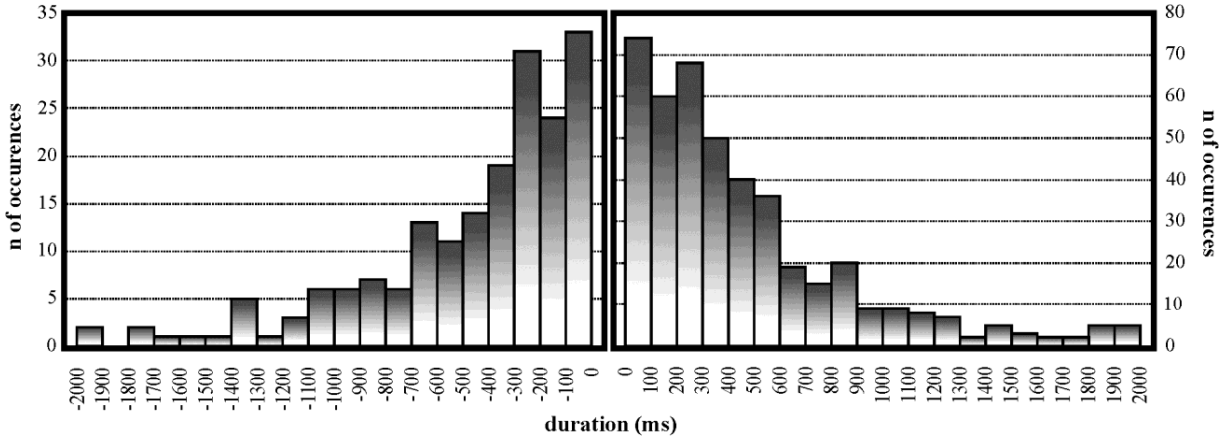


Figure 1: *A histogram of the durations of* OVERLAPS *(on the left) and* GAP TRANSITIONS *(on the right).*

frequent than overlaps (To), and this tendency holds true for individual dialogues as well, although in one of them the ratio amounted almost to 3:1. However, we found a discrepancy between the subjects and the experimenter: the Tg/To ratio was only 1.4 for the subjects, while it was as high as 4.6 for the experimenter (who was instructed not to interrupt the subjects if possible). In terms of percentage, gap transitions constitute 52 % of all turn-taking phenomena, whereas overlaps comprise 22 %.

Figure 1 shows the durational distribution of gap transitions and overlaps. Although both phenomena are inherently different in nature, they display strikingly similar temporal characteristics. The medians for Tg and To are 333 ms and 353 ms, respectively, and both distributions are massively skewed. 57 % of Tg's and 57.5 % of To's are shorter than 400 ms, while 89 % of Tg's and 88 % of To's are shorter than 1 s.

If we pool together gap transitions and overlaps from all speakers, the distributions of Tg × To show no significant difference (Mann-Whitney U test: $p = 0.83$). However, if we divide the transitions between the speakers and assign each to the speaker that follows, we discover that the experimenter and one of the subjects are differentiated in their realization of Tg's and To's – the subject in Dialogue 3 had significantly longer overlaps than gap transitions (Mann-Whitney: $p < 0.05$), while the experimenter had longer Tg's than To's (Mann-Whitney: $p < 0.01$). More importantly, if we compare the duration of overlaps of the subjects with those of the experimenter, the latter are significantly shorter (Mann-Whitney: $p < 0.05$).

Turn transitions in close succession, i.e. those realized with a gap/overlap not exceeding 10 ms (investigated for instance by [6]), constitute 1.7 % of all transitions in our material. According to [22], this boundary can be raised to 120 ms, which is the detection threshold for a no-gap, no-overlap transition in a dialogue. In this respect, smooth or close transitions were realized in 18.5 % of all cases.

3.2. Back-channelling

As expected given the nature of the interview, the back channels (Tbc) were more frequent (1.8 times) with the experimenter than the subjects. Back-channelling can be used as a supportive affirmation and encouragement on part of the experimenter in order to induce subjects to continue speaking. However, there were intra-speaker differences in the experimenter's back-channelling behaviour (see below, section 3.4). The number of back-channels on the part of subjects was too low to permit quantitative analysis.

Out of all turn-taking phenomena, back-channels constitute 26 %, so they occur more often than overlaps (22 %). Interestingly enough, the durational properties of back-channelling are not significantly different from gap transitions or overlaps (Kruskal-Wallis ANOVA: $p > 0.05$). The median duration of back-channels was 384 ms.

3.3. Pauses

The overall duration of within-turn pauses constitutes 16 % of the total duration of the speech material. Since the number of experimenter's pauses was low (due to briefness of her turns), in the following text only pauses of the subjects are reported.

In individual dialogues, breath pauses (Pb) constitute 9-14 % of the duration, while silent pauses (Ps) up to 5 % and hesitation pauses (Ph) only up to 3 % of the subjects' speaking time. Since hesitation pauses were infrequent, differed greatly in manifestation and were problematic in terms of identifying and labelling, we decided to exclude them from further analyses.

Breath pauses were significantly longer than silent pauses (Mann-Whitney: $p < 0.001$), which was the case for all subjects. For three of the subjects the median duration of a breath pause ranged between 425 and 470 ms, while one (D2) was significantly different with 572 ms (Kruskal-Wallis ANOVA: $H(3, n = 894) = 23.2$; $p < 0.001$). Interestingly, the same speaker had the lowest median duration of silent pauses (274 ms), while the other three speakers clustered between 370 and 401 ms.

Working under the paradigm of [1], pauses can be considered *transition relevant places* (TRPs) where the change of speakers does not occur (similarly to [30]). It would therefore be interesting to compare the duration of breath and silent pauses to gap transitions, where the change of speakers takes place.

We discovered a significant difference between the duration of breath pauses and gap transitions – breath pauses were considerably longer (Mann-Whitney: $p < 0.001$). This holds true for all dialogues but one (D1). The case of silent pauses and gap transitions was less clear, as in two dialogues (D1, D2) Tg's were longer than Ps's, whereas in D3 and D4 it was the other way round.

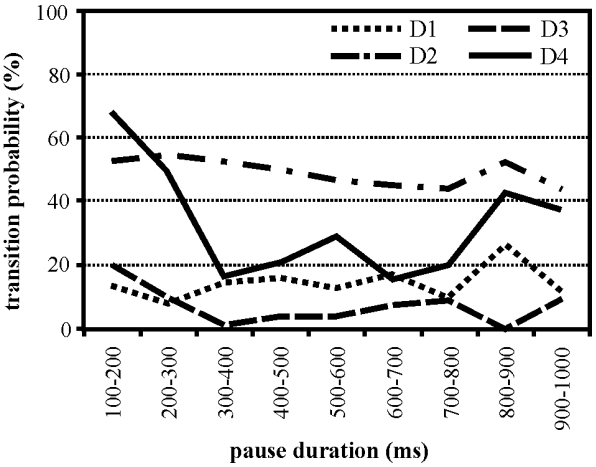


Figure 2: *The probability of turn transition as a function of pause duration. D1-D4 represent individual dialogues.*

We also computed transition probability by determining the percentage of gap transitions in all "silences" (Tg, Ps and Pb). Unlike [4], we did not find any drop in the percentage around 500 ms, nor an increase over 1500 ms. Instead, the percentage curve drops to about 20 % around 300 ms and stays at that level. Figure 2 shows the percentage in steps of 100 ms according to individual dialogues.

It is remarkable that between 300 and 1000 ms (over 1000 ms the number of cases starts to be too low) the percentage of gap transitions stays quite similar and highly dialogue-specific. For D1, gap transitions comprise around 15 % of all silent intervals, D2 has the highest percentage (48 % on average), D3 with only 5 % of Tg's represents the other extreme, and D4 fluctuates around 26 %.

3.4. The experimenter

Since the experimenter was the same in all dialogues, it presents an interesting opportunity to examine changes in her speech behaviour associated with change of dialogue partner.

The frequency of her turns paralleled the speaking behaviour of the subject, resulting in a substantial change in the average values from 1.7 turns per minute (D3) up to 10.5 turns per minute (D2). The duration of turns of the subject and the experimenter also seemed to be in a direct relationship, but the low number of dialogues prevented statistical verification.

Regardless of turn duration or frequency, it is remarkable to see that the experimenter uttered in all cases 58 % of the turns (the exact range was 58.1 to 58.8), leaving the remaining 42 % to the individual subjects. This could be the consequence of the task and of her role as interviewer.

Concerning the experimenter's transition behaviour, it can be seen from the data that her overlap strategy varied. Although in all four dialogues she produces approximately 1/3 of the overlaps, their duration differs significantly (Kruskal-Wallis ANOVA: $H(3, n = 57) = 8.5$; $p < 0.05$). Moreover, the duration and frequency of her back-channels differed as well. While in D3 the back-channels appear only 0.7 times per minute, in D2 it increases to more than three times per minute. This is in inverse relation to the length of the Tbc's – in D3 the back-channels were longest (median 593 ms), in D2 they were shortest (median 322 ms).

4. Discussion

Identical instructions to all participants resulted in a number of similar tendencies in the conversational structure of the dialogues, but several individual differences were also found.

Despite the fact that gap transitions, overlaps and back-channels are phenomena of a different nature and they manifest differently in the flow of speech, their temporal characteristics show remarkable similarities in our speech material.

The most frequent strategy for changing turns seems to be a transition with a silent gap between individual turns. Overlaps occured in our dialogues less often than reported in [6], where they constituted around 40 % of all transitions. We detected 30 % of overlaps; 22 % if back-channelling is also included as a transition phenomenon.

Concerning the findings of [5] and [6], who point out that the principle of minimizing gap and overlap postulated first by [1] may not be the main cause governing transition duration, our results show that although transitions without perceptible gap (that is ±120 ms) do not constitute the majority of transitions (only 18.5 %), the frequency of both gap transitions and overlaps generally does increase as their duration approaches zero.

If we compare the frequency of short transitions (without a perceptible gap), longer transitions and longer overlaps with the findings of [22: 511], our data are well in range of what the author found for 12 different languages. In our data, no-gap, no-overlap transitions constituted 18.5 % of all transitions, longer gap transitions 57.1 %, and longer overlaps 24.3 %. The author also noticed remarkable similarities in the ratio of these categories within language families – but in his material, Slavonic languages were not represented, so more research would be needed to see whether this hypothesis is valid also for them.

Durational characteristics of pauses and their relation to the turn-taking phenomena were also investigated. Gap transitions were significantly shorter than breath pauses. This may suggest that often the communication partner takes the turn before the other finishes breathing in. Despite breath pauses being physiologically conditioned, there also seem to be some idiosyncrasy in their duration. One of our speakers had a significantly longer duration of breath pauses and, interestingly enough, a significantly shorter duration of silent pauses at the same time. It is possible that this represents some kind of unconscious compensation on the part of the speaker. That the duration of pauses may be speaker-dependent and a useful tool for speaker identification has already been pointed out not only by forensic phoneticians (e.g. [11], [32]).

On the other hand, we could not replicate the findings of [4], who observed a marked drop in the probability of transition in pauses around 500 ms of length. We detected no such change in transition probability; however, this probability displayed remarkable and very stable inter-dialogue differences (shown in Figure 2).

In our experimental design, the experimenter received instructions concerning her conversational behaviour. The data suggest that she indeed performed her role as an interviewer – she uttered more turns and of shorter duration than the subjects (around 58 % in each dialogue). Furthermore, she took turns significantly less often by overlapping her communication partner; and, if she did, the overlaps were significantly shorter.

The conspicuous similarity of experimenter's turn percentage in all dialogues could be attributed to the experimental task. In future research we therefore plan to compare it with the controlled interview, which is the other conversational task from the VASST corpus, to see whether the percentage of turns changes. It is also possible to be a characteristics of the experimenter – again, we will examine the same dialogue task performed by other experimenters to verify this hypothesis.

Despite her consistencies mentioned above, the experimenter shows notable changes in her behaviour depending on the dialogue partner. This concerns mainly the duration of overlaps and the duration and frequency of back-channels. The notion of accommodation or entrainment may be evoked in this regard (see e.g. [7]), and these relationships should be investigated further.

The VASST corpus of contemporary Czech with its different speaking styles offers an excellent opportunity to research conversation behaviour in contrast to other speaking styles. In future studies the results will be verified on a larger amount of speech data, and the scope of the research should be enlarged to cover other prosodic phenomena in turn-taking, as well as to discover further individual patterns of speakers' behaviour.

5. Acknowledgements

The support of the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation is acknowledged. The first author was funded by a grant of the Czech Science Foundation (GACR 406/12/0298). The authors would like to thank all the students who participated in the recording and post-processing of the material. Special thanks also to Jan Volín for inspiring the paper and for his helpful insights and comments.

6. References

[1] Sacks, H., Schegloff, E. and Jefferson, G., "A simplest systematics for the organization of turn-taking for conversation", Language, 50(4): 696–735, 1974.

[2] Duncan, S., "Some signals and rules for taking speaking turns in conversations", Journal of Personality and Social Psychology, 23(1): 283–292, 1972.

[3] Ford, C. and Thompson, S., "Interactional units in conversation: Syntactic, intonational and pragmatic resources for the management of turns", in E. Ochs, E. Schegloff and S. Thompson [Eds.], Interaction and grammar, 134–184, Cambridge: Cambridge University Press: 1996.

[4] Wennerstrom, A. and Siegel, A. F., "Keeping the floor in multiparty conversations: Intonation, syntax and pause", Discourse Processes, 36: 77–107, 2003.

[5] Weilhammer, K. and Rabold, S., "Durational aspects in turn taking", Proceedings of the International Conference of Phonetic Sciences 2003, Barcelona, Spain, 2003.

[6] Heldner, M. and Edlund, J., "Pauses, gaps and overlaps in conversation", Journal of Phonetics 38: 555–568, 2010.

[7] ten Bosch, L., Oostdijk, N. and Boves, L., "On temporal aspects of turn taking in conversational dialogues", Speech Communication, 47: 80–86, 2005.

[8] Shriberg, E., Stolcke, A. and Baron, D., "Observations on overlap: Findings and implications for automatic processing of multi-party conversation", Proceedings of Eurospeech, vol. 2, Aalborg, Denmark: 1359–1362, 2001.

[9] Çetin, Ö. and Shriberg, E., "Analysis of overlaps in meetings by dialog factors, hot spots, speakers and collection site: Insights for automatic speech recognition", Proc. ICSLP, Pittsburgh: 293–296, 2006.

[10] Maclay, H. and Osgood, C. E., "Hesitation phenomena in spontaneous English speech", Word, 15: 1944. 1959.

[11] van Donzel, M. E. and Koopmans-van Beinum, F. J., "Pausing strategies in discourse in Dutch", Proceedings ICSLP '96, Philadelphia, USA, vol. 2: 1029–1032, 1996.

[12] Rose, R. L., "Crosslinguistic Corpus of Hesitation Phenomena: A corpus for investigating first and second language speech performance", Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), Lyon, France: 992–996, 2013.

[13] Boomer, D. S. and Dittmann, A. T., "Hesitation pauses and juncture pauses in speech. Language and Speech", 5: 215–220, 1962.

[14] Goldman–Eisler, F., "Pauses, clauses, sentences", Language and Speech, 15: 103–113, 1972.

[15] Campione, I. and Véronis, J., "A large-scale multilingual study of silent pause duration", Proceedings of Eurospeech 2002: 199–202, 2002.

[16] Local, J. and Kelly, J., "Projection and "silences": Notes on phonetic and conversational structure", Human Studies, 9: 185–204, 1986.

[17] Dankovičová, J., "The minimum pause duration in spontaneous speech", PROPH – Progress Reports from Oxford Phonetics 5: 17–24, 1992.

[18] Butcher, A., "Aspects of the speech pause: Phonetic correlates and communicative functions", Aipuk (Arbeitsberichte Institut für Phonetik Kiel): 15, 1981.

[19] Hieke, A., Kowal, S. and O'Connell, M., "The trouble with "articulatory" pauses", Language and Speech, 26(3): 203–214, 1983.

[20] Hansson, P., "Prosodic phrasing and articulation rate variation", Proceedings of Fonetik, TMH-QPSR, 44: 173–176, 2002.

[21] Goldman-Eisler, F., "Psycholinguistics: Experiments in spontaneous speech", New York: Academic Press, 1968.

[22] Heldner, M., "Detection thresholds for gaps, overlaps and no-gap-no-overlaps", Journal of the Acoustical Society of America 130(1): 508–513, 2011.

[23] Goldman-Eisler, F., "The distribution of pause durations in speech", Language and Speech, 4: 232–237, 1961.

[24] Ruder, K. F. and Jensen, P. J., "Fluent and hesitation pauses as a function of syntactic complexity", Journal of Speech and Hearing Research, 15: 49–58, 1972.

[25] Ferreira, F., "Effects of length and syntactic complexity on initiation times for prepared utterances", Journal of Memory and Language, 30: 210–233, 1991.

[26] Zellner, B., "Pauses and the temporal structure of speech", in E. Keller [Ed.], Fundamentals of speech synthesis and speech recognition, 41–62, Chichester: John Wiley, 1994.

[27] Zvonik, E. and Cummins, F., "The effect of surrounding phrase lengths on pause duration", Proceedings of Eurospeech 2003, Geneva, Switzerland: 777–780, 2003.

[28] Volín, J. and Weingartová, L., "Současný stav zkoumání zvukové stránky mluvních stylů", in preparation.

[29] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program], version 5.3.41, retrieved from http://www.praat.org/, 2013.

[30] Caspers J., "Local speech melody as a limitating factor in the turn-taking system in Dutch", Journal of Phonetics, 31: 251–276, 2003.

[31] Gravano A. and Hirschberg, J., "Turn-taking in task-oriented dialogues", Computer Speech and Language, 25: 601–634, 2011.

[32] Foulkes, P. and French, J. P., "Forensic speaker comparison: a linguistic-acoustic perspective", in P. Tiersma and L. Solan [Eds.], Oxford Handbook of Language and Law, 557–572, Oxford: Oxford University Press, 2012.

Segment Duration in Finnish as Imitated by Russians

Riikka Ullakonoja[1], *Mikko Kuronen*[2], *Pertti Hurme*[3], *Hannele Dufva*[2]

[1] Centre for Applied Language Studies, University of Jyväskylä, Finland
[2] Department of Languages, University of Jyväskylä, Finland
[3] Department of Communication, University of Jyväskylä, Finland

riikka.ullakonoja@jyu.fi, mikko.j.kuronen@jyu.fi, pertti.hurme@jyu.fi, hannele.dufva@jyu.fi

Abstract

The paper reports findings of a study in which Russian speakers without any prior knowledge of Finnish imitated utterances in that language, and, in particular, how they succeeded in imitating segmental duration. The data was analysed using acoustic measurements of segment duration as well as auditory analysis by four judges. The results show that Russian speakers faced difficulties in imitating some aspects of the complicated Finnish quantity system. On the other hand, many of the imitated words were judged as comprehensible.
Index Terms: duration, length, Finnish, imitation, language learning

1. Introduction

There has been a renewed interest in investigating the role of imitation in many fields that study social, cognitive and linguistic activity [1, 2]. To our knowledge, imitation has not been much used as an elicitation method in prosodic research. In this paper, imitation will be studied using a research design in which the participants were asked to imitate sentences of varying length in a language they do not know. We report the results of a pilot study of Russians imitating Finnish utterances, with the aim of focusing on how the participants succeeded in imitating the segment durations. The durational patterns in the imitations were investigated both by acoustic measurements and perceptual analysis. We also present preliminary reflections on the role of first language (L1) in imitating an unknown language and relate our findings to second/foreign language (L2) learning research.

Inspired by Hurme [3], who investigated how native speakers of Russian without any knowledge of Finnish succeeded in mimicking Finnish utterances, the present authors conducted a pilot study [4] where also the connection of the individual's working memory to the imitation ability was investigated. The results of both studies indicated that the success of imitation depended on the length of the utterance as measured in syllables and that the initial and final parts of the utterances were imitated best. In addition, the latter study found a positive correlation between the participants' score in the working memory test and the success of imitation. In this paper we analyse the prosody of the imitated utterances focussing on segmental durations. In contrast with Russian, in Finnish quantity is distinctive which also makes the results interesting from the point of view of second/foreign language learning.

1.1. Quantity in Finnish

Finnish has a complex quantity system. In two-syllable words alone there are eight possible combinations of /short/ and /long/ vowels and consonants [5]: from CVCV to CVVCCVV. Thus, *tule, tulle, tulee, tullee, tuule, tuulle, tuulee, tuullee* are all Finnish words, inflections of the verbs *tulla* (come) and *tuulla* (blow). Word stress is fixed on the first syllable of the word, and the main phonetic correlate of stress is a peak in F0 [6]. All vowels can occur phonemically long and short in all stressed and unstressed syllables [7], e.g. *sika* [sikɑ] (a pig), *siika* [si:kɑ] (a whitefish), *sikaa* [sikɑ:] (a pig, partitive) and *siikaa* [si:kɑ:] (a whitefish, partitive). Most consonants can occur as /long/ or /short/ between vowels too, making a single-double contrast [7] e.g. *muta* [mutɑ] (mud), *mutta* [mut:ɑ] (but). Even words like *muuttaa* [mu:t:ɑ:] (to change) are possible.

However, the phonetic reality of quantity in Finnish is even more complex [e.g. 8, 9, 10]. For instance, in CVVCV and CVCCV structures not only the long segments have longer duration but also the second-syllable vowel is shorter than in CVCV structures, where the second-syllable vowel can be phonetically characterized as long. Indeed, /short/ vowels can vary in duration from [very short] (CVVC*V* *kaato*), [short] (C*V*CV *kato*), [longish] (C*V*CCV *katto*) to [long] CVC*V* *kato*, depending on their position and the quantity pattern of the word [8]. Further, utterance length has an effect on segmental durations: segments tend to be shorter the more syllables the utterance consists of [11]. In recent studies also the tonal patterns in words with different quantity types and the possible role of F0 in maintaining the oppositions have been investigated [6, 12]. The possible role of F0 in perception of the quantity opposition has been shown by O'Dell [11].

1.2. Quantity in Finnish for learners

As quantity is a distinctive feature in Finnish, it is also often mentioned as one of the foremost difficulties of Finnish for L2 learners. However, the complexity and intricacy of the quantity system manifests also in native speakers' slower acquisition. It has been shown that while Finnish children learn to make the difference in the consonantal durations between CVCV/CVCCV patterns at an early age, they do not fully master those differences in the second-syllable vowels that are sub-phonemic but systematic even at the age of six [13].

As studies focussing on learners of Finnish as a foreign language have shown, quantity distinctions are notoriously difficult. Vihanta's [14] study of French learners of Finnish reading Finnish sentences containing minimal pairs showed that they often lengthened the final vowel (as in French) and failed to make the sub-phonemic durational difference in the second-syllable vowel in CVCV vs. CVVCV/CVCCV words. On the whole they tended to exaggerate the distinction between /short/ and /long/, leading to problems in understanding. Toivola [15] who studied the prosodic features and foreign accent in L2 Finnish spoken by Russian learners found out that they often produced either too short or too long segments in word-medial position, and segments that were too long in word-final position. The perception of Finnish quantity

by L2 speakers has been studied by Ylinen and colleagues [16, 17] who learned that the vowel quantity distinction was difficult to perceive even for Russians who had lived in Finland for several years.

In all, while the first extensive experimental acoustic analyses of Finnish quantity in native speakers' speech as well as tests of perception of the quantity distinction were conducted as early as in the 1970s [9], up till now there are few studies that have investigated the topic in L2 Finnish speech. Thus, the current study adds to the understanding of the Finnish quantity system as perceived and produced by non-speakers of Finnish.

1.3. Duration in Russian

In Russian, phoneme length is not distinctive (with the exception of some rare consonantal minimal pairs such as *strany – stranny)*. However, the vowel duration is used for another function, to signal the location of word stress, as stressed vowels are longer than unstressed ones vowels. Being distinctive, word stress can fall on any syllable and its position can differ in various forms of the same word. Proportionately, the durational difference between Finnish stressed long and short vowels is similar as between Russian stressed and unstressed vowels (about 2:1) [18]. However, in Russian the duration of the unstressed vowel depends also on its position relative to the stressed vowel, as there are two degrees of vowel reduction in unstressed vowels. The weak degree occurs in the syllable immediately preceding the stressed syllable, but can also occur in some other positions (e.g. word initial and phrase final positions) and has the duration of 1.25–1.4 times the duration of the other unstressed vowels (with the strong degree of reduction) [19]. The fact that in Russian the duration is one of three cues of word stress (the others being intensity and pitch) whereas in Finnish it mainly signals phonological length offers interesting possibilities for research.

2. Material and Methods

2.1. Material

Data was collected in an imitation task, following the design of [3]. Russian subjects were asked to orally imitate 30 auditory stimuli, each of which they heard only once, in a language they did not know. The stimuli were three-word Finnish utterances (with 4–11 syllables) previously recorded from a female speaker, a native Finnish speaker with a standard Finnish pronunciation. The stimuli aimed at echoing spontaneous utterances in spoken Finnish with a naturalistic prosody in mind. *Tili tuli tänään* (paycheck arrived today) and *yöllä saattaa tuulla* (at night it may rain) are examples. The stimuli were presented to the subjects in two different randomized orders in order to avoid fatigue and learning effects. The responses were recorded with an Edirol by Roland 4-bit Wave/MP3 R-09 digital recorder and a high quality Koss headset (sample rate 44.1 kHz, 16 bit resolution). A practice stimulus was presented before the experiment. Six native speakers of Russian (further R11, R12, R13, R14, R15, R16) were asked to imitate the stimuli. They were all female, aged 20–26, university students from St. Petersburg. Five of them were students of linguistics, and thus, had some phonetic training.

For the present analysis, we chose the stressed vs. non-stressed, short vs. long vowels (and long vs. short consonants) in the following disyllabic words: *päättää*[2] (to decide) – *päätä*[2] (decide, imperative form); *sata*[2] (hundred) – *saattaa*[2] (may); *tuulla*[3] (to blow) – *tulla*[2] (to come); *teetä*[3] (tea, partitive) – *teettää*[3] (to have made), *tili*[1] (pay check) – *tilli*[1] (dill); *tuli*[2] (fire) – *tulli*[2] (customs); *kisaa*[3] (competition, partitive) – *kissaa*[3] (cat, partitive). The numbers refer to the position of the word in the utterance (1=first word, 2=second word, i.e. middle, 3=last word). These words occurred in different utterance positions: the pairs listed above mostly (except for *tulla – tuulla*) occurred in the same sentential position, which makes their comparison possible.

2.2. Methods

First, the imitated utterances were submitted to auditory evaluation by the four authors, all native speakers of Finnish and experts in phonetics. Here, the judges listened to the model utterance and each imitation three times (each judge in a different randomized order) and rated the success of the imitation. The comprehensibility of both the whole utterance and each word were rated on a 1–5 Likert-type scale (0=missing, 1=completely against the model, 2=not completely comprehensible, 3=comprehensible, 4=rather good, 5=near-native). In addition, the judges were asked to submit verbal comments on the segmental and prosodic features that they found particularly disturbing or successful. The judges' numeric ratings ('c' in Figures 2–3) were highly consistent and reliable (Cronbach's alpha .91 at the utterance level, between .86–.92 at the word level).

Second, the acoustic analysis of segment durations were carried out in Praat [20]. First, the comprehensibly imitated words were annotated in the segmental level. The segment onsets and offsets were determined both visually in the spectrogram (as well as in the intensity and F0 curves) and auditorily, and marked manually. Words that were not comprehensibly imitated were discarded. In word-initial plosives the occlusion phase was excluded, but included in the position between the vowels. The durations and auditory ratings were analysed in different word types. The relative segment durations were calculated and expressed as the ratio between segment duration and the mean duration of the [very short vowel], second vowels in CVCCV and CVVCV structures, for each speaker. Due to the small sample size the data is not analysed using statistical tests at this point.

3. Results

3.1. Acoustic analysis of segment duration

First, we present the results of the acoustic analysis of segment duration in a) the first (stressed) syllable and b) the second (unstressed) syllable, further divided into three subgroups (Table 1). Four words were left out of the analysis either because they were not imitated by Russian speakers (*teetä*) or because they were the only occurrence of the particular word type *tuulla* (CVVCCV), *kisaa* (CVCVV), *kissaa* (CVCCVV).

Table 1. *Three categories of second syllable vowel duration following* [10]

Vowel duration	Words	Word type
very short	tulla, tilli, tulli päätä	CVCCV CVVCV
long	sata, tili, tuli	CVCV
very long	päättää, saattaa, teettää	CVVCCVV

Figure 1 shows the mean relative segment durations of the selected words as the proportion of the duration of the mean of the [very short] vowel duration produced by each speaker (see Table 1 for examples). Both Russian speakers and the Finnish native speaker make a distinction between short and long in the first-syllable vowel. However, there are differences in how the Russian subjects and the Finn produce duration in the second syllables, especially in case of the [very long] final vowel, imitated similarly as the [long] final vowel. As for consonants, the difference between /short/ and /long/ consonants is clearer in the speech of the Finnish speaker. The relative segment durations in the speech of the Finnish speaker are similar to those of other Finnish speakers as reported in [10].

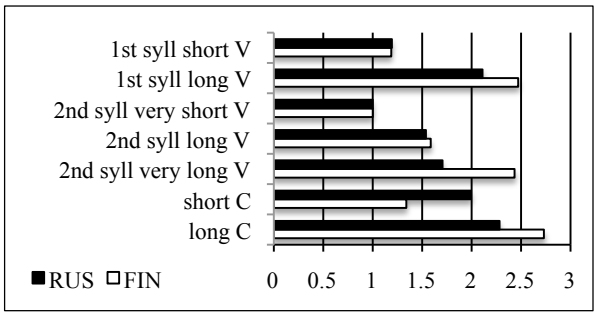


Figure 1: *Relative segment duration*

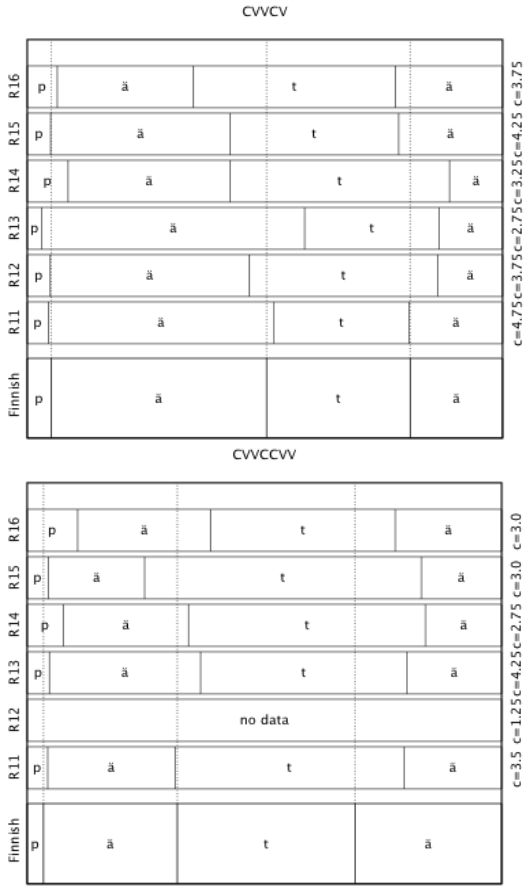


Figure 2: *Relative segment durations of the word päätä (above) and pääättää (below)*

Figure 2 shows the relative segment durations of the words *päätä* and *pääättää* in the Finnish model utterance and its Russian imitations, with time on the x-axis, speakers and comprehensibility ratings (c) on the y-axis. One of the Russian speakers (R11) matches the original durations very closely in her imitation of the word *päätä*. The other productions are less close to the target for all segments in both words, and the word *pääättää* is more difficult to imitate. As Figure 2 shows, the second syllable vowel in *pääättää* is too short in all imitations. However, often the preceding consonant is much longer than in the model. The word *pääättää*, consisting of three consecutive long segments, is structurally foreign to the Russian speakers, but interestingly, the subjects may aim at imitating the length they possibly hear by transferring the feature to the preceding consonant. Thus, Figure 2 confirms the finding illustrated in Figure 1 showing that a [very long] second syllable vowel is difficult to imitate.

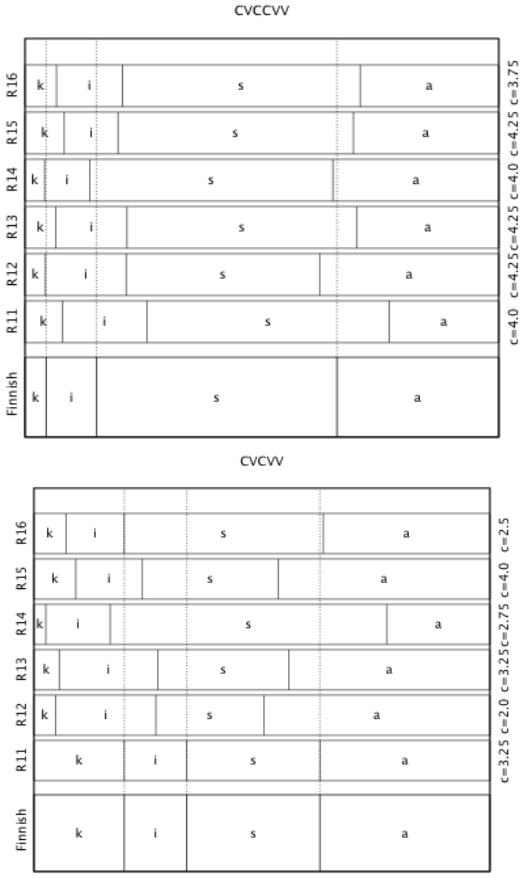


Figure 3: *Relative segment durations of the word kissaa (above) and kisaa (below)*

Next, we will investigate two words that were the only occurrences in their category, *kisaa* and *kissaa* (both in sentence accent position) and thus not included in the previous analysis. Figure 3 shows that some speakers clearly manage to imitate the durational patterns better than others. Speaker R14 matches all segment durations present in the Finnish model in the word *kissaa* and speaker R11 in the word *kisaa*. However, most Russian speakers fail to imitate the duration of the unstressed vowel in the second syllable for both words. This seems to interfere with the relative duration of the other segments in both words, similarly as in the word *pääättää* in

Figure 2. However, one possibility is that vowel duration is more difficult to imitate than that of consonants. In *kisaa* the speakers differ in their imitations: some (R14, R16) exaggerate the duration of the sibilant, others (R12, R13, R15) that of the word final long vowel. Also the duration of the first vowel can be exaggerated (R12, R13). Overall the durational pattern of the word *kissaa* was easier to imitate than *kisaa*, *päätä* or *päättää*, probably because it was pronounced slowly, with a strong sentence accent and some emotionality in the Finnish model. Furthermore, both *kisaa* and *kissaa* were in utterance final position, and thus easier to imitate than *päätä* and *päättää,* which were in utterance medial position, see [4]. Also, it can be speculated that the long fricative in the word *kissaa* may be easier for Russians to perceive and then imitate than the short plosive in *päätä*.

3.2. Perceptual evaluation of the segment duration

Next, we will discuss the perceptual evaluations of the imitations by the four judges. The verbal comments in the perceptual analysis suggest that whenever the judges perceived the participants' productions as recognizable on the whole, also the durational patterns of the model had been closely followed. Although there was some variation – and some poor imitations - the durational patterns in the subjects' productions quite consistently matched the native speakers' intuition of the boundary between /short/ or /long/ quantity. In line with the acoustic analysis, the word-final /long/ vowels were frequently, particularly after a long vowel in the first syllable, perceived as "too short".

Figure 4 shows the comprehensibility ratings (1–5) of the judges in different word types. Overall, most word types got a mean comprehensibility rating of over or near three (3=comprehensible). This means that despite the potential difficulties in the durational patterns discussed above, the speakers' imitations were generally comprehensible. Figure 4 includes all seven word types (see 2.1), although in some of them there is only one word as an example. In line with what was said above concerning Figure 3, the word *kissaa* (CVCCVV) was the easiest to imitate. Other successfully imitated word type was CVVCCVV (e.g. *päättää*). The least successful type CVVCCV was the word *tuulla*, the segmental content of which might be difficult for perception.

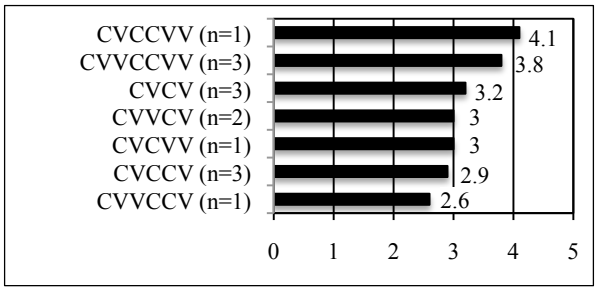


Figure 4: *Mean comprehensibility ratings (1–5) of the imitations according to the word type*

Obviously, the comprehensibility ratings are not based exclusively on successful imitations of durational patterns. Other factors, such as the position of the word within the utterance (see 2.1), the nature of the segments, and sentence stress are likely to influence as well. Nevertheless, it is obvious that a clear violation of the durational pattern is likely to result in low comprehensibility ratings especially when it concerns the first, stressed syllable of the imitated word. This can be seen in Figure 3, in the word *kisaa* where the violation in duration of the first vowel or sibilant appears to be associated with low comprehensibility ratings.

4. Discussion

The paper examined segment duration in imitated Finnish speech by Russians. The results seem to indicate that some features of the Finnish quantity system are indeed difficult to imitate. First, there are large individual differences. Second, there are differences caused by the word type and the position of the word in the utterance. In more particular, our results show that speakers of Russian were not successful in imitating the [very long] vowel in the (unstressed) second syllable, but produced a shorter vowel while prolonging the duration of the preceding consonant (Figures 2 and 3). As such a long vowel does not exist in Russian in similar positions, this is also an example of possible L1 influence. Length patterns in L1 have been shown to have influence on conveying quantity in L2 [21, 22]. Also, previous studies [14, 15] have shown that L2 learners of Finnish tend to produce word-final vowel durations as too long, suggesting that learning can lead to exaggerating the learned category, a phenomenon which is well-known from both L1 and L2 learning of different linguistic skills. Exaggeration of quantity categories in L2 Swedish is reported in [21]. Indeed, the word-final position may be particularly vulnerable for non-nativelike durations as even native speakers do not fully master the durational system by the age of six [13].

The findings indicate that closer examination is needed to study the role of such factors as, e.g., sentence stress, segmental content, tonal patterns [6, 12] and utterance type in the imitation of segment durations. Further, the imitation experiment with its interplay between articulation and perception – and between two languages, Finnish and Russian – offers new insights into cross-language phonetics and learning of L2 pronunciation, e.g. the effect of segments and prosodic features of L1 on L2 and possible differences in the acquisition of segmental and prosodic features.

5. Conclusions

The preliminary results suggest that more research is needed on the interplay between general and language-specific aspects in the ways prosodic features are perceived and articulated. In this experiment, data was elicited that aimed at mapping how subjects cope with the articulation of utterances in a language that is unfamiliar to them, using their auditory perception only. To see how prior knowledge about the language or its writing system exerts influence, new studies will be designed. Further studies will be conducted to examine the presence of L1 transfer in imitated utterances vs. second/foreign language learners' speech. In conclusion, we believe imitation experiments of an unknown language offer an interesting new tool for the study of the role of imitation in language learning and development.

6. Acknowledgements

We thank the students and staff at St. Petersburg State University for participating in and organizing the experiment, especially Dr. Nina Volskaya. We also wish to express our gratitude to the Academy of Finland for a travel grant (n:o 260539) to the first author, which enabled data collection.

7. References

[1] Yoshida, H., “Imitating and Repeating Foreign Language: lmplications for Language Teaching”, University of Kyoiko, 2009. Online: ir.lib.osaka-kyoiku.ac.jp/dspace/bitstream/.../3/ok_eibun_54_083.pdf, accessed on 21 Feb 2014.

[2] Reiterer S.M., Xiaochen, H., Sumathi T.A., Singh, N. C. “Are you a good mimic? Neuro-acoustic signatures for speech imitation ability”. Frontiers in Psychology, 4:782, 1–13, 2013.

[3] Hurme, P. ”Oudon kielen matkimisesta: Ihmisen kyvyistä ja rajoituksista imitoida oudon kielen lauseita”, in Fonetiikan paperit - Helsinki 1975, 19–35, 1975.

[4] Ullakonoja, R., Dufva, H., Hurme, P. and Kuronen M., (submitted) ”How to imitate an unknown language? Russians imitating Finnish”, Proceedings of the Finnish Phonetics symposium, Turku, 25-26th Oct 2013.

[5] Lehiste, I., “The function of quantity in Finnish and Estonian”, Language 41:3, 447–456, 1965.

[6] Suomi K., Toivanen J. and Ylitalo, R., “Durational and tonal correlates of accent in Finnish”, Journal of Phonetics 31, 113–138, 2003.

[7] Iivonen, A., ”Major features of standard Finnish phonetics”, in V. de Silva and R. Ullakonoja [Eds] Phonetics of Russian and Finnish, general description of phonetic systems, experimental studies on spontaneous and read-aloud speech, 47–65, Peter Lang, 2009.

[8] Wiik, K. and Lehiste, I., “Vowel quantity in Finnish disyllabic words”, Second International Congress of Fenno-Ugrists, Helsinki 1965, 569–574, 1968.

[9] Lehtonen, J., “Aspects of Quantity in Standard Finnish”, Studia Philologica Jyväskyläensia 6, Jyväskylä, 1970.

[10] Suomi, K., Toivanen, J. and Ylitalo, R., “Finnish sound structure. Phonetics, phonology, phonotactics and prosody”, Studia Humaniora Ouluensia 9, Oulu, 2008. Online: http://herkules.oulu.fi/isbn9789514289842/isbn9789514289842.pdf, accessed on 14 Dec 2013.

[11] O’Dell, M., “Intrinsic Timing and Quantity in Finnish”, Acta Universitatis Tamperensis 979, University of Tampere, Tampere, 2003. Online: http://tampub.uta.fi/bitstream/handle/10024/67344/951-44-5838-9.pdf?sequence=1, accessed on 14 Dec 2013.

[12] Vainio, M., Aalto, D., Järvikivi, J. and Suni, A., “Quantity and tone in Finnish lexically stressed syllables”, Proceedings of the Second International Symposium on Tonal Aspects of Languages, 121-124, 2006. Online: https://helda.helsinki.fi/handle/10138/24708, accessed on 15 Dec 2013.

[13] Hurme, P. and Sonninen, A., “Development of durational patterns in Finnish CVCV and CVCCV words”, in Hurme, P. [Ed] Papers in Speech Research 6, 1–14, Jyväskylä, 1985.

[14] Vihanta, V., ”Suomen äännekestot ranskalaisen suomenoppijan kannalta”, in Hurme, P. and Dufva, H. [Eds] Papers from the 14th Meeting of Finnish Phoneticians, Papers in Speech Research 7, Jyväskylä, 101–122, 1987.

[15] Toivola, M., ”Vieraan aksentin arvioiminen ja mittaaminen Suomessa”, Unigrafia, Helsinki, 2011. Online: https://helda.helsinki.fi/bitstream/handle/10138/27888/vieraana.pdf?sequence=1, accessed on 14 Dec 2013.

[16] Ylinen, S., Shestakova, A., Alku, P. and Huotilainen, M. “The perception of phonological quantity based on durational cues by native speakers, second-language users and nonspeakers of Finnish”, Language and Speech 48:3, 313–338, 2005.

[17] Ylinen, S., Shestakova, A., Huotilainen, M., Alku, P. and Näätänen, R. “Mismatch negativity (MMN) elicited by changes in phoneme length: A cross-linguistic study”, Brain Research 1072:1, 175–185, 2006.

[18] de Silva, V. “Quantity and quality as universal and specific features of sound systems: Experimental phonetic research on interaction of Russian and Finnish sound systems”, Studia Philologica Jyväskyläensia, Jyväskylä, 1999.

[19] Jones, D. and Ward D. “The Phonetics of Russian”, University Press, Cambridge, 1969.

[20] Boersma, P. and Weenink, D., “PRAAT: Doing phonetics by computer”, Version 5.3, [Computer program], 2013. Online: http://www.praat.org/ (University of Amsterdam, Amsterdam), accessed on 14 Nov 2013.

[21] Thorén, B. “The priority of temporal aspects in L2-Swedish prosody: Studies in perception and production.” Stockholms universitet, 2008.

[22] Thorén, B. “Durations of phonologically long segments in native and foreign accented Swedish.” Fonetik 2010, Lunds universitet, 2-4 juni 2010, 2010. Online: http://www.bossethoren.se/bosse_b_thoren_fon2010.pdf, accessed on 21 Feb 2013.

Savosavo word stress: a quantitative analysis

Candide Simard[1], *Claudia Wegener*[2], *Albert Lee*[3], *Faith Chiu*[3], *Connor Youngberg*[1]

[1]Department of Linguistics, SOAS, University of London, United Kingdom.
[2] Faculty of Linguistics and Literary Studies, University of Bielefeld, Germany.
[3]Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom.

cs75@soas.ac.uk, claudia.wegener@uni-bielefeld.de, kwing.lee.10@ucl.ac.uk, faith.chiu.11@ucl.ac.uk, cy4@soas.ac.uk

Abstract

This paper presents a quantitative analysis of stress in Savosavo (unclassified), an endangered language spoken on Savo Island, (Solomon Islands). Acoustic analyses comprise the measurements of F0, duration, and intensity for each syllable in a dataset carefully selected from elicited speech from one speaker only, aiming to test the effect of increasing morphological complexity on stress realization in a system that displays some variation. Statistically significant variation is found in all correlates between stressed and unstressed syllables, thus fitting with widely attested manifestations of stress cross-linguistically. Findings were further tested with a re-synthesis tool, to confirm our initial hypotheses. Our results demonstrate that the current annotation scheme is a reliable representation of the data, and that the qTA component embedded in PENTAtrainer is effective in modelling F0 contours, even with less controlled data as input. We will argue for the usefulness of instrumental phonetic investigations in describing lesser-known languages, to enhance our understanding of the characterization of the prosodic systems of the world's languages.

Index Terms: prosody, word stress, pitch, duration, Austronesian language, Oceanic language, prosodic typology

1. Introduction

Savosavo is an unclassified, Papuan language spoken on Savo Island (one of the Solomon Islands) by about 2500 people. It does not have any (close) relatives, and has been surrounded by Oceanic languages for at least several hundred years, hence there has been long-standing contact between Savosavo and many of its neighbours. Because of a shift in the younger generations to Pijin and English, Savosavo is an endangered language currently being documented by Wegener [1]. The speech data on which this study is based is part of the documentation of Savosavo, which is thus precious as the sole possible corpus of analysis.

Savosavo is a verb-final language, with postpositions, and adnominal modifiers preceding the head of an NP; it also has a marked-nominative system with case-marking enclitics on syntactic subject noun phrases, but no case-marking on object noun phrases, whereas on verbs only syntactic objects are cross-referenced by means of affixes or stem modification, while syntactic subjects remain unmarked.

In Savosavo, a syllable can either consist of only a vowel nucleus or a vowel nucleus and a consonant onset, i.e. the basic syllable structure is (C)V. Most roots consist of two or three syllables, but roots of four and more syllables also occur. Savosavo has been analysed as a stress language based primarily on auditory impression, with additional qualitative analyses in [1]. Generally, primary stress falls on the penultimate syllable of a root; initial syllables receive a secondary stress if primary stress is on a non-initial syllable. Impressionistically, stressed syllables have been associated with longer duration, clearer pronunciation, higher intensity and sometimes higher pitch. This paper aims to quantify such claims using data extracted from field recordings, however constrained the recording conditions. While lab speech may be desirable and advantageous for such analyses [2], we contend that field-based descriptions are still feasible. The Savosavo data demonstrates a consistency in acoustic correlates sufficient to corroborate its status as a lexical stress language. Its lexical prosody mirrors that of English, a well-studied stress language.

2. Methodology

2.1. Recording

The recordings of Savosavo being analysed are from Wegener (2007-2010), with one male native speaker serving as consultant. Recording took place in a secluded area of the village, though some background noise was unavoidable. The data is extracted from two elicitation sessions lasting about 35 minutes each, in which the subject was prompted by the linguist (using English) to provide a translation for various items, starting from a single word and increasing in morphological complexity e.g. a citation form for a verb, and then various inflections and some negative forms. 207 tokens were analysed in total.

2.2. Annotation and analysis

The initial transcription and annotation was done with ELAN [3] (Sloetjes & Wittenburg, 2008); the 207 selected tokens were converted into Praat [4] textgrid format. The assignment of the stressed syllable was based on the phonological stress assignment rules described above. Syllable boundaries were hand-labeled and glottal pulse data was generated and manually verified for missed or double marked vocal cycles in the wave form. The ProsodyPro script [5] was used to extract all the measurements including mean F0 (based on 10 evenly spaced F0 points from each labelled interval), max F0, duration, and mean intensity, and computed time-normalised F0 contours for each token. The resulting files from ProsodyPro were analysed to investigate the acoustic correlates of lexical prominence.

2.3. Analysis by synthesis

In order to verify the reliability of the stress assignment annotation scheme on which the above analysis is based, the

measurements were re-analysed with PENTAtrainer2 [6], a semiautomatic software package for the analysis and synthesis of speech melody, built upon the Parallel Encoding and Target Approximation model [7] also running on PRAAT [8]. In a preliminary phase, intervals were labeled according to the communicative function tested. The program then extracted the optimal parametric values for the tested communicative function through analysis by synthesis controlled by simulated annealing [6]. The tool assumes three model parameters controlling the F0 movement of each interval (here a syllable), including target slope (*m*), target height (*b*), and strength of target approximation (λ), where *m* and *b* specify the form of the pitch target and λ indicates how rapidly a pitch target is approached.

These results were used to generate F0 contours and compare them with the individual real utterances in the corpus.

In the current annotation scheme, two communicative functions were included, namely Stress condition (stressed (S) vs. unstressed syllable (U)) and Syllable location (Left, Medial, and Right edge of the utterance) as illustrated in Figure 1.

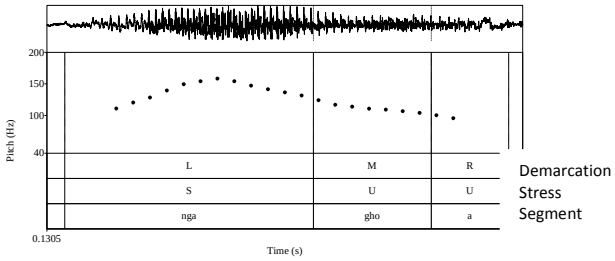


Figure 1. Example of functional annotation of recording for PENTAtrainer2, the top tier indicates the Syllable Location, the 2nd tier its stress condition and the 3rd tier the segmental content.

3. Results

3.1. Acoustic correlates of stress

The acoustic correlates of stress considered in this study are durational (ms.), intensity (dB) and mean F0 which is taken, for our purpose, as the direct correlate of pitch. The results shown in Table 1 indicate that stressed syllables have higher pitch with mean values of 112.06Hz to 96.36 Hz respectively; longer duration, with 212.91 ms to 176.09ms; and greater intensity, with 68.60dB to 64.29dB, than their counterparts. These results are statistically significant for all 3 correlates: F0 ($F=162.299$ (1, 1674) $p<0.001$; intensity ($F=48.933$ (1, 1696) $p<0.001$; duration ($F=146.349$ (1,1696) $p<0.001$.

Table 1. *Acoustic correlates of stress in Savosavo.*

Stress		Mean F0 (Hz)	Duration (ms)	Mean Intensity (dB)
S	Mean	112.06	212.91	68.60
	N	484	484	484
	Std. Deviation	23.50	56.39	9.91
U	Mean	96.36	176.09	64.29
	N	1192	1214	1214
	Std. Deviation	22.61	56.71	12.05

3.2. Analysis by synthesis

Functionally annotated data were fed into PENTAtrainer2, which computed the average parametric value of each interaction of the two functions (i.e. stress condition and syllable location). The results are shown in table 2. The learned values are compatible with our expectations that stressed syllables have a falling underlying target (shown by the negative *m* values) wherever they are located (left, medial or right periphery); these targets are higher than their unstressed counterparts (evident from the greater *b* values) when they are not word-initial. The anomalous Strength value for unstressed syllables at the right edge (100) should be disregarded, as it can be interpreted as a result of missing glottal pulses in the final syllable of some utterances, an artefact of field data commonly observed.

Table 2 *Mean target parameters of the Savosavo data learned through qTA modeling.*

Stress condition	Syllable Location (demarcation)	*Slope m*	*Height b*	*Strength λ*
S	Left	-42.32	1.74	20.15
	Medial	-52.09	-1.86	29.55
	Right	-29.56	4.89	5.34
U	Left	-0.5	2.39	36.86
	Medial	-5.8	-1.66	8.79
	Right	-26.84	-7.27	100

Subsequently, these averaged values were used to resynthesize the F0 contour of each utterance in the corpus. Mean RMSE (Root Mean Square Error is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled, since the RMSE is a good measure of accuracy, it is ideal if it is small) and Pearson's *r* (Correlation – often measured as a correlation coefficient – indicates the strength and direction of a linear relationship between two variables) of synthesis accuracy were respectively 2.48 and 0.87, these results are comparable to a study of English stress also using PENTAtainer2 [9] who report values of 3.97 ± 0.29 0.478 ± 0.028 for RMSE and R, respectively (for the encoding of the stress function only). The satisfactory accuracy of synthesis, even with only two functions annotated, demonstrates that the current annotation scheme is a reliable representation of the data, and that the qTA component embedded in PENTAtrainer is effective in modelling F0 contours.

4. Discussion

Some limitations due to the field-based nature of our data need be mentioned. First, the data is from only one speaker, and we are aware that in the context of a lab-based study, data from a lone speaker would not be deemed sufficient for a quantitative analysis; in our context it may indeed be a source of confounding factors such as speaker specific speaking style. Second, the male consultant in question uses a low mean F0 (100.7 Hz average across all utterances analysed), often close to his own pitch floor. As a result, some syllables were produced in non-modal phonation, hence did not contain glottal pulses. Third, where glottal pulses are not present, the duration of the syllable in question is determined by the annotator based on auditory

impression. The aforementioned limitations thus need to be taken into account when interpreting our results; nonetheless, we maintain that the acoustical correlates reported in the present paper serve as strong evidence to establish that Savosavo has stress, comparable to English, and unlike other languages, like Urdu, where lexical stress is marked by a lower F0 instead [10]. Further research will also investigate the acoustic difference between primary and secondary stress in Savosavo. Existing literature [1] has postulated the existence of secondary stress within the language, despite the claim of 'no or little [auditory] difference in realisation' between syllables carrying primary versus those with secondary stress. A possible avenue for future research would be to annotate for perceive primary and secondary stress and conduct descriptive statistics between these subgroups. An obvious limitation to this suggestion might be that the learning accuracy of the re-synthesis tool would be biased towards a system that distinguishes between three variables (Stressed, Unstressed and Secondarily Stressed) rather than one which uses two (as in this paper), rendering it incomparable with our above presented annotation scheme.

Finally, expanding from the word domain, continuing analyses will test how the encoding of lexical stress may interplay with sentential prosodic functions, such as modality and focus.

5. Conclusion

In this paper, we have provided quantitative evidence for the marking of lexical prominence in Savosavo through stress to complement previous qualitative analyses. We demonstrate that, even under adverse recording conditions, it is possible to carry out analyses using tools usually reserved for lab speech.

It is important to restate that this study forms the basis of further descriptive work on the prosodic system of Savosavo, so that after establishing the foundation for stress, we can define with more accuracy the nature of its syllable by analysing phonological processes such as syllable fusion. This description, in turn, will contribute to the ongoing areal research aiming to establish whether a historical change took place from a moraic to a syllabic system [11]. These findings will also be the basis for further investigations that will examine, for example, the interplay of word stress and sentence stress.

This paper argues for the usefulness of instrumental phonetic investigations in describing lesser-known languages. Our results demonstrate that the current annotation scheme is a reliable representation of the data, and that the qTA component embedded in PENTAtrainer2 is effective in modelling F0 contours. Finally, this paper also illustrates the usefulness of tools such as ProsodyPro for acoustic measurements, and PentaTrainer2 for hypothesis testing, tools that will eventually make a contribution towards our greater understanding of the characterization of the prosodic systems of the world's languages.

6. Acknowledgements

The authors would like to thank the Savosavo speakers for their continuing efforts to document their language and to make the recordings available for linguistic research. The present study is funded by DoBeS (Volkswagen Foundation) as part of the project 'Discourse and prosody across language family boundaries: two corpus based case studies on contact induced syntactic and prosodic convergence in the encoding of information structure'.

7. References

[1] Wegener, Claudia. A Grammar of Savosavo. Mouton Grammar Library, volume 61. 2012.

[2] Xu, Yi. In defense of lab speech. *Journal of Phonetics* 38: 329-336. 2000.

[3] Sloetjes, H., & Wittenburg, P. Annotation by category ELAN and ISO DCR. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (pp. 816–820). Marrakech, Morocco. 2008.

[4] Boersma, P. P. G., & Weenink, D. J. M. Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/ on 2013/11/30. 2013

[5] Xu, Y. ProsodyPro: A tool for largescale systematic prosody analysis. In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013) (pp. O1–1). Aixen Provence, France. 2013.

[6] Xu, Y. and Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57, 181-208. 2014.

[7] Xu, Y. Speech melody as articulatorily implemented communicative functions. Speech Communication, 46(34), 220–251. 2005.

[8] Boersma, Paul & Weenink, David (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.66, retrieved 9 March 2014 from http://www.praat.org/.

[9] Liu, F., Xu, Y., Promon, S., & Yu, A. C. L. Morpheme like prosodic functions: Evidence from acoustic analysis and computational modeling. Journal of Speech Sciences, 3(1), 85–140. 2013.

[10] Hussain, S. Phonetic correlates of lexical stress in Urdu. PhD Thesis. Northwestern University, Evanston, IL. 1997.

[11] Palmer, Bill. Shifting stress. Shifting stress: synchronic variation as a manifestation of diachronic change in Kokota. (Oceanic) prosody LAGB spring 2003.

The Perception of Prosodic Focus in Persian

Mortaza Taheri-Ardali[1], *Hamed Rahmani*[2], *Yi Xu*[3]

[1]Department of Linguistics, Institute for Humanities and Cultural Studies, Tehran, Iran
[2]Department of Linguistics, Radboud University, Nijmegen, Netherlands
[3]Department of Speech, Hearing and Phonetic Sciences, University College London, UK
m.taheri@ihcs.ac.ir, h.rahmani@let.ru.nl, yi.xu@ucl.ac.uk

Abstract

In a previous production experiment, post-focus compression (PFC) of F_0 and intensity were found to be present in Persian. It was also shown that F_0 and duration were the main correlates of prosodic focus in Persian. However, the perceptual relevance of PFC in Persian was not yet clear. The present paper reports the findings of an experiment on focus perception in Persian. Native speakers of Persian listened to sentences produced with focus in different positions as well as the neutral-focus sentence, and judged the presence and location of focus. Results show that final focus is identified much less well than other types of focus, and most of its confusion is with neutral focus. This shows that the presence of PFC is a main factor in recognizing prosodic focus in Persian.

Index Terms: prosodic focus, PFC, perception, post-focus, pre-focus, on-focus, F_0, intensity, Persian.

1. Introduction

There are different strategies to catch the attention of a listener to a particular portion of an utterance, i.e., to mark focus. This can be done both by syntactic and morphological means and by prosodic devices. Prosodically, many languages use phonetic variation in F_0, duration and intensity to mark focus. In particular, prosodic focus is realized not only by increasing F_0, duration and intensity of the on-focus component itself, but also by compressing the pitch range and intensity of the post-focus elements [5, 7, 16, 27, 28, 31]. There is also increasing evidence that post-focus compression as a perceptual cue plays a pivotal role in focus perception [11, 29]. It has been reported that if there are no F_0 peaks after an earlier peak in a sentence, it would be easier for listeners to perceive a non-final focus; otherwise listeners are prone to hear an additional late focus or no focus [15, 18]. It is also shown that when focus is not utterance-final, i.e., when PFC is applicable, its perceptual recognition is much easier, whereas final focus is often confused with neutral focus [3, 5, 13, 18]. Hence, findings from focus perception seem to lend further support to the importance of PFC [29]. However, the perceptual importance of PFC is not widely accepted, and much of the research on focus perception is still mainly concentrating on the focused words only [e.g., 20, 25]. There is therefore a need to explore further evidence for the importance of PFC in focus perception. In this paper, we present data of a perception experiment on Persian, for which PFC has been found in a recent study [22]. Before probing the perceptual effectiveness of PFC, a brief background of Persian prosody is provided in the next section.

2. Persian prosody

Persian, an SOV language with fairly free word order, is an Iranian language within the Indo-Iranian branch of the Indo-European family. Regionally, Persian has three major varieties: (1) the Persian of Iran (2) the Persian of Afghanistan, now called Dari and (3) the Persian spoken in Tajikistan in Central Asia. The variant used in this study is Iranian Persian, the official language of Iran and the mother tongue of about 60% (42 million) of Iran's population. It is worth noting that bilingualism and multilingualism are widely found in Iran [6].

There have been three central issues about Persian prosody. The first concerns word prosody. Persian has traditionally been described as a stress language. Abolhasanizade et al [1] recently found no marked phonetic difference between stressed and unstressed syllables independently of the presence of intonational pitch accents. The authors conclude that Persian word prosody involves a lexically-sensitive pitch accent assignment system that is more like Tokyo Japanese, which has no stress in the phonetic sense, than West Germanic, where stressed and unstressed syllables differ in durational and spectral properties.

The second issue relates to sentence prosody. Initial works, mostly based on the British tradition, have carefully documented the intonational patterns for various sentence types [24, 23]. In more recent literature, there has been a tendency toward the framework of autosegmental-metrical and intonational phonology [8, 14, 19]. A remarkable development to emerge from these studies concerns the issue of main stress in the sentence, generally referred to as the nuclear stress. Eslami [8] and Kahnemuyipour [12] have proposed a number of syntax-based rules to identify the location of the main stress within sentence.

The notion of syntax-based sentence stress, however, is contrasted by the findings related to the third issue, which concerns prosodic focus [1, 10, 19, 21]. According to these findings, any constituent in Persian can be contrastively focused, the only constraint being that focused elements cannot appear post-verbally. The phonetic properties of focus has been addressed in a number of experimental studies. The evidence provided by Sadat-Tehrani [19] points to greater pitch excursion and longer duration of focused elements. In a more detailed experiment, Abolhasanizade et al [1] show that focus has no significant durational and spectral effect and is expressed only by the differences in F_0. That the effect of duration and intensity is negligible is also maintained by Hosseini [10], who claims that F_0 is the only robust acoustic correlate of focus. As for the post-focal region, Abolhasanizade et al [1] provide evidence for PFC of F_0 and intensity. According to these authors, while the pitch range of the post-focal elements is phonetically reduced, the pitch accents are not deleted after the focus. This contradicts earlier report of complete post-focal de-accentuation [19, 21, 10]. The

existence of PFC in Persian is also found by Taheri-Ardali and Xu [22], who showed that focus not only increased the F_0 and duration of words under focus, but also decreased F_0 and intensity of post-focus words. In contrast, they did not find significant changes in on-focus intensity or post-focus duration.

Previous research has therefore established the existence of PFC in Persian. But there have been no empirical studies on the importance of PFC in the perception of focus in Persian, a gap the present study aims to fill.

3. Method

3.1. Stimuli

The sentence used in the perception experiment, as shown in Table 1, was taken from the previous production experiment [22]. The key words in the sentences consisted of mostly sonorant sounds to make sure that the F_0 contours were as smooth and connected as possible. The sentences were produced by five male speakers, who repeated each of the sentences five times in blocked random order. There were a total of 30 utterances from each speaker.

Table 1. *The target sentence of the experiment*

W1	W2	W3	W4	W5
Maha	baba-ye	nili-ro	lændæn	didim
we-PL	father-EZ	Nili-DO	London	see.PST-1PL

To elicit focus on a specific word in the production experiment, the target sentence was preceded by a sentence in parentheses. This focus cueing sentence was the same as the target sentence except the focused word, the verb and the ending word 'but' (Table 2).

Table 2. *The focus cueing sentences*

Focus	Focus cueing sentence (plus 'but')
W1	**They** didn't see Nili's father in London
W2	We didn't see Nili's **uncle** in London
W3	We didn't see **Amini's** father in London
W4	We didn't see Nili's father in **Tehran**
W5	We didn't **take** Nili's father to London

The stimuli for the current perceptual study were selected using mean standard deviation of F_0 across all repetitions of each speaker as arbitrary criteria, a method previously used in [5]. Speakers with minimum, maximum and median standard deviations were chosen. Then all tokens from these three speakers were used stimuli. In total, there were 6 foci x 5 repetitions x 3 speakers = 90 tokens.

3.2. Subjects

Five males and five females participated as subjects. All were native speakers of Persian with average age of 26.1 which is comparable to the age range of those who took part in production experiment. Each subject was paid in exchange for his/her participation in the test. They have also reported no hearing or speech disorders.

3.3. Listening Procedure

The experiment was conducted using ExperimentMFC in Praat software [2]. The listeners were instructed on how to choose the emphasized word. They listened to the stimuli once and then judged which word was focused. They were also told if none of the words was focused, the neutral focus choice must be selected. Before the start of the experimental trials, listeners had five practice trials without any feedback on the correctness of the answers.

3.4. Results

Figure 1 shows focus recognition rates of all six focus conditions, in percentage of correct identification. The overall rate of focus identification is fairly high. Compared to each other, the identification rate for the final focus (last word) is much lower than other focus conditions.

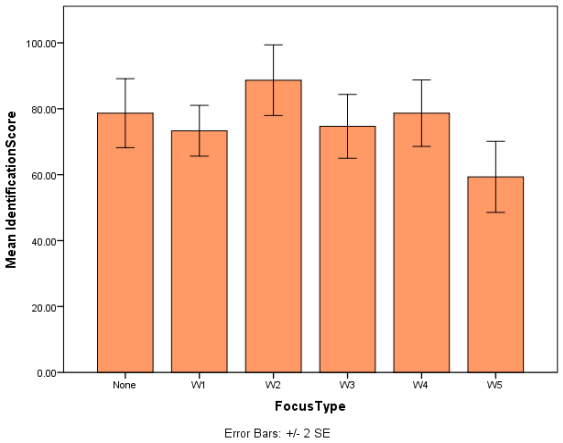


Figure 1. *Percentage of correct identification of neutral focus and focus on word 1-5. The error bars represent standard errors.*

Table 3 shows a confusion matrix of the focus perception experiment. It can be seen that when listeners identified focus location wrongly, they were prone to hearing no focus in the utterance. For example, 29.3% of the final foci that the listeners heard wrongly were heard as neutral focus. In contrast, there were almost no wrongly identified cases that were recognized as final focus.

Table 3. *Confusion matrix of focus perception (percent). Bold face indicates correct focus identification.*

heard as / original	none	W1	W2	W3	W4	W5
none	**78.6**	9.3	4.6	3.3	4	0
W1	24	**73.3**	2	0	0.6	0
W2	6.6	4	**88.6**	0.6	0	0
W3	16.6	2	4.6	**74.6**	2	0
W4	16	0.6	2	2	**78.6**	0.6
W5	29.3	2.6	0.6	3.3	4.6	**59.3**

Table 4 shows results of post-hoc pairwise comparisons between the focus conditions with Bonferroni adjustments. It can be seen that focus on Word 5 has significantly worse recognition rate than neutral, focus on Word 2 and focus on Word 4. In contrast, there is no significant difference between any other two focus conditions.

Table 4. *Results of post-hoc pairwise comparisons. The mean difference is significant at the .05 level.*

Focus Type (I)	Focus Type (J)	Mean Difference (I-J)	Sig.
None	W1	5.334	1.000
	W2	-10.000	.713
	W3	4.001	1.000
	W4	-.001	1.000
	W5	19.335*	**.009**
W1	W2	-15.334	.177
	W3	-1.333	1.000
	W4	-5.335	1.000
	W5	14.001	.598
W2	W3	14.001	.315
	W4	9.999	1.000
	W5	29.335*	**.004**
W3	W4	-4.002	1.000
	W5	15.334	.258
W4	W5	19.336*	**.015**

4. Discussion

Results from Table 3 show that recognition rates of all focus positions were high except for final focus. That the lowest recognition rate is for final focus is in line with many other studies where the final focus had the lowest rate of identification compared to other positions [3]. The most likely reason is that the lack of PFC impedes the easy recognition of focus in this position [13]. This is supported by Figure 2, which displays time-normalized mean F_0 contours of the three speakers whose utterances were used as the perception stimuli in the present study. It can be seen that the F_0 increase by final focus relative to neutral focus is just as substantial as in other focus positions. The only thing missing compared to the non-final focus is the compressed F_0 (and intensity) after the focused word, since there are no words following the final word. The second word has the highest rate of focus identification, followed by the fourth word and neutral focus.

The high recognition rate of focus on the second word can be explained from a syntactic perspective. The basic pattern of Persian sentence prosody is that every major class word receives an accent. From the neutral-focus contour in Figure 2 we can see that all words are accented except the second word, i.e. *babaye* 'father-EZ'. Generally speaking, under some discourse conditions, the head noun of a definite noun phrase may be unaccented when post-modified, as is the case with the head noun *babaye* inside the noun phrase *babaye nili* 'Nili's father' [19]. This usual lack of pitch accent on the second word apparently had a consequence on the focus perception. That is, when asked to put focus on the second word, speakers in fact accented a constituent that would receive no accent in the neutral pronunciation of the sentence. Thus, the focus on the second word involves one extra structural manipulation (i.e. accenting a structurally unaccented word) compared to that on the other words in the sentence which are structurally accented words. This may have made the second word even more salient than the other focused words, and so helped the subjects to more efficiently identify focus on this word.

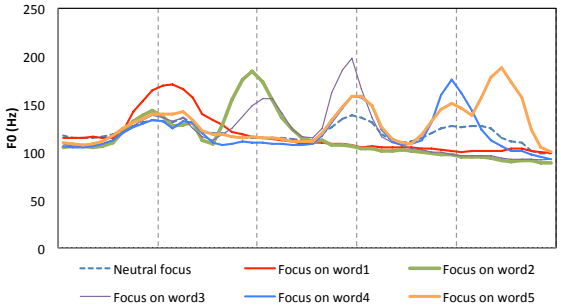


Figure 2. *Time-normalized mean F_0 contours of all the sentences uttered by speakers with* minimum*, maximum and median standard deviations. Each curve is an average of 25 repetitions.*

In Persian, like Uygur [26], initial focus was mostly confused with neutral focus (24%). But in terms of confusion of final focus with neutral focus, with 29.3% wrong identification, Persian acts like Beijing Mandarin [5].

Speakers of Iranian and Turkic languages (both PFC languages) have been in contact since pre-Islamic times [17]. However, to compare with each other, phonetic realization of prosodic focus in Persian and Turkish as two exemplars of Iranian and Turkic languages is not similar in pre-, on- and post-focus regions. To name a few, contrary to Persian, Turkish does not show significant changes in F_0 in on-focus region. And it was also observed that in Turkish pre-focus region is raised in F_0 while in Persian there is no change in mean F_0 in this region. Furthermore, in Turkish, unlike Persian, the acoustic differences in the three above-mentioned regions are not independent of the position of focus. It is worth noting that Azeri language as a Turkic language which is also spoken in Iran is heavily influenced by Persian as an Iranian language. Thus, since Azeri has a common proto language with Turkish, on the one hand, and its close contact with Persian on the other hand, the prosodic focus of Azeri and its cross-comparison with Persian and Turkish are worth investigating.

In addition, since Persian has been in close contact with Arabic and even some part of the population (2%) in Iran is Arab, findings from the study of prosodic focus for Arab-speaking areas and its comparison with Persian and the other dialects of Arabic like Hijazi, Lebanese [4] and Egyptian [9] might reveal further implications.

5. Conclusions

Considering the results of the present perception experiment, it can be concluded that post-focus compression of F_0 and intensity is a highly important acoustic cue for the recognition of prosodic focus in Persian [30]. Its presence in non-final position leads to an average of over 78% focus recognition, whereas its absence in final position leads to less than 60% of focus recognition. Therefore, the overall recognition rate is 75% somewhere between the PFC languages like Beijing

Mandarin [5] and Uygur [26] with 90%, and languages without PFC like Taiwanese with 60% [5]. This perception experiment thus provides another piece of evidence that Persian, like English, Beijing Mandarin, Japanese, Turkish and Tibetan, can be categorized as a PFC language. This finding therefore adds yet another piece to the overall picture of focus typology.

6. Acknowledgements

We thank all subjects who helped us to carry out this research.

7. References

[1] Abolhasanizadeh, V., Bijankhan, M. and Gussenhoven, C., "The Persian pitch accent and its retention after the focus". Lingua 122, 1380-1394, 2012.

[2] Boersma, P., "Praat, a system for doing phonetics by computer". Glot International 5:9/10: 341-345, 1992-2013.

[3] Botinis, A., Fourakis, M. and Gawronska, B., "Focus identification in English, Greek and Swedish". In Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, 1557-1560, 1999.

[4] Chahal, D., "Phonetic cues to prominence in Lebanese Arabic". In Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona. pp. 2067-2070, 2003.

[5] Chen, S.-w., Wang, B. and Xu, Y., "Closely related languages, different ways of realizing focus". In proceedings of Interspeech 2009, Brighton, UK, 1007-1010, 2009.

[6] Comrie, B., *The world's major languages*, 2ed edition, Routledge, London, 2009.

[7] Cooper, W., S. Eady. and P. Mueller., "Acoustical aspects of contrastive stress in question-answer contexts", JASA 77(6): 2142-2156, 1985.

[8] Eslami, M., Šenaxt-e næva-ye goftar-e zæban-e farsi væ karbord-e an dær bazsazi væ bazšenasi-ye rayane'i-ye goftar [The prosody of the Persian language and its application in computer-aided speech recognition]. PhD thesis, Tehran University, 2000.

[9] Hellmuth, S., "Focus-related pitch range manipulation (and peak alignment effects) in Egyptian Arabic". In Proceedings of Speech Prosody 2006, Dresden, Germany. pp. PS4-12-164, 2006.

[10] Hosseini, A., "L1 interference in L2 prosody: Contrastive focus in Japanese and Persian". Journal of Logic Language and Information 03/2013; 11:55-67, 2013.

[11] Ipek, C., "Phonetic realization of focus with no on-Focus pitch range expansion in Turkish". In Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 2011.

[12] Kahnemuyipour, A., *The syntax of sentential stress*. Oxford: OUP, 2009.

[13] Liu, F. and Xu, Y., "Parallel encoding of focus and interrogative meaning in Mandarin intonation", Phonetica 62: 70-87, 2005.

[14] Mahjani, B., "An Instrumental study of prosodic features and intonation in Modern Farsi (Persian)", M.Sc. thesis, retrieved from: http://www.ling.ed.ac.uk/teaching/postgrad/mscslp/archive/dissertations/2002-3/behzad_mahjani.pdf. 2003.

[15] Mixdorff, H., "Quantitative tone and intonation modeling across languages". In proceedings of International Symposium on Tonal Languages, Beijing, 137-142, 2004.

[16] Pell, M. D., "Influence of emotion and focus on prosody in matched statements and questions", Journal of the Acoustical Society of America 109: 1668-1680, 2001.

[17] Perry, J., "Turkic-Iranian contacts: linguistic contacts," Encyclopedia Iranica, online edition, 2013, available at http://www.iranicaonline.org/articles/turkic-iranian-contacts-i-linguistic (accessed on 27 December 2013).

[18] Rump, H. H. and Collier, R., "Focus conditions and the prominence of pitch-accented syllables", Language and Speech 39: 1-17, 1996.

[19] Sadat-Tehrani, N., "Intonational grammar of Persian", doctoral dissertation. Manitoba: University of Manitoba, 2007.

[20] Sahkai, H., Kalvik, M.-L. and Mihkla, M., "Prosody of contrastive focus in Estonian," in Proceedings of Interspeech 2013, Lyon, France, 315-319, 2013.

[21] Scarborough, R., "The intonation of focus in Farsi". UCLA working papers in phonetics, No. 105, pp. 19-34, 2007.

[22] Taheri-Ardali, M. and Xu, Y., "Phonetic realization of prosodic focus in Persian", in Speech Prosody 2012, Shanghai, 326-329, 2012.

[23] Towhidi, J., "Studies in the phonetics of Modern Persian. Intonation and related features" (=Forum Phoneticum 2). Hamburg: Helmut Buske [originally dissertation, London 1973], 1974.

[24] Vahidian-Kamyar, T., *Næva-ye goftar dær farsi [Melody of speech in Persian]*. Mashhad, Ferdowsi University Press, 2001.

[25] van Heuven, V. J., and de Jonge, M., "Spectral and temporal reduction as stress cues in Dutch," Phonetica 68, no. 3:120-132, 2011.

[26] Wang, B., Qadir, T. and Xu, Y., "Prosodic encoding and perception of focus in Uygur". Chinese Journal of Acoustics. 2013. (in Chinese)

[27] Xu, Y., "Effects of tone and focus on the formation and alignment of F_0 contours", Journal of Phonetics 27, 55-105, 1999.

[28] Xu, Y., "Speech melody as articulatory implemented communicative functions", Speech Communication 46, 220-251, 2005.

[29] Xu, Y., Xu, C. X. and Sun, X., "On the temporal domain of focus", International Conference on Speech Prosody 2004, Nara, 2004.

[30] Xu, Y., Chen, S.-w. and Wang, B., "Prosodic focus with and without post-focus compression (PFC): A typological divide within the same language family?" The Linguistic Review, 29, 2012.

[31] Xu, Y. and Xu, C. X., "Phonetic realization of focus in English declarative intonation", Journal of Phonetics 33, 159-197, 2005.

Interpreting Final Rises: Task and Role Factors

Catherine Lai

Centre for Speech Technology Research,
School of Informatics, University of Edinburgh, United Kingdom
clai@inf.ed.ac.uk

Abstract

This paper examines the distribution of utterance final pitch rises in dialogues with different task structures. More specifically, we examine map-task and topical conversation dialogues of Southern Standard British English speakers in the IViE corpus. Overall, we find that the map-task dialogues contain more rising features, where these mainly arise from instructions and affirmatives. While rise features were somewhat predictive of turn-changes, these effects were swamped by task and role effects. Final rises were not predictive of affirmative responses. These findings indicate that while rises can be interpreted as indicating some sort of contingency, it is with respect to the higher level discourse structure rather than the specific utterance bearing the rise. We explore the relationship between rises and the need for co-ordination in dialogue, and hypothesize that the more speakers have to co-ordinate in a dialogue, the more rising features we will see on non-question utterances. In general, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data, and when we link them to speaker states such as uncertainty or submissiveness.

Index Terms: Intonation, task-oriented dialogue, rises.

1. Introduction

The question of what prosody contributes to meaning is a key problem for both automated spoken language understanding and theories of semantics and pragmatics. In particular, a good number of studies have investigated how utterance final pitch rises and falls relate to epistemic and affectual states of speakers and how this relates to tasks such as dialogue move detection. Such studies generally examine how prosody affects the interpretation of the carrier utterance in the immediate context, e.g. whether a cue word is interpreted as a backchannel or not. While the local context clearly has a large effect on how prosody is interpreted, we would also like to know what impact higher level features such as task and role have as well.

One reason we expect higher level features to affect the interpretation of prosody follows from the incongruence of findings based on single dialogue types. For example, a correlation between rises and backchannels has been reported in map task dialogues in Bari Italian [1], Swedish [2], and Dutch [3], as well as in other game corpora in English [4, 5, 6]. However, these sorts of results are absent from studies of more free-form conversational dialogues in English [7, 8, 9] and Hindi [10]. We would like to know whether the differences in rise distributions from these backchannel studies extend to other sorts of dialogue moves, and if so, why.

To examine this, we look at the IViE (Intonational Variation in English) corpus [11]. The corpus contains speech of various styles including isolated read sentences as well as spontaneous conversational and task-oriented dialogues (the map task) from speakers of urban regions of the United Kingdom. In this paper we look at the boundary intonation of speakers from Cambridge (i.e. Standard Southern British English) in these different modes of speech. The motivation for looking at this dialect in particular is that out of those included in the corpus the intonation pattern for this region's declarative statements has been found to be the most pervasively falling or low at the boundary in read speech. Thus, rises are more likely to be seen as deviations from the canonical. When we observe rises, we expect them to mean something more than just a phonological boundary. So, looking at this data enables us to look at the effects of task and role on the frequency of rises, as well as giving as a more general view on how task-oriented and conversational dialogue differ.

2. Background

Direct comparisons of conversational and task-oriented speech have mainly focused on the greater need for affectual/emotional modelling in the former [12, 13]. While automatic role recognition has received more attention recently [14, 15], studies have not generally investigated in any detail how prosody varies with different role/move categories. However, some investigations of this type have been carried out with the goal of improving expressive speech synthesis. For example, [16] find that 'Assess' moves in the AMI corpus were produced with tenser voice quality, while project managers had higher average F0 and vocal effort. Dominant participants exhibited 'louder' voice quality features in [17]. While these studies provide broad descriptions for specific roles, they don't look at the contribution of intonation features like terminal rises in any detail.

Theoretical studies have analyzed rises as expressing uncertainty [18, 19, 20] or submissiveness [21, 22]. So, we might expect to find more rises in the productions of participants in socially submissive roles rather than leader type roles. However, empirical studies suggests that the distribution of rises depends heavily on situational and cultural conventions. For example, in a qualitative analysis of sorority speech, rises were used by senior members to take and hold the floor in monologues, while they were perceived as expressing uncertainty in narratives by uninitiated members of the group [23]. Similarly, in a comparison of several dialogue types, rises were found to be more prevalent in dialogues where one person has a socially dominant role, e.g. academic supervision versus informal office conversations [24] . Moreover, it was the socially dominant participant who produced the rises.

The latter study doesn't conditionalize over different move types, so it's not clear whether the more one-sided conversations simply involve more questions. In particular we would like to know when rises occur on sentence types that are canon-

ically falling in the dialect we are examining, e.g. declarative statements (informing moves) and imperatives (instructions) [25, 11]. Rises have also been analyzed in terms of contingency [26, 27], hearer dependence [28] about the rise carrying utterance. So, we would like to know whether the distribution of rises in a dialogue can be explained in terms of whether moves need explicit ratification or not. This would predict a higher number of affirmatives following rising moves. More generally, we would like to know if the distribution of rises can be adduced from the turn-taking structure of the dialogues, i.e. whether rises give or hold the floor.

3. Experimental Setup

3.1. Data

The IViE corpus was developed to systematically study differences across regions, speakers, and styles [11]. As mentioned previously, we look at data from Cambridge speakers as the most consistently 'falling' dialect in the collection. Twelve speakers (6 male, 6 female) from each region were recorded between 1997-2000. The speakers were 16 years old at the time of recording and had been born in and grown up in the region. The recordings include a mixture of read and spontaneous speech, of which we use the following:

- *Map task (map):* Each participant was given a map of a small town. Participants took one of two roles: Instruction *giver* and *follower*. The goal was for the giver to explain a pre-defined route around town on their map to the follower, who traced it out on their own map. The task ended when the route was completed to the satisfaction of both participants. Maps differed in place names and locations of landmarks, so speakers had to work to establish common ground. Speakers were separated by screens so they could not see each other. More details about this task can be found in [29].
- *Free conversation (conv):* Participants discussed smoking, face-to-face. Speakers had the same role, which was simply a *participant*

3.2. Segmentation and Annotations

Only short portions from four speakers were transcribed and annotated at sentence type level for each of these dialogue types in the official IViE release, so additional annotation was undertaken. The dialogues were manually segmented into utterances corresponding to whole meaning units rather than phonological phrases (cf. [30]). This was done as a conservative measure of the frequency of rises. Sentential segmentation delimited whole propositions including any embedding. Similarly, imperative utterances mapped to one action (i.e. one segment of the route). A number of sub-sentential clauses also formed separate utterances, e.g. an NP or VP as an answer or a modifier separated by an affirmative, which were as marked as XPs. Utterances were labelled with sentence (syntactic) and move types:

- Sentence type: Declarative (dec), Imperative (imp), Polar question (yno), Wh-question (whq), Tag (tag) question, Affirmative (affirm), Negative (neg), Cue word (cw), If antecedent (IFA), XP (XP).
- Broad dialogue moves: Affirm, Neg, Contra (direct contradictions), CW (cue words), Inform, Instruct, Q (non-syntactically marked question), YNQ (polar question), WhQ (wh-question), Tag (Tag-question), sync (synchronize).

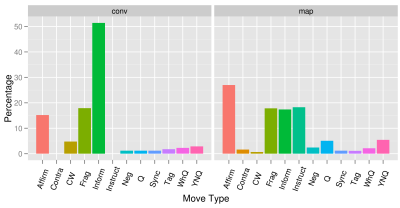


Figure 1: *Proportions of moves:* $N_{conv} = 430$, $N_{map} = 1287$.

The rationale for using such broad move types was to keep the annotation in terms of easily identifiable categories which could be refined in the future. In many cases, one sentence type dominated a move type (e.g. wh-questions). The main points of variation were in the declaratives which we see as instructions, informing moves, and questions, amongst other moves. Similarly, instructions, while primarily imperative in form, were also expressed as declaratives, polar questions or if-antecendents ('If you could go to the church'). The *sync* category captured utterances in the map task like 'You should be at the Anne's Arms' which were not quite questions, instructions or inform moves. A backchannel category was not included as it was not clearthat the distinction could be reliably made [31]. Moreover, it did not seem that any of the affirmatives in the map task could be clearly classified as simple signals of attention. In the investigations to follow we will concentrate on the most populous and easy to identify categories: Affirmative, Instructions and Inform moves. Utterance segmentation of the dialogues resulted in 430 and 1287 utterances for the conversational and map task sets respectively. The distribution of moves is shown in Figure 1.

3.3. Boundary pitch features

The target area for analysis was the stretch of speech from the last prominence rather than the last word. Extension away from the last word was generally due simply to stress assignment in compounds (e.g. *bowling* alley) or deaccenting of pronouns, (e.g. *about* it). Utterances with speaker overlap at the target were excluded from the prosodic analysis (3% conv, 4% map).

The F0 contour data on target segments was extracted using the Praat autocorrelation method. Parameter settings were automatically determined using the method described in [32]. Utterances which produced less than 5 F0 points were discarded. The F0 data was normalized into semitones relative to the median F0 value (Hz) for each speaker using data produced in all IViE tasks including read speech [11]. F0 contours were smoothed using a Butterworth filter and contours were approximated using Legendre polynomial decomposition of order 4 (cf. [33]). Instead of making categorical judgements about shape, we will instead look at first three coefficients where LC1 increases with overall pitch *height*, LC2 increases with positive contour slope (or *tilt*), and LC3 with *convexity*. Positive LC2 and LC3 indicate a fall-rise contour with an overall rising trend, negative LC2 and LC3 indicate a rise-fall contour, while values close to the origin indicate a flat contour. Previous work has described the relationship between these features and ToBI perceptual labels [34].

4. Results

4.1. Rises across dialogue types

Figure 2 shows the distribution of values for LC2 ad LC3 by move type. Overall, the map-task data can be characterized as

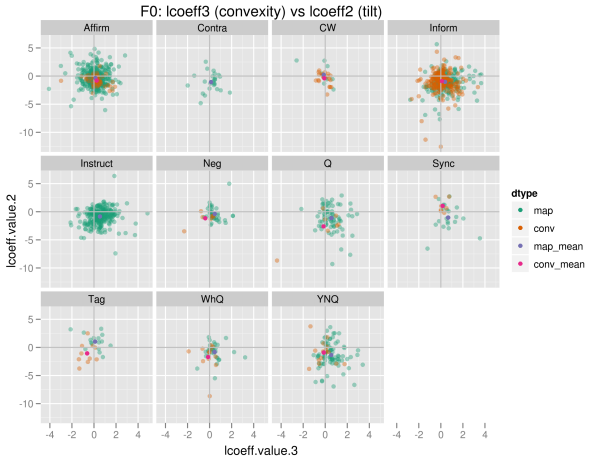


Figure 2: *Tilt (LC2) and Convexity (LC3), by move type, with means.*

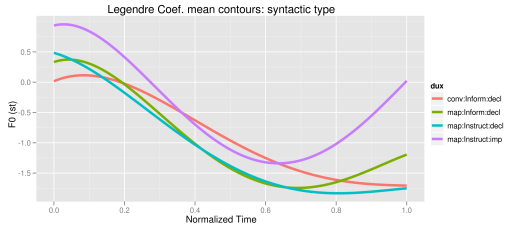


Figure 3: *Mean contours Instruct and Inform moves based on Legendre coefficients grouped by syntactic type.*

having more rising features, with these mainly coming from instructions and affirmatives. We see that Instruct moves are mostly situated in the positive LC3 space, indicating a fall-rise contour. The distribution of affirmatives in the map task extends into the positive LC2 space, indicating rising tilt.

Inform and Instruct moves make up 44% of the utterances in the map task, while 66% of conversation moves were Informs. These moves provide most of the 'new' content in the dialogue and are canonically falling in Southern Standard British English [25]. So, this subset of moves are good indicators that task affects the distribution of rising features. Figure 3 shows mean contours for Inform and Instruct moves which are declaratives, as well as the imperatives for comparison. Within the Instruct moves, syntactic imperatives are particularly rising compared to declarative instructions. We see that, on average, Inform declaratives are more rising in the map task than in conversational speech.

In order to quantify this we model the relationship between Legendre polynomial coefficients and dialogue factors (role, move, and syntactic type).[1] We fit (non-nested) multilevel linear models predicting values of LC1, LC2 and LC3. The model parameters were estimated using the package `lmer` in `R`. We only see significant effects for role when we look at convexity (LC3). The effect estimates and confidence intervals shown in Figure 4 indicates that being an instruction giver (1) increases convexity, while simply being a conversational participant (3) decreases it. In terms of moves, there is a signficant positive effect on convexity for imperatives (i.e. fall-rise). The effect of

[1]Note: Role encodes the task.

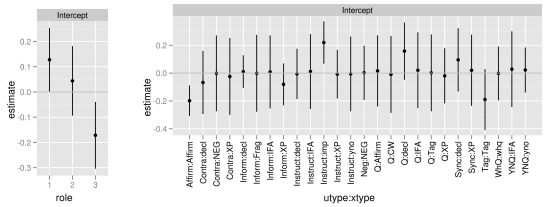


Figure 4: *F0 convexity (LC3) parameter estimates.*

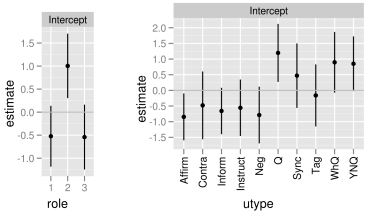


Figure 5: *Speaker change parameter estimates: 1=giver, 2=follower, 3=conversational participant.*

Affirm moves was negative on convexity, but positive on height and tilt, with a positive value for the instruction follower. This again points to there being more rising affirmatives in the task-oriented dialogue. Interestingly, yes/no and declarative questions have a negative relationship with tilt. This suggest that the specific questioning use of rises is less in play in these sorts of dialogues.

4.2. Turn-taking

Since Cambridge imperatives and declaratives are usually described falling at the boundary we would like to know whether the rises we see in the map task data can be attributed to other discourse factors like turn-taking. To do this, we fit parameters for multilevel logistic regression models (stay=0, switch=1). We compare a model containing speaker, role and move factors with one extended with Legendre coefficients as predictors (adding syntactic indicators did not improve the model fit).

Looking at the parameter estimates in Figure 5, we see that being the instruction follower (role 2) increases the probability of a switch by 25%. The trend for the other two roles is to hold the floor. In terms of move type, we see that the broad class of question moves increase the probability of switching, while content adding moves decrease it, although we saw previously that these generally had a falling tilt. With respect to move-role interaction, affirmatives produced by the instruction giver are likely to result in stays. That is, instruction givers seem to use affirmatives as a ready signal. The effects of the other move-role combinations are quite small in comparison.

Figure 6 shows significant positive effects for LC2 and LC3 (estimated coefficients: 0.12 and 0.2 respectively.) However, the magnitudes of these effects are relatively small compared to the effects of role and move. For example, when LC2 equals 1 we get an approximate 3% increase in probability of a switch, where the 95% interval of values in the observed data for LC2 is $(-4.15, 2.30)$ (LC3: $(-1.39, 2.12)$). So it seems that having higher tilt or convexity nudges up the probability of a speaker switch, but the contribution is not as strong as that of move or role. As we would expect, question type moves are generally turn giving irrespective of whether the utterance has a rising or falling boundary. To see whether rising features have different

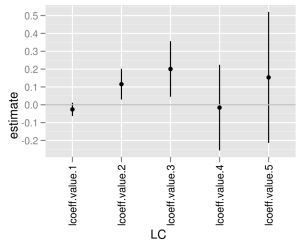


Figure 6: *Turn-taking: LC data as individual level predictors.*

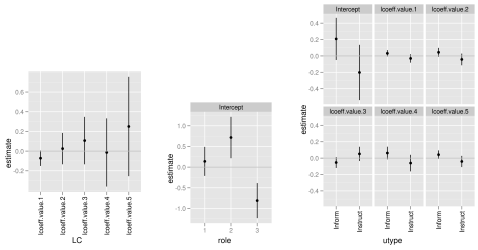


Figure 7: *Predicting Affirmatives after Inform/Instruct moves with LC features as individual level predictors and predictors on the move group.*

effects on different move types for turn-taking we extend the model to allow coefficients for the LC values to vary by move type. However, the difference between the previous two models is very small: DIC decreases from 1809 to 1802 where the new model adds many more parameters. So, it does not appear that rising features have much of a role to play in determining who takes the floor.

4.3. Eliciting Affirmatives

To check whether rising features signal a need for explicit ratification/agreement, we fit a multilevel logistic model predicting whether or not Inform or Instruct moves are followed by Affirm moves with the same predictors as above but coding an Affirm response as 1, and other responses as 0. If rises do signal a need for ratification we would expect to see that the probability of an affirmative increase with LC2 and LC3. After controlling for the higher level dialogue features, we see that the effects of the contour shape features to be, again, dwarfed by the effect of role. Estimated coefficients for the pitch features are around ± 0.05 at the move level, resulting in about a $\pm 1\%$ change in the likelihood of an affirmative response for every LC coefficient unit (cf. Figure 7). The effects are similarly close to zero at the individual level. On the other hand, being the map task follower, again, increases the probability of the next move being an affirmative by about 18%. In general, we don't see that pitch shape features on content moves are predictive of whether or not that move will be explicitly ratified.

5. Discussion

Our goal was to find out if higher level effects like task, role and move type had an effect of whether an utterance was produced with rising features. Looking at the Cambridge IViE data, it seems that we do get more rising features in the task-oriented speech than the free conversation, mostly with respect to instructions and affirmatives. While we saw that significant positive effects of tilt and convexity coefficients for speaker changes and production of affirmatives, we also saw that these effects were dwarfed by the effect of role. So, whatever the rising features are doing on these turns, it does not seem that their main role is to manage turn-taking. Instead, turn-taking strategy seems mostly dictated by the higher level, task structure of the dialogue.

How do our results fit with other analyses of rises? It is hard to reconcile the data with accounts that link rises to *propositional* attitudes (i.e. attitudes about the content bearing the rise). For example, we wouldn't want to associate rising features with how we usually think of epistemic uncertainty (contra [18]'s Maxim of Pitch), nor would be want to associate them with social submissiveness [21] or lack of speaker commitment [28], since these features are predominantly used by by the instruction giver, i.e. the situationally dominant speaker. In fact, if the instruction giver is uncertain about anything, it's not about the actual route. Instead it seems to be about whether the follower can or will carry out the task, i.e. discourse structural uncertainty. Similarly, the instruction giver is at the mercy of the follower in terms of task completion. So, at a glance we might say that the map-task shows more rising features because it just has more contingent elements (cf. [27, 26]. However, we saw that rises don't seem to elicit ratification in terms of explicit affirmative responses. So, if rises do signal contigency it is not necessarily about the current utterance.

An alternative is that the speaker deploys rising features because it is important to attend to whether a task is open or closed, since each subtask is dependent on the subtask before it. That is, we get more rises because there is a more well defined subtask structure (cf. [30]) and participants need a high level of common ground co-ordination in order to reach the end-goal of the dialogue. The need for co-ordination is much less pressing in conversational speech where participants are basically voicing opinions. So, the notion of contingency we are dealing with is at a higher level than accepting single instructions. From this point of view, explicit affirmation (rise or not) is a good strategy for the map-task follower, but rises primarily signal that there is more to come [35]. Thus we expect to see more rises in dialogues where greater quality of co-ordination is required.

6. Conclusion and Future Work

In this paper we saw that that higher level discourse factors, like task and role, have an effect on whether an utterance is produced with rising features or not. Overall, we found that content providing utterances in map-task dialogues had greater convexity than those from the conversational dialogue. Most of this seemed to come from instruction moves which often had a distinct fall-rise shape. While the rate of Affirmative moves was higher in the map task, we didn't find any strong link between rising features – higher LC2 and LC3 – and affirmative responses, or more generally speaker switches or stays. This state of affairs sits best with discourse structural analyses of rises, rather than notions like submissiveness or uncertainty. It appears that the more speakers have to co-ordinate through verbal signals, the more rising features we expect to see. So, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data.

Further work looking at the relationship between frequency of rises and the overall quality of task-completion, as well as comparision to other dialects, especially default rising ones such as Belfast English, and styles, will help complete the picture of where intonation fits into semantic/pragmatic theories.

7. References

[1] M. Savino, "The intonation of backchannels in italian task-oriented dialogues: cues to turn-taking dynamics, information status and speakers attitude," in *5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 2011.

[2] M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé, "Prosodic features in the vicinity of silences and overlaps," in *Proc. 10th Nordic Conference on Prosody*. Citeseer, 2008, pp. 95–105.

[3] J. Caspers, "Melodic characteristics of backchannels in dutch map task dialogues," in *Sixth International Conference on Spoken Language Processing*, 2000.

[4] B. Hockey, "Prosody and the role of okay and uh-huh in discourse," in *Proceedings of the eastern states conference on linguistics*. Citeseer, 1993, pp. 128–136.

[5] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of ICPhS 2007*, 2007, pp. 1065–1068.

[6] A. Gravano, "Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue," Ph.D. dissertation, Columbia University, 2009.

[7] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, and M. Meteer, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[8] N. Ward, "Pragmatic functions of prosodic features in non-lexical utterances," in *Speech Prosody 2004*, vol. 4, 2004, pp. 325–328.

[9] K. Truong and D. Heylen, "Disambiguating the functions of conversational sounds with prosody: the case of'yeah'," in *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), 2010.

[10] S. Prasad and K. Bali, "Prosody cues for classification of the discourse particle" hã" in hindi," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] E. Grabe, "Intonational variation in urban dialects of english spoken in the british isles," in *Regional variation in intonation*, P. Gilles and J. Peters, Eds. Linguistische Arbeiten, Tuebingen, Niemeyer, 2004, pp. 9–31.

[12] N. Campbell, "Developments in corpus-based speech synthesis: Approaching natural conversational speech," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 376–383, 2005.

[13] Y. Wilks, R. Catizone, S. Worgan, and M. Turunen, "Review: Some background on dialogue management and conversational speech for dialogue systems," *Computer Speech and Language*, vol. 25, no. 2, pp. 128–139, 2011.

[14] A. Vinciarelli, F. Valente, S. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus," in *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2011, pp. 374–379.

[15] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 238–247.

[16] M. Charfuelan and M. Schröder, "Investigating the prosody and voice quality of social signals in scenario meetings," *Affective Computing and Intelligent Interaction*, pp. 46–56, 2011.

[17] M. Charfuelan, M. Schröder, and I. Steiner, "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings," in *Interspeech'10*, 2010.

[18] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1-2, pp. 31–43, 2002.

[19] M. Nilsenova, "Rises and falls. studies in the semantics and pragmatics of intonation," Ph.D. dissertation, University of Amsterdam, 2006.

[20] B. Reese, "Bias in questions," Ph.D. dissertation, University of Texas at Austin, 2007.

[21] A. Merin and C. Bartels, "Decision-Theoretic Semantics for Intonation," Universitt Stuttgart and Universitt Tübingen, Tech. Rep. Bericht nr. 88., 1997.

[22] C. Gussenhoven and A. Chen, "Universal and Language-Specific Effects in the Perception of Question Intonation," in *Sixth International Conference on Spoken Language Processing*. ISCA, 2000.

[23] C. McLemore, "The pragmatic interpretation of english intonation: Sorority speech," Ph.D. dissertation, University of Texas at Austin, 1991.

[24] W. Cheng and M. Warren, "//CAN i help you //: The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English," *International Journal of Corpus Linguistics*, vol. 10, no. 1, pp. 85–107, 2005.

[25] A. Cruttenden, *Intonation*. Cambridge: Cambridge Univ Press, 1997.

[26] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgen, and M. Pollack, Eds. Cambridge: MIT Press, 1990.

[27] C. Gunlogson, "A question of commitment," *Belgian Journal of Linguistics*, vol. 22, no. 1, pp. 101–136, 2008.

[28] M. Steedman, "Information Structure and the Syntax-Phonology Interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, September 2000.

[29] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[30] M. Swerts and R. Geluykens, "Prosody as a marker of information flow in spoken discourse," *Language and speech*, vol. 37, no. 1, pp. 21–43, 1994.

[31] C. Lai, "Prosodic Cues for Backchannels and Short Questions: *Really?*" in *Proceedings of Speech Prosody 2008, Campinas, Brazil, May 2008*, 2008.

[32] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction," in *Presented at the 2nd PanAmerican/Iberian Meeting on Acoustics, Cancun, Mexico, 15-19 November 2010*, 2010.

[33] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *The Journal of the Acoustical Society of America*, vol. 118, p. 1038, 2005.

[34] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech*, vol. 50, no. 3, pp. 281–310, 2007.

[35] C. Bartels, *The intonation of English statements and questions: a compositional interpretation*. Routledge, 1999.

Spontaneous speech corpus data validates prosodic constraints

Philippe Martin

UMR 7110, LLF, UFRL, Université Paris Diderot, ODG, Place Paul Ricœur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

Abstract

In the Autosegmental-Metrical model, the prosodic structure is defined as a hierarchy of Accent Phrases (AP). Groups of AP form intermediate prosodic phrases ip, which in turn are grouped into Intonation Phrases IP, and finally sequences of IP form the sentence intonation unit. In this hierarchy several constraints affect the prosodic structure, such as the AP 7 syllables rule, the stress clash conditions, eurhythmicity and syntactic clash.

These constraints have been established essentially from read sentences data. They lead to an experimental justification in the observed synchronization of AP's syllabic chunking by Delta brain waves.

This paper investigates the validity of the prosodic structure constraints on spontaneous speech data in French, as well as the adequacy of the Delta waves characteristics to synchronize AP data.

Index Terms: prosodic structure, accent phrase, spontaneous speech, Delta waves, eurhythmy.

1. Introduction

In the classical Autosegmental-metrical approach, the prosodic structure is defined as a hierarchical organization of minimal prosodic units, the Accent Phrases (AP, aka prosodic words, rhythmic groups, etc.), a sequence of syllables which contain one lexical stress (or a metrically strong syllable). Groups of AP's form intermediate prosodic phrases ip, groups of ip form Intonation Phrases IP, and groups of IP form the complete sentence intonation. In this hierarchy, the whole sentence intonation can eventually be reduced to one single IP, which may contain a single ip, which itself may include a single AP.

Furthermore, the prosodic structure is constrained by a set of rules [8]:

a) AP 7 syllables rule;
b) Stress clash;
c) Eurhythmy;
d) Syntactic clash.

The maximum number of 7 syllables per AP was already mentioned by Meigret [12]. Stress clash pertains to avoidance of two consecutive stressed vowels in sentence realizations [12]. Eurhythmy ([3], [8], [14]) determines the tendency to either balance the number of syllables in successive IP's (or possibly ip's), or compensate the duration of enunciation of successive IP's containing an unbalance number of syllables. Finally, syntactic clash defines AP's allowed alignments with sequences of grammatical categories, which for instance cannot group a Verb followed by a determinant (something like *which for the* or *followed by a* in a single AP). In these examples, in the syntactic structure, syntactic units (words) are dominated immediately by nodes that group (i.e. are fathers) units which do not belong to the same AP.

Another characteristic pertains to the composition of AP, assumed to contain a single lexical word (Verb, Noun, Adjective or Adverb) possibly accompanied by grammatical words (Pronoun, Conjunction…) [2]. The validity of the rule will be evaluated as well.

2. Testing the hypotheses

The prosodic structure constraints originate mostly from observations pertaining to read sentences built by linguists. The goal of this paper is to evaluate the validity of these constraints for spontaneous speech, and also test a hypothetical cognitive explanation for each of the constraints.

Briefly stated, the cognitive hypothesis assumes that Delta brain waves are synchronized by stressed syllables (ending accent phrases in French) much as syllabic perception is synchronized by Theta waves [5],]6]. This synchronization would operate even if stressed syllables are not in final position.

Delta waves frequency varies from 1 Hz to 4 Hz, i.e. their periods vary from 250 ms to 1000 ms. This suggests that Delta waves are responsible for the conversion of sequences of syllables stored in short-time memory into a higher level linguistic unit, corresponding to AP's [6], [12]. This process timing is limited by the extreme values of Delta periods, whereas minimal period of Theta waves, which synchronize the perception of syllables, is about 100 ms (10 Hz).

The Delta wave hypothesis would be validated for constraints a) and b) if the observed AP longest and shortest duration would not exceed the Delta wave period values, whereas eurhythmicity would be explained if a compression effect would affect AP syllable duration, the shortest AP containing longest syllables, and the longest AP the shortest syllables.

Finally, the syntactic clash constraint could be validated by the total absence of realizations violating this alignment condition in the corpus. In addition, the validity of the AP composition with lexical words will be questioned, as occasionally some examples show AP containing only grammatical words.

These hypotheses are central in the prosodic incremental storage concatenation process proposed in [11]. In this model, acoustically and phonologically differentiated prosodic events trigger various transfer in the listener memory: a) transfer of syllables into another part of memory synchronized by Theta brain wave, b) transfer of syllabic chunks (correspond to AP's) into another short term memory synchronized by Delta waves, and c) concatenation of these sequences of partial processing into an interpretation module synchronized by differentiated prosodic events (various melodic contours).

3. Data analysis

To test the prosodic structure constraints, French data were selected as the absence of lexical stress in French may a priori

lead to more variations in AP number of syllables, as one single Accent Phrase can contain more than one lexical word, as in *le bilan des ventes* "the stock sales" pronounced rapidly, with two nouns *bilan* and *ventes* belonging to the same AP.

Analyzed data were taken from the C-PROM corpus [16]. C-PROM is a transcribed, aligned and annotated corpus, developed among other applications, to evaluate syllabic prominences in French. It includes 24 recordings belonging to 6 different speech styles of francophone speakers originated from Belgium, France and Switzerland. Only French speakers were retained in this study. Details and transcription formats can be viewed on line.

Primary data were then transferred to WinPitch [17], whose routines allow a direct analysis into an Excel sheet of results, giving automatically in one single mouse click (Fig. 1 and Fig. 2):

a) The number of syllables in each AP;
b) The overall AP duration in ms;
c) The average syllabic duration in a given AP;
d) The AP stressed vowel duration;
e) The API vowel transcription;
f) The AP transcription in API;
g) The AP orthographic transcription;
h) Time references of the events.

	A	B	C	D	E	F	G	H	I	J
1	Nb Syl	Duration	Dur / Syl	Vowel Dur	Vowel	Center	Syllables	Text	Start	End
2										
3										
4	4	1193	298	225	y	1.972	sɛtaptity*	cette aptitude	1.607	2.047
5	4	897	224	39	a	4.372	tẽkõmɛmpa*	incon- même pas	4.294	4.385
6	6	731	121	54	e	5.103	ãvidpɔlemike*	envie de polémiquer	4.974	5.121
7	6	857	142	75	ã	5.96	paskəʃyzokuʁã*	parce que suis au courant	5.858	5.985
8	4	797	199	192	u	6.757	mẽtnãdətu*	maintenant de tout	6.587	6.821
9	7	1058	151	155	ã	7.815	safɛkãɛmvẽsẽkãk*	ça fait quand même vingt cinq ans que	7.622	7.909
10	6	896	149	76	ê	8.711	ʃekɔnɛasebjê*	je connais assez bien	8.589	8.737
11	4	575	143	104	ê	9.286	sɛtekʁivê*	cet écrivain #	9.19	9.321
12	6	1238	206	75	ã	11.737	tẽkõtɛstabləmã*	incontestablement	11.63	11.762
13	5	846	169	89	ɛ	12.583	ẽtʁɛgʁãpɔɛt*	un très grand poète	12.524	12.738
14	3	719	239	98	u	13.302	avãtu*	avant tout #	13.077	13.335
15	3	467	155	128	ã	15.893	jaləskã*	il|y a le	15.72	15.936
16	4	655	163	84	ɛ	16.548	dalləpʁɔblɛm*	le problème #	16.421	16.683
17	5	794	158	106	u	18.485	dəplyzãplyluʁ*	de plus en plus lourd	18.304	18.61
18	4	700	175	72	ã	19.185	puiskəmẽtnã*	puisque maintenant #	19.096	19.209
19	3	897	299	187	õ	20.93	sapasjõ*	sa passion	20.484	20.993
20	3	402	134	69	ɛ	22.282	ʒãdanjɛl*	Jean Daniel	22.153	22.396
21	3	381	127	100	ɛ	22.663	mədizɛ*	me disait	22.536	22.697
22	4	859	214	272	ã	23.522	ynʃozmaʁã*	une chose marrante	23.257	23.613
23	2	259	129	35	y	25.001	sɔkyp*	occupe	24.96	25.054

Figure 2. Example of Excel sheet of primary results (*nar-fr speaker*)

The speech styles of the corpus are:

lec-fr: oral reading;
cnf-fr: university conference;
nar-fr: narrative, life story;
pol-fr: political discourse;
jpa-fr: radio news.

The iti-fr itinerary style of the corpus was not retained as recordings were considered too short, whereas the lec-fr recording are kept for comparison with the non-read styles.

Since French has no lexical stress, only boundary tones (in AM terminology) are observed. Their effective realization is sometimes difficult to establish as linked to the effective stressed characteristics of given syllables [10], but the error rate can be estimated at less than 5 %.

The original labelling of stressed syllable into two degrees of stress, noted p and P, were carefully revised. As few occurrences were found questionable, informal perception tests were conducted for possible corrections and adjustments.

An example of AP stress syllable revision is given below.

In cnf-fr, a segment was originally transcribed as:
de deux cent soixante-cinq phrases, dədøsãswasãtsœ̃fʁaz*
with seven syllables pronounced in 1491 ms. But listening more carefully two AP were actually realized :
De deux cent dədøsã* and *soixante-cing phrases* swasãtsœ̃fʁaz (« of two hundred sixty five sentences") with two accent phrases, of respectively 3 and 4 syllables.

The AP's duration are taken from the right boundary of the syllable vowel to the next stressed vowel right boundary. When the stressed syllable is preceded by a pause, the end of the pause is retained as the starting time reference to measure the duration of the current AP ended by the next stressed syllable.

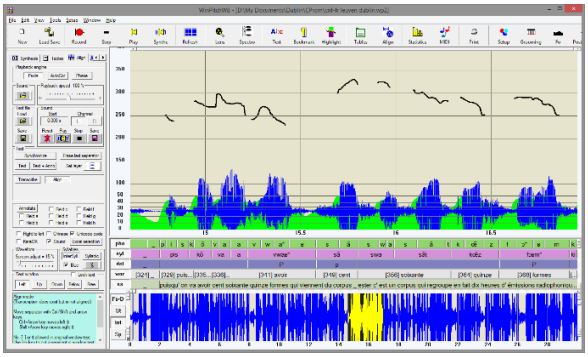


Figure 1. Example of WinPitch display [17]. The second transcription tier displays the syllables in API, with perceived prominence indicated by a star (*cnf-fr speaker*)

4. Results

Table 1 gives the following results pertaining to the hypotheses to be tested:

a) Longest AP duration;
b) Shortest AP duration;

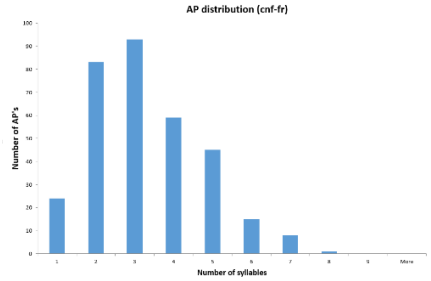
c) Number of stress clash violation.

These values were computed for each recording styles individually, in order to evaluate the possible influence of speech styles on the results.

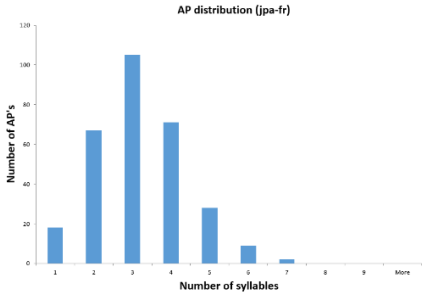
Table 1. Longest and shortest AP duration (in ms) and AP duration vs. number of syntactic clash violation for five speech styles.

Duration	Lec	Cnf	Nar	Pol	Jpa
AP Max	1260	1134	1058	975	1258
AP Min	435	277	354	438	241
Synclash	0	0	0	0	0

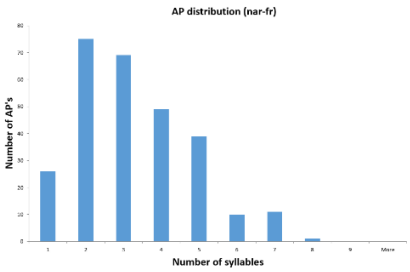
Histograms giving the distribution of the number of syllables per accent phrase are given Fig. 3. They are very similar for the five styles retained.



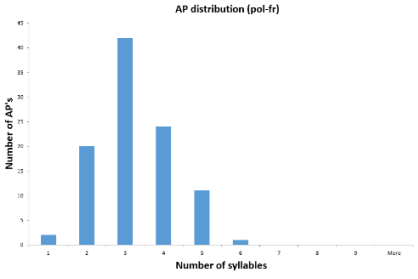
Distribution of cnf-fr number of syllables in AP's



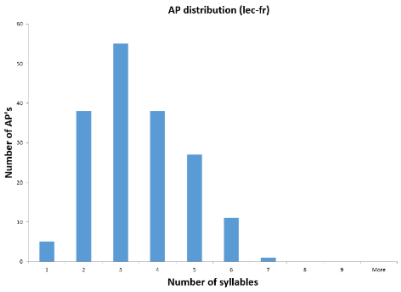
Distribution of jpa-fr number of syllables in AP's



Distribution of nar-fr number of syllables in AP's



Distribution of cpol-fr number of syllables in AP's



Distribution of lec-fr number of syllables in AP's

Fig. 3. Distribution of the number of syllables for the C-PROM styles retained

The only noticeable difference pertains to pol-fr (political speech), which uses a more restrained distribution of AP number of syllables similar to lec-fr, i.e. 3-4 vs 1-7 or 1-8 for the other styles. This suggests that pol-fr style was, at least partially, read speech.

Fig. 4 and Fig. 5 give the corresponding distributions of shortest and longest AP duration in function of the five styles considered.

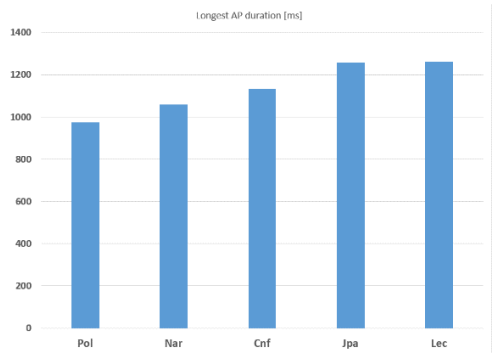


Fig. 4. Longest AP duration in ms

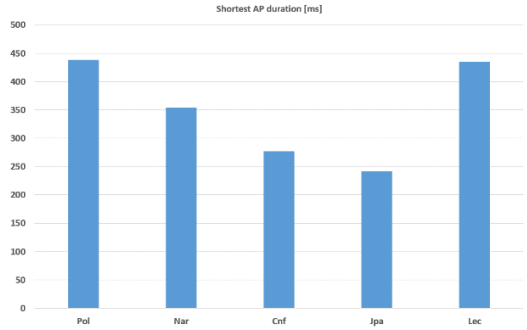


Fig. 5. Shortest AP duration in ms

These distributions show that the spoken news style uses the shortest AP's, whereas the reading and political recordings favor longer minimal AP duration.

The regression line of Fig. 6 demonstrate the compression of AP's duration in function of their number of syllables. When this number increases, the average duration of syllables is reduced allowing a single AP to contain up to about 7 syllables.

The regression lines of the other styles are not shown, as being similar to the one presented (cnf-fr).

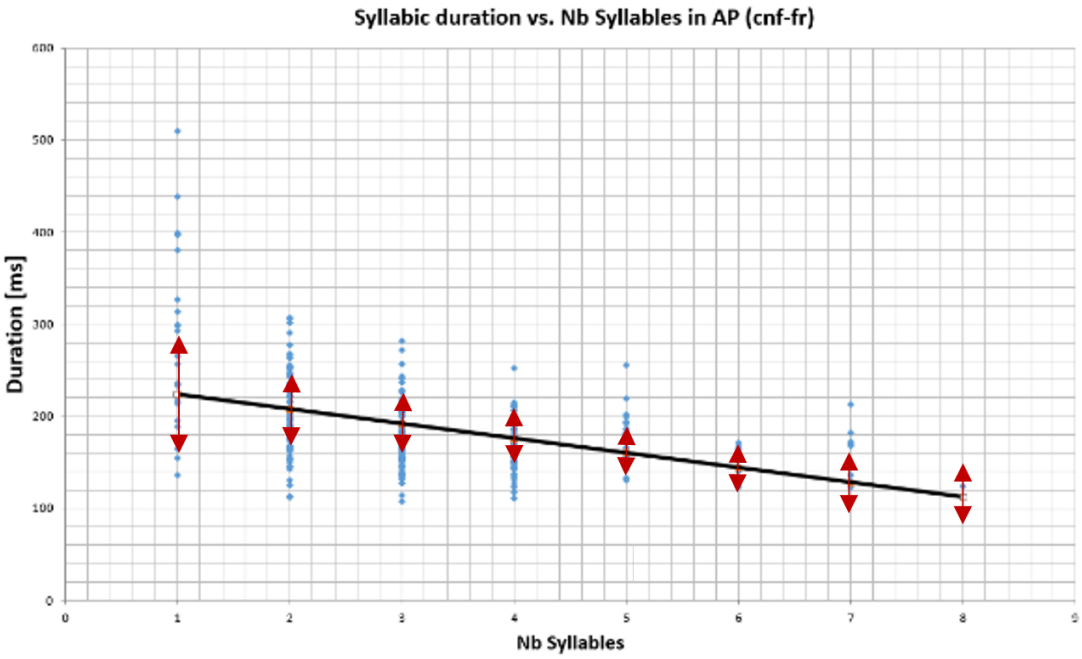


Figure 6. Syllabic duration in function of the number of syllables in AP regression line (*nar-fr*). Standard deviation is indicated by double arrows.

These regression lines correspond to the results given in [12] and [1], the longest the AP duration, the shortest are the syllables they contain.

5. Interpretation of results

1.1 Maximum and minimum values of AP's duration

Table 1 and the following figures suggest that Delta brain waves synchronization hypothesis is validated, as the minimum and maximum values do correspond convincingly to the maximal and minimal values of AP duration values.

1.2 Eurhythmy

Fig. 6 shows that eurhythmy is obtained by compression of the syllabic duration for long AP's.

1.3 AP's grammatical words

Table 2. Number of AP's containing only closed class words (Conjunctions, determinants, etc.) over the total number of AP's in each recording.

	Lec	Cnf	Nar	Pol	Jpa
Cases	**4/175**	**18/223**	**2/290**	**0/100**	**11/300**
	2%	**8%**	**0.6%**	**0%**	**3.6%**

This table shows clearly that the hypothesis pertaining to the AP content is invalidated. Not only can an AP contain more than one open class word in French, but spontaneous speech data include a relatively large number of occurrences of AP's with only grammatical words.

6. Conclusion

The analyzed data on various styles of spontaneous speech data validate the proposed explanation for prosodic structure constrains, namely:

a) The maximum number of syllables in a given AP is indeed of the order of 7 to 8, but the actual limit is given by the largest possible Delta wave period, about 1200 ms. Examples of AP containing up to 12 syllables in are found for example in [7], but even in this fast speech rate case their duration is below the Delta brain wave limit of 1200 ms;

b) Successive stressed syllables are found ("stress clash", corresponding in French to one single syllable AP following the first AP) but there is a minimal amount of time between two consecutive stressed syllables (actually between two consecutive stressed vowels). This observation confirms the hypothesis about Delta brain waves synchronizing the perception of AP, in the case at a maximum frequency, i.e. a minimal period of about 250 ms.

c) Cases where eurhythmy is obtained at the expense of congruence of the prosodic structure with syntax are rare so the eurhythmic compensation is done by compressing the syllabic duration in AP with many vowels. This was already observed empirically in [4], [8], [15] and more recently in [1]. One of the reason why balancing of the number of syllables is not frequent in spontaneous data may pertain to the fact that such balancing requires preplanning essentially possible for read speech (cf. the

read phrasing [*Marie adore*] [*les chocolats*] vs. the spontaneous [*Marie*] [*adore les chocolats*]). It seems that speakers realize eurhythmic phrasing when the syntactic constraint is weak or absent, i.e. for enumeration, short read sentences, etc.

d) No cases of syntactic clash were observed;

e) However, occurrences of AP containing no lexical words and only grammatical word are not infrequent.

The next step in this research would concern other Romance languages with lexical stress, and later tone languages such as Mandarin with no lexical stress.

Romance languages other than French may show the coexistence of a lexical stress and a tone boundary sometimes combines on the same AP final syllable (in Italian for example). These two prosodic events may play the same role (or complement each other) in the storage concatenation process proposed by [11].

7. References

[1] Avanzi, Mathieu, Lucie Rousier-Vercruyssen et al. (2013) C-PROM-Task. A New Annotated Dataset for the Study of French Speech Prosody, Proceedings TRASP 2013, Aix-en-Provence, 27-30.

[2] Beckman, Mary E. & Janet B. Pierrehumbert (1986) Intonational structure in Japanese and English, Phonology Yearbook 3, 255-309.

[3] Dell, François (1984) L'accentuation dans les phrases en français, in Dell F., Hirst D. & Vergnaud J.R. (éds), *Formes sonores du langage*, Hermann, Paris, 65-122.

[4] Fónagy, Ivan & Magdics, Klara (1960) Speed of utterances in phrases of different lengths, Language and Speech, 3, 179-192.

[5] Friederici, Angela & Wartenburger, Isabell (2010) Language and brain, *Cognitive Science*, (10) 150-159.

[6] Ghitza1, Oded, Giraud, Anne-Lise and Poeppel, David (2013) Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence, Frontiers in Human Neuroscience, www.frontiersin.org, January2013, Volume 6, Article340.

[7] Lehka Irina. & David Le Gac (2004) Etude d'un marqueur prosodique de l'accent de banlieue, *Actes des XXIIIème Journées d'Etudes sur la Parole*, avril 2004, Fèz, Maroc

[8] Malécot, André, Johnston, R., & Kizzlar, P. A. (1972) Syllabic rate and utterance length in French. Phonetica, 26, 235-251.

[9] Martin, Philippe (1986) Structures prosodiques et structures rythmiques, Actes des 13ème JEP, Aix-en-Provence, l986.

[10] Martin, Philippe (2005) La transcription des proéminences accentuelles : mission impossible ? *Revue PFC*, septembre 2005.

[11] Martin, Philippe (2009) *Intonation du français*, Armand Colin, Paris, 256 p.

[12] Martin, Philippe (2013) Contraintes phonologiques de l'intonation de la phrase réinterprétées à la lumière des recherches récentes en neurophysiologie, *La Linguistique*, 2013/1.

[13] Meigret, Louis (1550) *Le treté de grammere francoeze*, Réimpression chez Slatkine, Genève, 1972.

[14] Pasdeloup, Valérie (2004) Le rythme n'est pas élastique : étude préliminaire de l'influence du débit de parole sur la structuration temporelle, Actes des JEP 2004, Fès (Maroc) - 19-22 avril 2004.

[15] Wioland, François (1984) Organisation temporelle des structures rythmiques du français parlé, *Bulletin de Linguistique de Lausanne*, 6, 293-322.

[16] C-PROM (2010) Corpus libre de parole multigenre, https://sites.google.com/site/corpusprom/

[17] WinPitch, www.winpitch.com

Hearing the Structure of Math: Use and Limits of Prosodic Disambiguation for Mathematical Stimuli

Michael Phelan

Department of Linguistics, The Ohio State University, USA
phelan@ling.osu.edu

Abstract

Listeners use the prosodic cues of an utterance to help determine its syntactic structure, but how does this process happen in the specialized domain of mathematics? Mathematical expressions can contain deeply embedded structures, and listeners encounter read mathematical expressions (RMEs) far less frequently than other potentially ambiguous utterances. How does experience with listening to math affect our ability to hear the structure of an RME via its prosody? Are there limits to the amount of structure we can pull out of the prosody of an utterance?

A perception experiment was conducted with subjects aged 7-59 to help answer these questions. Participants heard recordings of RMEs and attempted to determine which of two or more mathematical structures the reader intended. When subjects chose between two options for phrases like *nine times A minus two*, they chose the mathematical expression that had bracketing matching the prosody of the utterance. However, for more complex phrases like *the square root of sixteen over A plus twelve*, results were at chance. Age played a surprising role: subjects' performance increased dramatically from age 7 to 16, but adults' performance varied widely. This is attributed to variation in exposure to read mathematics.

Index Terms: speech perception, acquisition of prosody, prosody of mathematics

1. Introduction

If you have just heard someone say *two plus three squared*, should you be thinking of 11 or 25? The present study shows that the answer largely depends on the prosody of the utterance. Speakers use prosodic manipulations to communicate the intended grouping of words and phrases in their everyday English utterances, and listeners make use of these manipulations when determining which of several possible syntactic structures the speaker intended [1][2]. This paper extends these findings to the domain of read mathematical expressions (RMEs) and explores variations in the way listeners of different ages interpret certain prosodic phrasing. Finally, limitations on the use of prosodic phrasing for disambiguation are investigated with the use of complex RMEs that could correspond to more than two possible mathematical structures.

Read mathematical expressions provide and ideal test case for investigations of prosodic disambiguation for several reasons. First, the written form of a mathematical expression unambiguously shows the structural relationships between terms of the expression at a glance, while the utterance used to describe the mathematical expression can be ambiguous between two (1) or more (2) structures.

(1) Nine times A minus two: $(9 \cdot A) - 2$ $9 \cdot (A - 2)$

(2) The square root of sixteen over A plus twelve:

$$\frac{\sqrt{16}}{A+12} \qquad \frac{\sqrt{16}}{A}+12 \qquad \sqrt{\frac{16}{A}+12}$$

$$\sqrt{\frac{16}{A+12}} \qquad \sqrt{\frac{16}{A}+12}$$

Unlike everyday English stimuli, none of the mathematical structures corresponding to RMEs are inherently more or less plausible, and frequency effects are unlikely to bias listeners towards one selection or another.

Importantly for the purposes of this paper, readers in the same speech community have widely varying levels of experience with mathematical stimuli. Children who have long since attained adult-like levels of proficiency in using prosodic cues to disambiguate everyday English utterances may not yet have come across mathematical expressions like (2), and adults who do not work in education or math-heavy fields may not have dealt with such expressions in many years. On the other hand, older children and teenagers frequently encounter such expressions in the classroom, and thus may have an easier time disambiguating utterances of these expressions.

Finally, research to date on listeners' use of prosody for disambiguation has largely focused on whether a phrase should be interpreted with high or low attachment to previous parts of of the sentence. Comparatively little work has been done on prosodic disambiguation of utterances with many possible interpretations. The ease of generating such complex stimuli with mathematical utterances such as (2) can thus shed some light on the limitations of what listeners are able to disambiguate when forced to rely on prosodic cues alone.

2. Background and Assumptions

2.1. Previous research

Study of the prosodic disambiguation of syntactic ambiguities in non-mathematical speech goes back many years. Speakers have been shown to use lengthening and pausing to mark the intended structures of sentences with prepositional phrase and relative clause attachment ambiguities, complex noun phrase coordination ambiguities, and many others [3][4]. Listeners have likewise been shown to be sensitive to these

manipulations, at least when the interpretations differed in their syntactic bracketing [1][2].

A few previous studies investigated prosodic disambiguation of mathematical stimuli. Streeter [5] and Wagner [6][7] used simple mathematical stimuli like (3) and found that both speakers and listeners use the same prosodic disambiguation strategies as with non-mathematical stimuli.

(3) a. $A + (E \cdot O)$
b. $(A + E) \cdot O$

Only a few authors have used more complex mathematical stimuli, and all did so only in production. [8] and [9] both found longer pauses at locations likely to be marked with parentheses, though the only prosodic events coded by [8] were pauses over 300ms in length and [9] used a single speaker who consciously tried to prosodically disambiguate very complex stimuli.

In a production experiment with both simple (1) and complex (2) stimuli, [10] found that speakers produced RMEs using four distinct prosodic patterns. When intending structures like (3a), speakers most often used a strong prosodic break early in the utterance to group later terms together. When intending structures like (3b), speakers used a late prosodic break to group early terms together. Because the prosodic structure of these two sorts of utterances matched their syntactic structure, [10] referred to these utterances as having "cooperating prosody". In a few cases, speakers did the opposite, producing RMEs where the placement of breaks in prosodic structure grouped terms contrary to their grouping in syntactic structure, which [10] referred to as "conflicting prosody". Finally, many speakers produced evenly sized prosodic breaks throughout the utterance. Though the prosodic structure in these utterances did not necessarily group terms together in any particular way, different prosodically flat structures were used in different ways by speakers. When flat prosodic structures were used to indicate structures like (3b), the prosodic breaks were significantly more likely to be large, either all intonational phrase (IP) breaks or all intermediate (ip) phrase breaks. When flat prosody was used with structures like (3a), the breaks tended to be small (∅, or prosodic word-level breaks). [10] referred to these prosodic patterns as "big flat" and "little flat". The current study uses stimuli from [10] to determine whether these four distinct prosodic patterns are interpreted by listeners the same way they are apparently intended by speakers, and whether age and the complexity of the mathematical expression affects listeners' ability to determine the structure the speaker intended.

2.2. Assumptions about prosodic processing

Accounts of the processing of prosodic breaks fall into two main camps: those assuming the Strict Layering Hypothesis of [11], and those like [12] that allow for recursion in the prosodic description of an utterance. Data from experiments like the one presented here can bear on this question, allowing investigators to test the limitations on what sorts of utterances can be reliably disambiguated via prosody. Both sorts of theories would agree that utterances of phrases like (1) should be easily disambiguated by listeners, while more complex phrases like (2) could quickly reach the limit of the number of psychologically distinct levels of prosodic phrases in theories that assume the Strict Layer Hypothesis. The number of prosodic layers available to the speaker should set a hard limit on the degree to which multiply-embedded mathematical phrases could be reliably disambiguated. Under theories like [12] that are not subject to this restriction, speakers and listeners attempting to disambiguate utterances of phrases like (2) could make use of continuously varying relative differences in boundary strength. Drop-offs in perception accuracy for these more complex utterances would need to be explained by limitations on working memory capacity, and thus should be more gradual.

3. Experiment

To answer the questions raised above, a perception experiment was conducted with 29 subjects recruited from a local science museum. Subjects ranged in age from 7 to 59 (mean 25.2, sd 14.9) and participated in the experiment in groups of two to seven. None reported hearing problems, and all subjects whose data are presented here were native English speakers.

3.1. Stimuli

Three sets of mathematical stimuli were used. Two sets were selected from the recordings made for [10], in which naïve college students were asked to read twelve pairs of expressions like (4) and (5) with similar structures but different numbers.

(4) a. $9 \cdot (A - 2)$
b. $(10 \cdot A) - 2$

(5) a. $\sqrt{\frac{16}{A+12}}$
b. $\sqrt{\frac{81}{A}+72}$

Speakers in [10] were asked to avoid using terms like *the quantity, parentheses,* and *all of that,* and were told not to rearrange the terms of the expression, but were not told of the possible ambiguity in their speech. The two members of each pair (4a, b; 5a, b) were separated by at least four expressions with different structures, to prevent intentionally contrastive readings.

The first set of stimuli were chosen from pairs where one speaker produced cooperating prosody on both RMEs in a pair. The second set were chosen from pairs where a speaker produced cooperating prosody on one member of the pair and conflicting prosody on the other. A third set of stimuli were used in which one member of each pair was produced with IP breaks after each number or variable (called big flat, as in [10]) and the other member had only prosodic word level breaks throughout (little flat). Since such pairs were rarely produced by the same speaker in [10], these stimuli were recorded for this study by the experimenter.

There were 66 total mathematical items, 44 of which allowed only two structures as in (1), and 22 of which allowed either three or five possible structures, as in (2). All three sets

of stimuli were intermixed, such that the same expression was never repeated without at least five others expressions in between. The 44 "easy" items appeared first, followed by the 22 "hard" items.

A fourth set of stimuli consisted of eight non-mathematical English sentences like (6), which contained ambiguities due to prepositional phrase attachment, relative clause attachment, or complex NP conjunction.

(6) Leslie photographed the manager with the iPad.

English stimuli were recorded by prosodically trained speakers unaffiliated with the study, who were told to disambiguate by putting an IP break either after the verb (for low attachment) or before the preposition (for high attachment) in PP attachment ambiguities and similarly for other ambiguity types. These eight non-mathematical English stimuli were given at the end to assess subjects' understanding of the general pattern of prosodic disambiguation in English.

3.2. Procedure

Subjects were given paper packets showing the possible mathematical structures that corresponded to each utterance. They were told they would hear people reading math problems and were to circle which problem they thought the subject meant, or, for the non-mathematical stimuli, which of two paraphrases they thought the subject intended. Stimuli were played back to groups of subjects via a laptop situated in front of the group, at a volume comfortable for all participants. There was a brief pause after each item to allow subjects to make their selections: six seconds following "easy" items, twelve seconds for "hard" items, and ten seconds for non-mathematical items.

4. Results

4.1. Mathematical results – Easy trials

Over all trials in which the subjects had a choice between exactly two mathematical structures, if the prosodic structure grouped two terms together, subjects were significantly more likely to pick the mathematical structure that grouped those terms together. Trials with cooperating prosody, where the grouping of terms in prosodic structure matched the grouping in syntactic structure, saw subjects selecting the intended mathematical expression on 67% of trials, significantly higher than chance (binomial test, $p < 0.001$). On trials in which the utterance was produced with conflicting prosody, where the prosody did not match the syntactic grouping the speaker intended to convey, subjects followed the prosody, choosing the mathematical expression that matched the prosody on 59% of trials (binomial, $p < 0.05$).

For the easy subset of problems with flat prosodic structures, listeners were significantly more likely to interpret "big flat" utterances containing IP breaks between each term as indicating left-branching syntactic structures (as in (3b)) than right-branching syntactic structures, doing so on over 80% of trials (binomial, $p < 0.001$). "Little flat" structures, however, were not consistently seen as marking one type of mathematical structure. Subjects selected left-branching and right-branching mathematical structures at essentially the same rate (binomial, $p > 0.3$) on these trials.

4.2. Mathematical results – Hard trials

On trials where subjects were asked to pick between more than two answer choices, there was an significant difference in performance between problems that allowed for multiply-embedded constituents and those that did not. When subjects did not have to consider multiply-embedded structures, accuracy resembled that of easy problems, with choices following prosodic structure on over 60% of trials. However, when the utterance was consistent with multiply-embedded structures, as in (2) and (5), subjects were no better than chance at selecting the expression the reader had intended to convey (binomial, $p > 0.4$).

4.3. Non-mathematical trials

Overall accuracy on the eight non-mathematical items was reasonably high at 79%, significantly better than chance (binomial, $p < 0.001$). However 10 of the 28 subjects did score at or slightly below chance on these eight items, showing that understanding of the general English prosody-syntax correspondence rules is not universal. There was no significant correlation between age and accuracy on the non-mathematical items ($t = 1.717$, $p > 0.09$), though the two youngest subjects were the least successful on these questions.

In general, subjects who were most successful on the non-mathematical trials were also more successful on the mathematical trials ($t = 3.470$, $r = 0.56$, $p < 0.005$).

4.4. Interaction with age

There was an overall trend for accuracy in selecting the intended mathematical expression to improve with age ($t = 2.073$, $r = .37$, $p < 0.05$), though figure 1 shows that there is more going on than a simple trend to get better with age. Subjects improve dramatically from age 7 to 16, while adults are scattered all over the range. On non-mathematical trials adults did significantly better than children (64.4% vs 57.5%, $t = 2.234$, $p < 0.05$).

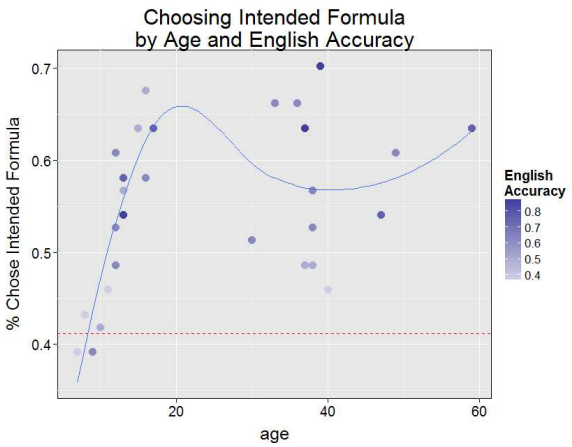


Figure 1. Effects of age and non-mathematical English trials accuracy in selecting intended form. Line of best fit is decidedly non-linear. Dashed line shows chance performance.

5. Discussion

In a listening experiment where they were asked to choose which of two or more possible mathematical expressions a speaker had intended to convey, subjects assumed that a larger prosodic break early in an utterance was meant to group later terms together, whereas a larger prosodic break later in an utterance was intended to group earlier terms together, as has been shown to be the case when listeners use prosody to disambiguate non-mathematical utterances.

Listeners correctly chose the intended mathematical form when there were only two options, or when there were more than two options but answers did not involve multiply-embedded phrasings. For trials where multiply-embedded structures were possible, listeners performed at chance. Under theories that follow the Strict Layer Hypothesis, this poor performance on complex problems could be attributed to listeners only have a small fixed number of psychologically distinct levels of prosodic phrasing. When the number of levels of embedding exceeds the number of prosodic phrasing levels available, disambiguation becomes difficult or impossible. Theories that reject the Strict Layer Hypothesis and allow recursion in prosodic phrasing would have to assume that this drop off in performance was due to higher memory load requirements for these complex mathematical expressions. The role of working memory in prosodic disambiguation could be tested by adding distractor tasks that increase memory load independent of the prosodic disambiguation task.

The striking effects of age evident in figure 1 suggests that familiarity with listening to or working with algebraic expressions like those used in the experiment has a strong impact on a listener's ability to use prosodic cues to disambiguate read mathematical expressions. The youngest subjects, just seven and eight years old, performed at or below chance. This was not surprising, given that these subjects were unlikely to have ever encountered expressions as complex as (4a,b), let alone (5a,b). [13] found that five to seven year olds could use prosodic breaks to choose between two different groupings in phrases like *pink and green and white,* but only when the pauses were quite pronounced (680ms), much longer than almost all pauses in stimuli here. Likewise unsurprising was the increase in performance for older children, who have had more experience with algebra and with prosodic disambiguation in general. What jumps out is the widely varying ability of adults in figure 1. While the results of the non-mathematical English trials clearly show that adults are able to use prosodic cues to disambiguate utterances, the presence of mathematical expressions seems to have greatly impacted adults' ability to apply the rules of prosody-syntax correspondence. Syntactically, the utterances used in the non-mathematical English trials are very similar or identical to those used in the mathematical trials, so differences between these two sets of results must be due to the fact that the stimuli were mathematical. It is unlikely that adults did not understand the structure of the mathematical expressions they were choosing between, since spacing, parentheses and other conventions of written mathematics indicate the grouping of terms quite clearly. It may be the case that adults, who generally do not have much exposure to algebraic expressions after graduating high school, are out of practice in applying known rules of prosody-syntax correspondence to the mathematical domain. There are few if any previous studies showing this sensitivity to a particular domain for prosodic disambiguation, and follow-up studies are underway with adult math teachers and college students majoring in math and engineering, who have much more exposure to mathematical expressions in their daily life.

6. References

[1] Lehiste, I., Olive, J., & Streeter, L. Role of duration in disambiguating ambiguous sentences. Journal of the Acoustical Society of America. 60, 1199-1202. 1976.

[2] Snedeker, J., & Casserly, E. Is it all relative? Effects of prosodic boundaries on the comprehension and production of attachment ambiguities. Language and Cognitive Processes, Vol. 25, 1234-1264. 2010.

[3] Lehiste, I. Phonetic disambiguation of syntactic ambiguity. *Glossa* Vol. 7, 107-122. 1973.

[4] Cooper & Paccia-Cooper. *Syntax and Speech*. Harvard University Press.1980.

[5] Streeter, L. "Acoustic determinants of phrase boundary perception." *Journal of the Acoustical Society of America*, Vol. 64, 1582–1592. 1978.

[6] Wagner, M. "Prosodic options or syntactic/semantic choices?"presented at the Workshop on Prosody and Meaning. Barcelona, Spain, 2009.

[7] Wagner, M., & Crivellaro, S. "Relative Prosodic Boundary Strength and Prior Bias in Disambiguation." Speech Prosody, University of Chicago. 2010.

[8] O'Malley, M., Kloker, D., & Dara-Abrams, B. "Recovering parentheses from spoken algebraic expressions." *IEEE Transactions on Audio and Electroacoustics,* Vol. 21 No. 3, 217-220, 1973.

[9] Holm, B., Bailly, G., & Laborde, C. Performance structures of mathematical formulae. in International Congress of Phonetic Sciences. San Francisco, USA. 1999.

[10] Phelan, M. The Prosody of Algebra and the Algebra of Prosody: Prosodic Disambiguation of Read Mathematical Formulae. Proceedings of Speech Prosody 2012, Shanghai. 2012.

[11] Selkirk, E. O. *Phonology and Syntax: The relation between sound and structure.* Cambridge,MA: MIT Press. 1984.

[12] Wagner, M. Prosody and Recursion. Ph.D. Dissertation, Massachusetts Institute of Techonology. 2005.

[13] Beach,C., Katz, W., & Skowronski, A. Children's processing of prosodic cues for phrasal interpretation. *Journal of the Acoustical Society of America,* 99, 1148-1160. 1996.

Robust Pitch Estimation using Ensemble Empirical Mode Decomposition

Sujan Kumar Roy[1,2], *Md. Khademul Islam Molla*[2] *and Keikichi Hirose*[3]

[1] University of Concordia, ON, Canada, [2]University of Rajshahi, Rajshahi, Bangladesh
[3]Graduate School of Information Science and Technology, The University of Tokyo, Japan

khademul.cse@ru.ac.bd, hirose@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents an efficient pitch estimation algorithm for noisy speech signal using ensemble empirical mode decomposition (EEMD) based time domain filtering. The dominant harmonic of noisy speech is enhanced to make pitch period more prominent. The normalized autocorrelation function (NACF) of the modified signal is then decomposed into time varying subband signals using EEMD. In contrast to the ordinary EMD, it does not introduce any mode mixing during decomposition. The subbands containing pitch component are selected and separated yielding partially reconstructed signal. The pitch period is determined from thus separated signals. The experimental results show that the proposed algorithm performs better compared to other recently reported algorithms in noisy environment.

Index Terms: ensemble empirical mode decomposition, filtering, pitch estimation

1. Introduction

Pitch information is an important prosodic feature which is used in many speech processing applications including speech enhancement, automatic speech recognition, analysis and modeling of speech prosody, low-bit-rate speech coding [1]. Although there are many pitch estimation algorithms (PEAs), the development of an efficient PEA is still demanding.

Some PEAs implemented in time domain contained poor accuracy of pitch estimation [2]-[5]. Autocorrelation Function (ACF) based algorithm is introduced in [2]-[3]. The performance of ACF method is basically depended on pitch peak in the autocorrelation domain which may be quite difficult to perform against noise, quasi-periodic nature of the speech signals [6]-[7]. Normalized autocorrelation function (NACF) based algorithm has been presented in [2] which provides better results than ACF but suffer from robustness. A weighted autocorrelation (WAC) based method has been presented in [3]. It is easy to enhance pitch peak by dividing ACF with the AMDF in a lower portion than succeeding peaks but this leads the algorithm to cause a serious problem called double pitch error. Signal reshaping technique with the improvement of specific harmonic is presented in [6]-[7]. The dominant harmonic (DH) of the noisy speech signals is determined by using a DFT based method and boosted the amplitude of DH in the analyzing signal which is called dominant harmonic modification (DHM). This technique also fails to perform under noisy environment. The data adaptive techniques of pitch estimation are introduced in [8]-[11] using EMD [12, 13]. The ordinary EMD suffers from the 'mode mixing' effect decreases pitch estimation performance. To overcome the mode mixing problem, Wu and Huang [14] proposed a modification to the EMD algorithm what is termed as EEMD.

In this paper, a novel pitch estimation algorithm is proposed using the combination of DHM and EEMD. DHM plays an important role to overcome the double pitch error while EEMD acts as data adaptive time domain filtering to separate the signal containing only the pitch information which minimizes the pitch estimation error.

2. Pitch Estimation Algorithm

The noticeable improvement of the EMD based PEA is possible by overcoming the mode mixing problem. At first, conventional low-pas filtering is performed on the noisy speech signal. If $s(n)$ and $v(n)$ denote the speech and additive noise signals respectively, the observed speech signal $x(n)$ can be represented as: $x(n) = s(n) + v(n)$.

The noise effect can be reduced significantly by pre-filtering the observed signal $x(n)$ in the Fourier domain. As the pitch range of speech signal is well known to be 50-500Hz, a significant portion of the high frequency components is filtered out in frequency domain. The resultant signal is termed as pre-filtered speech (PFS) and represented as $\psi(n)$ which contains less noise. In the second phase of pre-processing, the DHM is applied to the PFS signal which is described in the following subsection.

2.1. Dominant harmonic enhancement

The dominant harmonic (DH) can be estimated as the one which has the largest amplitude in the Fourier domain [6].

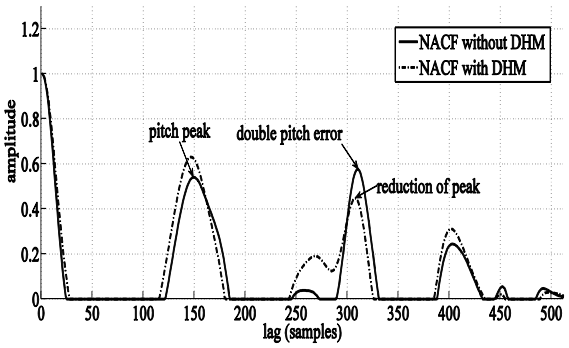


Figure 1: *Pitch peak enhancement using DHM for a voiced frame of a female speaker with 0dB SNR. NACF causes double pitch error but DHM resolves that problem. The true pitch is 151 in samples.*

Let $y_{dh}(n)$ denotes the DH present in the PFS signal $\psi(n)$. The signal after dominant harmonic enhancement denoted by $y(n)$ is then given by:

$$y(n) = \psi(n) + \bar{y}_{dh}(n) \quad (1)$$

where,

$$\bar{y}_{dh}(n) = \rho[y_{dh}(n) - |y_{dh}(n)|] \quad (2)$$

and

$$y_{dh}(n) = \theta_{dh}\cos(\omega_{dh}n + \delta_{dh}) \quad (3)$$

where, θ_{dh}, ω_{dh}, δ_{dh} are amplitude, frequency and phase of the dominant harmonic respectively, ρ is an arbitrary constant, and |.| denotes the absolute value. The parameter ρ controls the mixing ratio to be chosen appropriately. Figure 1 demonstrates the effects of the DHM which shows that the conventional NACF of $\psi(n)$ causes double pitch error while the use of DHM to $\psi(n)$ overcomes such error.

2.2. EEMD in pitch detection

EMD suffers from mode mixing effect which indicates that the oscillations of different time scales coexist in a given IMF, or that oscillations with the same time scale have been assigned to different IMFs [14]. Due to this problem, EMD does not guarantee the accurate frequency scale separation and pitch information can be mixed between two successive IMFs which degrades the pitch detection accuracy. EEMD utilizes the scale separation principle of the EMD. The principle of the EEMD is to add white noise into the signal with many trials. The noise in each trial is different, and the added noise can be canceled out on average, if the number of trials is sufficient. Let $\phi(n)$ is the NACF of $y(n)$. For signal $\phi(n)$, original EMD is defined as [12]:

$$\phi(n) = \sum_{k=1}^{N} c_k(n) \quad (4)$$

where $c_k(n)$ is the k^{th} IMF. EEMD decomposes $\phi(n)$ by first construction an ensemble of signal samples $\phi_m(n)$ by adding to $\phi(n)$, M independent copies of finite amplitude white-noise $\eta_m(n)$, i.e.,

$$\phi_m(n) = \phi(n) + \eta_m(n), \ (m=1, 2, \ldots, M), \quad (5)$$

Applying EMD to $\phi_m(n)$ and repeat until the trial number with different added white noise series of the same power at each time, the new IMF combination $c_k^{(m)}$ is obtained, where k is the iteration number and m is the IMF scale.

$$\phi_m(n) = \sum_{k=1}^{N} c_k^{(m)}(n), \qquad (m=1, 2, \ldots, M), \quad (6)$$

where, N is the trial number.

The ensemble means is calculated as:

$$\tilde{c}_k(n) = \frac{1}{M}\sum_{m=1}^{M} c_k^{(m)}(n), \qquad (k=1, 2, 3\ldots), \quad (7)$$

A large number (M) of samples and white noise of finite amplitude are required to force the ensemble to exhaust all possibilities in the sifting process. In this way, possible mode mixing is effectively removed, and components of different time scales embodied in the original signal are well collated in proper IMFs, whose frequency bands essentially approximate those dictated by the dyadic filter banks [14]. The effect of the added white noise should decrease following the well-established statistical rule:

$$\beta_n = \frac{\beta}{\sqrt{M}} \quad (8)$$

or

$$\ln \varepsilon_n + \frac{\varepsilon}{2}\ln M = 0 \quad (8)$$

where, M is the number of ensemble members, β is the amplitude of the added noise and β_n is the final standard deviation of error, which is defined as the difference between the input signal and the corresponding IMF(s) [14].

Mode mixing problem is illustrated in the left column of Figure 2, where a clean speech segment of 100 ms length is decomposed by EMD. The three IMFs with higher energy are shown. The appearance of oscillations of dramatically disparate scales in IMF_3 is clear. Another example can be seen in IMF_4, where two oscillations are marked with circles. These oscillations are very similar to those on IMF 5. On the right side of Figure3, in IMF_4 and IMF_5 there is no mode mixing, where $M=50$ and $\beta=0.1$ were used for generating the resulting IMFs decomposed by EEMD.

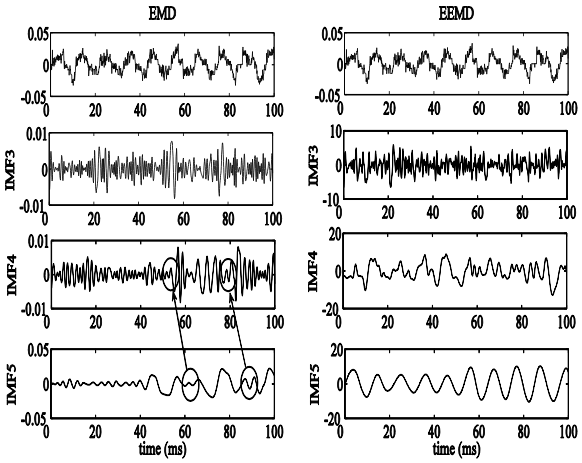


Figure 2: *A clean speech segment analyzed by EMD (left col-umn) and EEMD (right column). The corresponding IMFs 3 to 5 are shown. In IMF4 and IMF5of EMD where "mode mixing" occurs are marked with circles.*

The pitch detection performance using DHM method often reduced especially for low SNR speech signals [6]-[7]. To mitigate this shortcoming effects, EEMD based data adaptive time domain filtering is applied to the NACF of the DH modified signal $y(n)$ (equation-2) and a partial reconstruction is made from the EEMD domain to calculate the actual pitch period. The reconstructed signal $\lambda(n)$ is

obtained as $\lambda(n) = \sum \tilde{c}_j(n)$, where $\tilde{c}_j(n)$ are the IMFs and its fundamental periods are within the pitch range 50-500Hz. Let *ξ(n)* is the NACF of *y(n)*. The NACF *ξ(n)* and its corresponding partially reconstructed signal *λ(n)* are shown in Figure 3. It is observed that the pitch peak is more prominent in *λ(n)* and overcomes double pitch error.

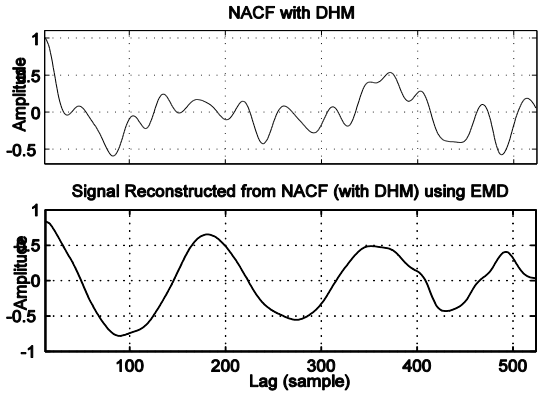


Figure 3: *NACF of the signal after performing the DH enhancement (above), the partially reconstructed signal from NACF with DHM using EMD (bellow). The double pitch error is eliminated (bellow).*

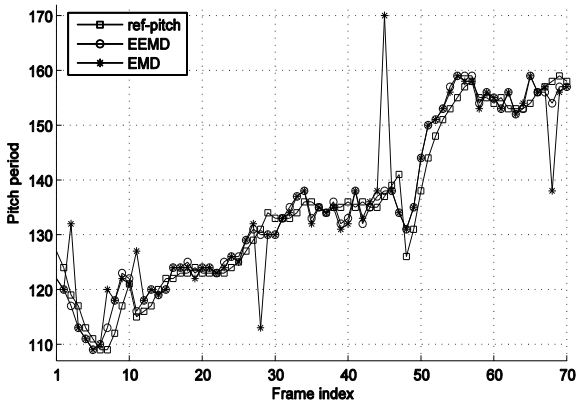


Figure 4: *Performance Comparison of pitch period estimated by EEMD (proposed) and EMD method with the reference pitch for female speech. The proposed EEMD based method performs better than EMD.*

2.3. Proposed pitch detection algorithm

The proposed algorithm for pitch estimation can be summarized as:

- Apply pre-filtering to the noisy speech signal to remove a significant portion beyond the pitch range 50-500Hz.
- Perform signal modification using DHM on PFS and then apply NACF to the modified signal.
- Apply EEMD to the obtained NACF.
- Sum up the IMFs with fundamental period lies within the specified pitch range for partial.
- Take the right half of the reconstructed signal.
- The amplitude at zero-lag is selected as the starting index of the pitch period.
- Find the next highest peak from the right half of the reconstructed signal.
- Calculate the pitch period from the difference between the starting index and the next highest peak index.

The detected pitch response based on the proposed algorithm has shown in Figure 4.

3. Experimental Results and Discussion

The performance of the proposed method is tested using the Keele pitch extraction reference database obtained from ftp://ftp.cs.keele.ac.uk/pub/pitch/ which contains the sampling frequency of 20 kHz with 16-bit resolution. Note that the value of the parameter ρ in Eq. (3) was set to 0.5. Both male (M2~M3) and female (F2~F3) mature speakers' speech are used here. There are 2650 'clearly voiced' male frames and 3227 'clearly voiced' female frames that is a total of 5877 analysis frames are used for experiment. White Gaussian noise and babble noise are used to corrupt the test signal to investigate the robustness. In the experiment, each 25.6 msec analysis frame is weighted by a 512-point rectangular and 10 msec is used as frame shift to generate the reference pitch values given in the database. If the estimated pitch for a frame deviates from the reference by >20%, we recognize the error as a gross pitch error (GPE).

The performance of the proposed EEMD based algorithm is compared with recent four PEAs - EMD based method [11], DHM based method [6], and conventional NACF method [2] and WAC [3] in presence of white noise as shown in Table I. It is observed that the proposed EEMD based algorithm performs better and closer to the original reference pitch (as shown in Figure 4) than all the mentioned PEAs for a wide range of SNRs(-5dB to 30dB). Figure 5 shows the average %GPE for male and female speakers respectively from which we see that the proposed PEA improves pitch estimation accuracy in noisy environment especially with low SNRs.

Table 1. *Performance comparison of EEMD algorithm with recently reported PEAs for male and female data. White noise is used to corrupt the speech signal.*

SNR(dB)		**-5**	**0**	**10**	**20**	**30**
M2	EEMD	6.72	5.51	2.02	1.54	1.53
	EMD	13.78	7.69	4.13	2.92	2.67
	DHM	14.26	7.45	2.99	2.26	1.94
	NACF	22.36	14.74	6.96	5.26	4.94
	WAC	25.20	15.23	7.37	6.17	6.07
M3	EEMD	3.12	0.99	0.21	0.14	0.07
	EMD	7.77	2.40	0.35	0.14	0.14
	DHM	12.28	4.73	1.12	0.56	0.35
	NACF	23.37	11.51	3.38	1.69	1.27
	WAC	24.29	12.07	3.38	1.20	0.98
F2	EEMD	5.51	1.98	0.93	0.88	0.55
	EMD	10.98	5.85	1.32	1.05	0.72
	DHM	10.97	6.39	1.93	1.15	0.93
	NACF	21.48	11.91	4.24	2.04	1.70
	WAC	23.05	11.58	3.97	1.93	1.59
F3	EEMD	5.51	2.97	0.98	0.28	0.21
	EMD	11.67	5.37	1.49	0.50	0.49
	DHM	11.38	6.01	2.05	0.70	0.49
	NACF	21.49	12.16	4.73	2.19	1.69
	WAC	21.78	7.85	4.38	1.64	1.62

In the second phase of experiment, we evaluate and compare the pitch detection performances of the proposed PEA with EMD [11, 15] and DHM [6] in noisy environment with bubble noise as illustrated in Table II. The proposed EEMD method provides better results than EMD and DHM based methods for the whole range of SNRs (-5dB to 30dB). The average performance comparison for this experiment has been shown in Figure 6 and it is clearly observed that the proposed PEA exhibits better performance than EMD and DHM based approach.

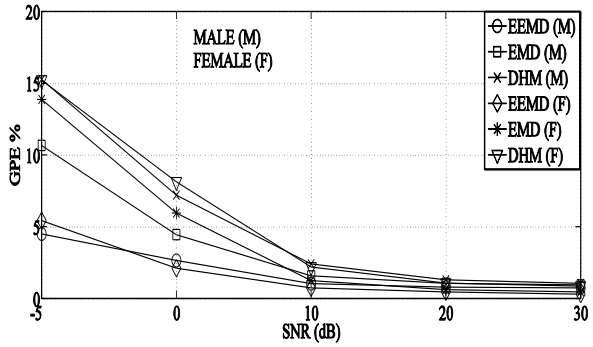


Figure 5: *Average performance results in terms of %GPE for white noise ('M' and 'F' denote male and female, respectively).*

Table 2. *Performance comparison of EEMD algorithm with recently reported PEAs for male and female data. Babble noise is used to corrupt the speech signal*

SNR(dB)		-5	0	10	20	30
M2	EEMD	20.65	10.93	5.26	2.26	1.54
	EMD	24.71	18.15	5.26	2.76	2.75
	DHM	52.83	32.98	9.64	3.81	3.08
M3	EEMD	21.66	14.74	2.04	0.56	0.07
	EMD	29.52	19.49	4.02	0.71	0.28
	DHM	57.98	38.42	8.55	1.06	0.35
F2	EEMD	13.18	5.62	1.48	1.10	0.56
	EMD	58.30	32.98	5.63	1.49	0.93
	DHM	73.08	47.21	9.82	2.31	0.94
F3	EEMD	8.69	5.23	1.06	0.35	0.28
	EMD	63.72	41.02	9.12	2.75	0.49
	DHM	67.05	38.76	6.44	1.98	0.78

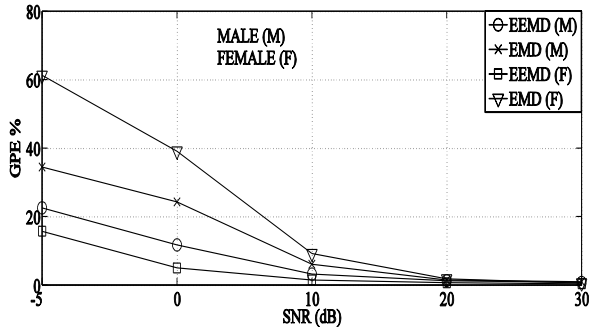


Figure 6: *Average performance results in terms of %GPE for babble noise ('M' and 'F' denote male and female, respectively).*

4. Conclusions

It is observed that the proposed method provides better performance in terms of % GPE than other recently reported PEAs in bubble and white noise environment for a wide range of SNRs(-5dB to 30dB). The EEMD based filtering efficiently extracts the signal components containing pitch information. Then pitch period is determined from partially reconstructed signal of EEMD domain. The pitch peaks become more prominent in the reconstructed signal. The EEMD is being free of mode mixing problem, pitch period does not overlap between any two IMFs and in most cases the detected pitch responses are almost closer to the original reference pitch. This strategy proves the superioty of the proposed algorithm. Also its performance in low SNRs is extremely better than EMD, DHM, NACF and WAC based PEAs. Finally, all sorts of experimental results prove the advantages of the proposed algorithm.

5. References

[1] W. Hess, Pitch Determination of Speech Signals: Algorithms and Devices, Springer, Berlin, 1983.

[2] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking”, Proc. IEEE ICASSP, pp.361-364, 2002.

[3] T. Shimamura and H. Kobayashi, “Weighted Autocorrelation for Pitch Extraction of Noisy Speech”, IEEE Trans. Speech and Audio Proc., 9(7):727-730, 2001.

[4] M. C. Dogan and J. M. Mendel, “Real-time robust pitch detector”, Proc. of IEEE ICASSP, 1, 129-132, 1992.

[5] N. Abu-Shikhan and M. Deriche, “A novel pitch estimation technique using the Teager energy”, Proc. of ISSPA, 1, 135-138, 1999.

[6] M. K. Hasan et. al., “Signal reshaping using dominant harmonic for pitch estimation of noisy speech”, Signal Processing, 86(5):1010-1018, 2005.

[7] M. K. Hasan, C. Shahnaz and S. A Fattah, “Determination of pitch of noisy speech using dominant harmonic frequency”, Proc. IEEE Int. Symposium on Circuits and Systems, 2, pp.556-559, 2003.

[8] H. Huang and J. Pan, “Speech pitch determination based on Hilbert-Huang transform”, Signal Processing, 86(4):792-803, 2005.

[9] Z. Yang, D. Huang and L. Yang, “A novel pitch period detection algorithm based on Hilbert-Huang transform”, LNCS 3338, pp. 586-593, Sinobiometrics, 2004.

[10] M. K. I. Molla, K. Hirose, N. Minematsu and M. K. Hasan, “Pitch Estimation of Noisy Speech Signals using Empirical Mode Decomposition”, Proc. of EUROSPEECH 2007.

[11] Sujan Kumar Roy, Md. Khademul Islam Molla, Keikichi Hirose and Md. Kamrul Hasan, “Harmonic modification and data adaptive filtering based approach to robust pitch estimation”, International Journal of Speech Technology, Volume 14, Number 4, 339-349, DOI: 10.1007/s10772-011-9112-6.

[12] N. E. Huang et. al., “The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis”, Proc. Roy. Soc. London A, Vol. 454, pp. 903-995, 1998.

[13] P. Flandrin, G. Rilling and P. Goncalves, “Empirical mode decomposition as a filter bank”, IEEE signal processing letters, Vol. 11, No. 2, pp.112-114, 2004.

[14] Z. Wu and N.E. Huang, “Ensemble Empirical Mode Decomposition: a noise-assisted data analysis method,”Ad-vances in Adaptive Data Analysis, vol. 1, pp. 1-41, 2009.

[15] S. K. Roy, M. K. Islam, K. Hirose and M. K. I. Molla, “Dominant harmonic modification with data adaptive filter based algorithm for robust pitch estimation”, Proc. of IEEE Int. Symposium on Circuits and Systems (ISCAS), 2011.

The Information Structure–Prosody Language Interface Revisited

Mónica Domínguez[1], *Mireia Farrús*[1], *Alicia Burga*[1], *Leo Wanner*[2,1]

[1]TALN Group, N-RAS Research Centre
Department of Information and Communication Technologies
Universitat Pompeu Fabra
[2]Catalan Institute for Research and Advanced Studies (ICREA)
{monica.dominguez|mireia.farrus|alicia.burga|leo.wanner}@upf.edu

Abstract

Several grammar theories relate information structure and prosody, highlighting a major correspondence between theme and rheme, and intonation patterns. Although these theories have been successfully exploited in some specific speech synthesis applications, they are mainly based on short default-order sentences, which limits their expressiveness for real discourse with longer sentences and complex structures. This paper revises these theories, identifying cases in which they are valid, and providing a new proposal for cases in which a more complex model is needed. Specifically, our experiments performed on real discourse from the Wall Street Journal corpus show that we need a model that: (1) foresees a hierarchical theme/rheme structure, and (2) introduces, apart from the traditional theme and rheme, a new element—the specifier.

Index Terms: information structure, thematicity, theme, rheme, prosody, ToBI.

1. Introduction

The influence of the information structure on intonation is widely reported in literature [1, 2, 3] under the heading of the "semantics-syntax-intonation language interface" [4, 5, 1, 6, 7]. Steedman [1] proposes a grammar theory that relates three different fields: syntax, semantics and intonation. Based on the theory stated by Beckman and Pierrehumbert [8] on intonation and information structure, Steedman establishes a main and recurrent correspondence between theme and rheme on the one side and intonation patterns on the other side. This correspondence has already been exploited experimentally in speech synthesis applications that serve as front ends in dialogue engines [9, 10]. However, it is not obvious that the experience gained in these applications can be transferred to, for instance, monologue generation: works such as [1] are based on rather short sentences with a simple structure and a default word order (SVO for English), which are not common for the genre of monologues. If we want to generate natural speech based on real discourse information, the hypotheses put forward in these works must be applicable to long sentences with complex syntactic structures as well. There are no descriptive studies on real data that provide evidence for their applicability or offer sound arguments that help revising these theories.

The aim of this paper is to test these theories at a small scale with a two-fold objective: (i) to validate them and, in case discrepancies between their proposed thematicity–intonation correlation and the observed correlation are identified, to determine when and why these discrepancies occur; and (ii) to propose a model that has the potential to capture better the thematicity–intonation correlation, especially in the case of complex linguistic constructions. Achieving this objective will also be instrumental for the establishment of a valid methodology for dealing with large corpora for the description of prosody at a deep structure level. To this end, this paper draws upon real discourse extracted from the Wall Street Journal and recorded in a professional setting.

The structure of this paper unfolds as follows. In Section 2, the theoretical background related to the annotation of both information structure and prosody is briefly explained. Section 3 presents the study in which we relate information structure and intonation patterns in the spirit of the "classical" theories, as well as the most outstanding findings in this respect. Section 4 suggests how at least some of the challenges encountered in the previous section can be met by drawing upon a more elaborated notion of information structure. The summary of our experiments and the conclusions we draw from them are sketched in Section 5.

2. Theoretical background

In what follows, we briefly introduce the notions of information structure and the information structure–prosody interface.

2.1. Information structure

The Information Structure (IS) (also known as Topic-Focus Articulation, TFA [11] in the Prague School [12], and Communicative Structure, CommStr, in the Meaning-Text Theory [6]) determines the "communicative" segmentation of the meaning of an utterance. This makes it central to the semantics–syntax–intonation interface [4, 5, 1, 6, 7] and therefore also to Natural Language Processing (NLP).

Steedman's work (which will be referred to in this paper as "the classical approach") is based on the interpretation of IS as a two-partite *thematicity* structure, with *theme* (that part of an utterance which connects it to the rest of the discourse) and *rheme* (what the utterance contributes to that theme) [1, p.655]. According to Steedman, it is also possible to have discontinuous themes and rhemes and all-rhematic sentences.[1] Consider an example of theme (Th)/rheme (Rh) distribution taken from [1]:[2]

(1) *Q: I know what Marcel SOLD to HARRY.*

[1]In fact, for Steedman, "the majority of themes in everyday utterances are null themes" [1, p.678], i.e., the majority of sentences are all–rhematic.

[2]For theme/rheme determination in a sentence, Steedman pictures, as is common in the field, the sentence in question as an answer to a hypothetical question.

But what did he GIVE to FRED?
A: (Marcel GAVE)$_{Th}$ (a BOOK)$_{Rh}$ (to FRED.)$_{Th}$

Mel'čuk [6] proposes a more complex thematicity structure. Thus, on the one hand, he distinguishes, apart from the traditional theme and rheme (whose definition, in general terms, coincides with Steedman's), a specifier element (SP), which sets up the context of the utterance; and, on the other hand, he defines thematicity over propositions rather than over sentences. The second feature implies that thematicity is *per se* a hierarchical structure: if a proposition is embedded, its thematicity partition will be embedded as well. In sentences containing coordinated propositions, there is a parallel thematicity structure (one partition by proposition)[3].

In further contrast to Steedman, Mel'čuk assumes that apart from existential and zero-argument propositions, which are all-rhematic, any proposition has at least theme and rheme. Consider an example of theme (T)/rheme (R)/specifier (SP) distribution in the sense of Mel'čuk:

(2) $\{$*[Years ago]$_{SP}$, [he]$_{T}$ [collaborated with the new music gurus Peter Serkin and Fred Sherry in the very countercultural chamber group Tashi,* $\{$*[which]$_{T(P2)}$ [won audiences over to dreaded contemporary scores like Messiaen's Quartet for the End of Time]$_{R(P2)}\}_{P2}]_{R}\}_{P1}$

2.2. The classic information structure–prosody interface

Although all three elements of prosody (intonation, rhythm and stress) are equally important from a theoretical point of view, intonation is the most relevant feature for speech synthesis applications, since its correct prediction helps to obtain naturalness and variability in the generated speech [13]. As has been argued by many authors, among them by Beckman and Pierrehumbert [8] and Steedman [1], intonation is also directly correlated with IS. Beckman and Pierrehumbert identified six pitch accents and classified them as theme-rheme markers (see Table 1).

Table 1: *Pitch markers of theme and rheme (stated by [8]).*

	patterns
theme	L+H*, L*+H
rheme	H*, L*, H*+L, H+L*

As far as pitch accents are concerned, Beckman and Pierrehumbert suggest that the characteristic bitonals for theme and rheme are L*+H and H+L* respectively. Steedman [1] builds upon this theory and hypothezises on complete pitch accent/boundary tone (PABT) patterns, claiming that:

> *the intonational phrase L+H* LH%* [a clearly increasing Low-High pattern] *(among others) is associated with the theme, whereas the H*L% and H*LL%* [clearly decreasing High-Low] *tunes (among others) are associated with the rheme*; cf. [15, p.275, 16, p.28, 17].

Accordingly, Steedman correlates the theme/rheme in example (1) with intonation patterns (IP) as follows:

(3) *Q: I know what Marcel SOLD to HARRY.*
But what did he GIVE to FRED?
A: (Marcel GAVE) (a BOOK) (to FRED.)
L+H LH% H* L L+H* LH%*

In order to test Steedman's theme/rheme-IP correlation hypotheses, we have grouped together ToBI patterns resulting from automatic labeling and a further reduction process into three cathegories according to their final intonation curve typology, i.e. 'falling', 'rising' and 'flat'. We assume that our automatic labeling [20], based upon one main pitch accent within each intonational phrase, may not always coincide with detailed manual labeling accounting for all intonation events occurring along the pitch contour line, such that a need for broader categories arises in order to establish a comparable ground between our automatic labeling and traditional theories. Table 2 summarizes the collection of patterns we used in this experiment.

Table 2: *Intonation patterns classified according to their final intonation curve type.*

	patterns
falling	H*L%, L*+HL%, H*+LL%
rising	L*H%, L*+HH%, H*+LH%, H+L*H%
flat	L*L%, H*H%, H+L*L%

3. Validating the classic interface

In order to validate the classic information structure–prosody interface defined in terms of the correlation between theme/rheme of a sentence and its prosodic patterns, we carried out a number of experiments. In the context of these experiments, a non-expert native speaker of standard American English was instructed to read 109 sentences from the corpus annotated with information structure (composed of around 450 sentences from the Penn TreeBank).[4] The selection of those 109 sentences was based on the variation and complexity of their deep-linguistic (and thus also information) structures, making them prosodically interesting and useful for our study. The speaker was recorded in a professional recording studio to guarantee the quality of the sound signal. In a first stage, those sentences were annotated prosodically. In a second stage, the prosodic and theme/rheme patterns of these 109 utterances were assessed and contrasted with the patterns as prognosticated by Steedman's proposal.

3.1. Prosody annotation

Among several prosody annotations models, ToBI (Tone and Break Indices) [17] is the most widely used for annotating and adapting a markup language for open-source speech synthesizers such as Festival [18]. In our experiments, intonation patterns have also been annotated following the ToBI annotation convention. ToBI labels account for pitch accents (PA) within the intonational phrase (IP) and significant boundary tones (BT) within the sentence, which perfectly suits the purpose of these experiments. However, in contrast to most annotation exercises, which use a manual ToBI annotation,[5] we developed an own au-

[3]The hierarchical relations in a given thematicity segmentation are in practice controlled by parentheses (e.g. 'T(T)' will stand for "theme within the theme" and 'R(T)' for "rheme within theme". In coordinations and subordinations, each proposition is pointed out, e.g. T(P2) stands for the theme of the second proposition.

[4]The Penn Treebank was chosen as corpus base because it already contains the annotation at different linguistic levels (semantic and syntactic). Thus, we just added the information structure level to obtain all the information we need.

[5]The manual ToBI annotation has the advantage of being reliable and highly descriptive, but on the other hand it is subjective and very time consuming.

tomatic annotation interface that is based on AuToBi as initial labeling stage [19]. Since AuToBI only labels sentences word by word and our aim is to describe intonation patterns within intonational phrases (see [20] for details), we have processed the data automatically in a second stage to obtain a single ToBI pattern that is *a posteriori* manually validated and matched to the corresponding IP in the utterance. The advantage of working at the IP level is that we can correlate it with other layers, especially with IS.

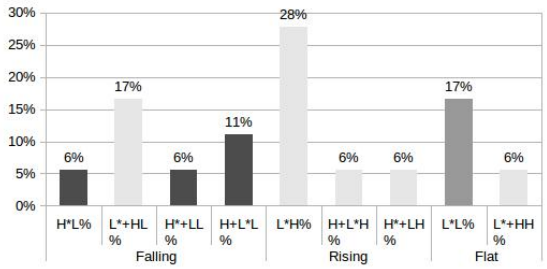
3.2. Assessment of the IS–prosody correlation

As already mentioned above, in the scope of our IS–prosody correlation assessment, we aimed to observe to what extent the classical proposals on the information structure–prosody interface can be applied to general (monologue) discourse with its rather complex sentential (and thus also thematic and prosodic) structures.

We analyzed pitch accents within the intonational phrase and complete intonation patterns (PABT — the combination of a pitch accent and a boundary tone). The analyzed themes include the set of patterns shown in Figure 1. All IPs are taken into account for calculating the percentages, regardless their position. It can be observed that rising PABT patterns together with L*+H tones represent a reasonable amount of the total (63%).[6]

(4) shows an example where the short theme, "Mister Kuehn", matches the pattern L*+H as expected by [1]; the same can be observed in (5), where the non-subjectival theme, "for some players", also matches the L*+H pattern.

Figure 1: *Theme ToBI Patterns.*



That is, as Steedman [1] claims there is indeed a tendency in themes to contain a rising pattern, at least in our recording, which includes only one speaker performing a reading task—although somewhat less than 40% of themes do not reflect the intonation pattern suggested in [1].

(4) *"Mister Kuehn, the company said, will retain the rest of the current management team".*

T	**Mister Kuehn**			
	H+L*H%			
R	**the company said**	**will retain the rest**	**of the current**	**management team**
	L*L%	L*+HL%	L*L%	L*L%

(5) *"For some players the lure is money up to fifteen thousand dollars a month".*

T	**For some players**			
	L*+HL%			
R	**the lure**	**is money**	**up to fifteen thousand**	**dollars a month**
	L*H%	L*H%	L*+HL%	L*L%

[6] It obviously remains to be proved that this tendency is kept in a big recorded corpus with several speakers and different registers, such as spontaneous speech.

However, in both (4) and (5), the rhemes do not show the expected pattern. In (4), there is no explanation for the L*L% pattern of the IP "the company said" and in (5), there is no explanation for any of the rising patterns found in the rheme span. Consequently, these two examples suggest that [1]'s approach to include everything, apart from theme, into a flat rheme span lacks the prediction accuracy we would need for speech synthesis applications.

4. Towards a more accurate IS–prosody interface

[1] is based on a linear dimension of thematicity. However, the study of our recorded material suggests that if we apply a hierarchical three-partite thematicity structure in the sense of Mel'čuk, we may be able (i) to find a justification for the discrepancies we saw in (4) and (5) between the prognosticated and the observed rheme patterns; (ii) propose a more accurate modelization of the intonation–thematicity correlation for the about 40% of non-coincident patterns within the theme span captured in Figure 1. Consider in (6) and (7) the sentences already cited in the examples in (4, 5), with a thematicity annotation as suggested by Mel'čuk.

(6) *"Mister Kuehn, the company said, will retain the rest of the current management team".*

T1	**Mister Kuehn**			
	H+L*H%			
SP1	**the company said**			
	L*L%			
R1	**will retain the rest**	**of the current**	**management team**	
	L*+HL%	L*L%	L*L%	

(7) *"For some players the lure is money up to fifteen thousand dollars a month".*

T1	**For some players**		
	L*+HL%		
T1(R1)	**the lure**		
	L*H%		
R1(R1)	**is money**	**up to fifteen thousand**	**dollars a month**
	L*H%	L*+HL%	L*L%

As already pointed out in Section 2.1, the notion of thematicity in the sense of Mel'čuk's includes, apart from theme and rheme, specifier elements. We have observed in our corpus that the identification of specifier elements provides very stable intonation patterns, being L*L% the commonest pattern, especially in reported speech. The IP "the company said" is a specifier (SP) and indeed carries this pattern. Another example in (8) also shows that the specifier element introduced in our information structure-prosody interface can be intonationally characterized as a separate entity from theme and rheme. This observation reinforces the tripartite division proposed by Mel'čuk.

(8) *"There is a large market out there hungry for hybrid seeds, he said".*

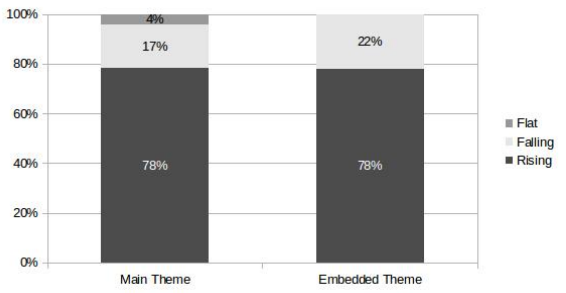
R1	**There is a large market out there**	**hungry**	**for hybrid seeds**
	L*+HL%	L*H%	L*L%
SP1	**he said**		
	L*L%		

In (7), the rheme element of (5) is hierarchically decomposed into an embedded theme/rheme structure. "The lure" becomes an embedded theme of the rheme and "is money up to fifteen thousand dollars a month" becomes the rheme of the rheme. As a consequence, we can at least explain why "the lure" carries a rising pattern: it is theme.

Going back to the results presented in Figure 1, a deeper analysis has been carried out to shed some light on the 40% of patterns within the theme span that do not coincide with the L*+H pattern in that they are either flat or falling and thus subsumed by the category Steedman called "among others". These diverging patterns are mostly found in complex structures, i.e., long themes containing more than one IP, coordinated sentences, subordinated clauses, etc. Consequently, an information structure model that can cater for complex utterances may be suitable to target this linguistic reality.

In order to test whether Mel'čuk's proposal of thematicity can be used as such a model, another experiment has been carried out. Thematicity, in accordance with Mel'čuk's definition, has been labeled in our selected corpus of 109 sentences following the guidelines outlined in [21]. Taking into account only theme spans which are prosodically marked up, we have classified their intonation patterns into rising, falling and flat, as shown in Table 2. The pattern L*+H L% is included into the rising pattern classification due to the fact that the state of the art considers this rising PA as characteristic of theme tunes. As a result, we have found that both main and embedded themes contain rising intonation patterns at equal rates of 78%. Consequently, we can affirm that embedded themes behave as main themes in terms of intonation when they are prosodically marked up. If we take into account the total of spans analyzed, main themes containing a rising pattern will characterize 34%, and when we add embedded themes we reach 50%. Therefore, we can conclude that a hierarchical approach to thematicity has the potential to provide more clues than traditional approaches when attempting to predict intonation patterns under complex communicative conditions.

Figure 2: *Comparison between main and embedded theme spans.*



In Section 3, we have already shown several examples where both main and embedded themes are characterized by rising tunes. The fact that data from a broader selection of utterances at an intra-speaker level also show this tendency sets a sound ground for further insight into the characterization of diverging intonation patterns based upon a hierarchical IS-prosody interface.

However, despite these advances in the explanation of the IS-prosody correlation there is still a substantial amount of cases that call for further investigation. These are, first of all, cases where a whole theme is not intonationally marked: 19% of primary themes and 49% of embedded themes. The example in (9) shows a sentence where the theme is a deaccented subject pronoun ("he") that, intonationally, forms part of the first IP of the rheme.

(9) *"Nevertheless he said he is negotiating with plant genetic to acquire the technology to try breeding hybrid cotton".*

SP1	**Nevertheless**				
	H*L%				
SP2	**he said**				
	L*L%				
T1	**he**				
R1	**is negotiating**	**with plant genetic**	**to acquire the technology**	**to try breeding**	**hybrid cotton**
	L*+HL%	H*+LL%	H+L*H%	H+L*H%	L*L%

The characterization of this kind of thematicity is also worthwhile to be born in mind, and further experiments will aim to find out if there is a characteristic IP in those cases. So far, we have observed that IPs containing a deaccented theme tend to bear a rising pattern, but the tendency is not so clear, and it seems that more layers interact at this level. Therefore, further research needs to be carried out, also including more speakers of the same dialect in order to test inter-speaker consistency and thus be able to draw more definite conclusions.

5. Conclusions

We have presented results of a descriptive study on a limited set of sentences from a wider corpus, attempting to determine which intonation patterns better characterize thematicity in real utterances, with the ultimate goal to build a model suitable for use in speech synthesis applications. We have observed that classical theories on IS–prosody interface are partially applicable in that themes of a specific (simple) nature have been proved to be characterized by a rising tune. On the other hand, the flat theme/rheme interpretation prevailing in these theories fails to explain complex linguistic structures. Drawing upon more advanced proposals on information structure, we have shown that further descriptive work needs to be done in order to accurately and concisely describe the IS–prosody correlation. Complex and hierarchical thematicity structures as well as the introduction of specifiers into the thematicity structure are bound to render positive results. Furthermore, the tri-partite division and the possibility of hierarchy are features of thematicity that facilitate a fine-grained communicative partition of complex utterances and thus a more detailed projection between the different layers of the semantics–syntax–intonation interface (or, more especifically, a more accurate description of the prosodic patterns related to each span). Our work can also be considered as detailing Steedman's proposal where he remains vague, stating that the patterns he identifies are only a few "among others".

The ultimate goal of developing a model combining prosodic and communicative structures for speech synthesis requires a deeper insight into descriptive studies on how these two linguistic layers interact. A good understanding of the structure and sequence of intonation patterns, as well as rare and/or exceptional cases will hopefully provide clues to more efficient NLP.

6. Acknowledgements

Parts of this work have been funded by a grant from the European Commission under the contract number FP7-ICT-610411. The second author is partially funded by a grant from the Spanish Ministry of Economy and Competitivity in the framework of the Juan de la Cierva fellowship program (JCI-2012-12272).

7. References

[1] Steedman, M., "Information structure and the syntax-phonology interface", Linguistic Inquiry, 4(31):649–685, 2000.

[2] von Heusinger, K., "Intonation and information structure". Habilitation dissertation, University of Konstanz, 2007.

[3] Büring, D., "Semantics, intonation and information structure". The Oxford Handbook of Linguistic Interfaces, ed. by Gillian Ramchad and Charles Reiss. Oxford University Press, 2007.

[4] Lambrecht, K., "Information structure and sentence form: Topic, focus and the mental representations of discourse referents". Cambridge University Press, Cambridge, 1994.

[5] Hajičová, E., Partee, B. and Sgall, P., "Topic-Foucs Articulation, Tripartite Structures, and Semantic Content. Kluwer Academic Publishers, Dordrecht, 1998.

[6] Mel'čuk, I. A., "Communicative Organization in Natural Language: The semantic-communicative structure of sentences". Benjamins Academic Publishers, Amsterdam, 2001.

[7] Erteschik-Shir, N., "Information Structure: The Syntax-Discourse Interface". Oxford University Press, Oxford, 2007.

[8] Beckman, M. and Pierrehumbert, J., "Intonational structure in Japanese and English". Phonology Yearbook, 3: 255–310, 1986.

[9] Moore, J.D., Foster, M.E., Lemon, O. and White, M., "Generating tailored, comparative descriptions in spoken dialogue". In Proceedings of FLAIRS-04, 917–922, Miami Beach, USA, 2004.

[10] White, M., Clark, R.A.J. and Moore, J.D., "Generating tailored, comparative descriptions with contextually appropriate intonation". Computational Linguistics, 36(2): 159–201, 2010.

[11] Sgall, P., "Functional sentence perspective in a generative description of language". Prague Studies in Mathematical Linguistics, 2: 203–224, 1967.

[12] Daneš, F., "Zur linguistischen Analyse der Textstruktur". Folia Linguistica, 4: 72–78, 1970.

[13] Llisterri, J., Carbó, C., Machuca, M.J., De la Mota, C., Riera, M. and Ríos, A., "El papel de la lingüística en el desarrollo de las tecnologías del habla". VII Jornadas de Lingüística, 137–191, 2003.

[14] Steedman, M., "Structure and Intonation". Language, 68: 260–296, 1991.

[15] Steedman, M., "Surface Structure, Intonation, and Focus". In Ewan Klein and Frank Veltman (eds.), Natural Language and Speech, Proceedings of the ESPRIT Symposium, Brussels, 21–38, 260–296. Dordrecht: Kluwer, 1991.

[16] Steedman, M., "The Syntactic Process", MIT Press, Cambridge MA, 2000.

[17] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: a standard for labelling English prosody". Proceedings of the ICSLP, vol. 2, 867-870, Sydney, Australia, 1992.

[18] Steedman, M., "Using APML to specify intonation". Magicster Project Deliverable 2.5. University of Edinburgh, 2005. Available at http://www.ltg.ed.ac.uk/magicster/deliverables/annex2.5/apml-howto.pdf

[19] Rosenberg, A., "AutoBI - a tool for automatic ToBI annotation". Proceedings of Interspeech, 146–149, 2010.

[20] Domínguez, M., Farrús, M., Burga, A. and Wanner, L., "Automatic extraction of prosodic patterns for speech synthesis applications". Submitted to Speech Prosody Conference 2014.

[21] Bohnet, B., Burga, A. and Wanner, L., "Towards the Annotation of Penn TreeBank with Information Structure". Proceedings of the Sixth International Joint Conference on Natural Language Processing, 1250–1256, Nagoya, Japan, 2013.

Prosody is in the hands of the speaker

Bahia Guellai [1], *Alan Langus* [2], *Marina Nespor* [2]

[1] Laboratory of Ethology Cognition and Development, University Nanterre, France
[2] SISSA Language Cognition and Development Lab, Trieste, Italy

bahia.guellai@gmail.com

Abstract

It has been suggested that speech and hand gestures could form a single system of communication that facilitates the interaction between the speaker and the listener. What kind of information do gestures carry? In the present study, we tested the possibility that spontaneous gestures accompanying speech carry prosodic information. Results show that gestures provide prosodic information as adults are able to perceive the congruency between a low-pass filtered – thus unintelligible - speech stream and the gestures of the speaker. These results suggest that prosody is not a modality specific phenomenon and can be perceived in spontaneous gestures that accompany speech.

Index Terms: prosody, hand gestures, speech perception.

1. Introduction

Human language is a multimodal experience: it is perceived through both the ears and the eyes. Adults automatically integrate auditory and visual information as evidenced by the McGurk effect [1], and seeing someone talking improves performances on speech intelligibility tasks [2]. This visual information involved in speech is not limited to the lips and the mouth but includes also the movements of the head [3, 4]. Other regions of the body could also give information about speech. Indeed, when interacting with others, people usually also produce spontaneous gestures while talking. What is exactly the role of these gestures that accompany speech? A line of research evidenced that gestures accompanying speech ease the speaker's cognitive load and gesturing help solving diverse tasks in mathematics and spatial problems [5, 6]. Gestures are also believed to aid the conceptual planning of messages as well as facilitate lexical access [7, 8]. This suggests that gestures and speech go 'hand-in-hand' from the earliest stages of cognitive development. In this view, gestures should carry the same structure as spoken language. One way to test this possibility is to look at prosody, an essential aspect of language.

In the auditory modality, prosody is characterized by changes in duration, intensity and pitch [9]. Interestingly, some part of the grammatical structure of human language is automatically mapped onto prosodic structure during speech production [10]. An interesting issue is whether prosody is modality specific or not. Since it has been shown to characterize sign languages as well [11], prosody cannot be restricted to the oral modality. It is therefore possible that the grammatical structure of language is not only automatically mapped to the acoustic speech signal but also to the spontaneous gestures accompanying speech.

Adult listeners use prosodic cues for various tasks that range from segmenting speech, to constraining lexical access [12], to disambiguating sentences that have more than one meaning (e.g., [bad] [boys and girls] vs. [bad boys] [and girls]) [10]. If some elements of grammatical structure are automatically mapped also to the spontaneous gestures accompanying speech, we should ask whether listeners use these gestures while processing the speech signal.

Thus, while there is evidence suggesting a direct link between the prosody of the speech signal and the spontaneous gestures that accompany speech, it is unclear whether listeners can use these cues provided by gestures when perceiving speech audio-visually. In the present study, we investigate the role of gestures as prosodic cues in speech perception.

2. Method

2.1. Experiment 1

In this first experiment, we explored whether gestures carry prosodic information. We tested Italian-speaking participants in their ability to discriminate audio-visual presentations of lowpass filtered Italian utterances where the gestures either matched or mismatched the auditory stimuli. While low-pass filtering renders speech unintelligible, it preserves the prosody of the acoustic signal [13]. This guaranteed that only prosodic information was available to the listeners.

2.1.1. Participants

We recruited 20 native speakers of Italian (15 females, mean age 24 ± 5) from the subject pool of SISSA – International School of Advanced Studies (Trieste, Italy). Participants reported no auditory, vision, or language related problems. They received a monetary compensation.

2.1.2. Stimuli

We used sentences that contain the same sequence of words and that can be disambiguated using prosodic cues from one of two different levels of the prosodic hierarchy. The disambiguation could take place at the Intonational Phrase (IP) level – the higher of these two constituents, coextensive with intonational contours – signaled through final lengthening and pitch resetting [10]. For example, in Italian, *Quando Giacomo chiama suo fratello è sempre felice* is ambiguous because depending on the Intonational Phrase boundary *è sempre felice* (*is always happy*) could refer to either *Giacomo* or *suo fratello* (*his brother*): (1) [Quando Giacomo chiama]IP [suo fratello è felice]IP (*When Giacomo calls him his brother is always happy*); or (2) [Quando Giacomo chiama suo fratello]IP [è felice]IP (*When Giacomo calls his brother he is always happy)*. Alternatively, the disambiguation could take place at the Phonological Phrase (PP) level where phrase boundaries are signaled through final lengthening. The Phonological Phrase extends from the left edge of a phrase to the right edge of its head in head-complement languages (e.g. Italian and English); and from the left edge of a head to the left edge of its phrase in complement-head languages (e.g. Japanese and Turkish) [10]. An example of a phrase with two possible meanings is *mappe di città vecchie* that is ambiguous in Italian because depending on the location of the PP boundaries, the adjective *vecchie* (*old*) could refer to either *città* (towns) or

mappe (maps): (1) [mappe di città]PP [vecchie]PP (old *maps of towns*); or (2) [mappe]PP [di città vecchie]PP (*maps of old towns*). The presentation of the two types of sentences – those ambiguous at the IP level and those ambiguous at the PP level - was randomized across subjects. We video recorded two native speakers of Italian – a male and a female – uttering ten different ambiguous Italian sentences (see Table 1). The speakers were unaware of the purpose or the specifics of the experiments. The speakers were asked to convey to an Italian listener the different meanings of the sentences using spontaneous gestures. The videos of the speakers were framed so that only the top of their body, from their shoulders to their waist, was visible (see Figure 1). Thus the mouth – i.e. the verbal articulation of the sentences – was not visible. Two categories of videos were created from these recordings using the Sony Vegas 9.0 software. One category corresponded to the 'matched videos' in which the speakers' gestures and their speech matched and the second category corresponded to the 'mismatched videos' in which the gestures were associated with the speech sound of the same sequence of words, but with the alternative meaning. A total of 80 videos were created (each of the sentences was uttered twice). We ensured that, in the mismatched audio-visual presentations, gestures and speech were temporally aligned so that the beginning and the end of the gestures were aligned with the beginning and the end of the speech act. To remove the intelligibility of speech but to preserve prosodic information, the speech sounds were low-pass filtered using the Praat software with the Haan band filter (0-400 Hz). As a result it was impossible to detect from speech which of the two meanings of a sentence was intended. The resulting stimuli had the same loudness of 70 dB.

Table 1. *Example of a sentence with two different meanings depending on its prosody.*

Sentence	**Meaning 1**	**Meaning 2**
Quando Giacomo chiama suo fratello è sempre felice.	Giacomo è felice.	Suo fratello è felice.
When Giacomo calls his brother is always happy.	*Giacomo is happy.*	*His brother is happy.*

Figure 1: *Examples of the stimuli presented.*

2.1.3. Procedure

Participants were tested in a soundproof room and the stimuli were presented through headphones. They were instructed to watch the videos and answer – by pressing a key on a keyboard – whether what they saw matched or mismatched what they heard (i.e., [S] = yes or [N] = no). A final debriefing ensured that none of the participants understood the meaning of the sentences.

2.1.4. Results

The results show that participants correctly identified the videos in which hand gestures and speech matched (M=81.9, SD=11.03: t-test against chance with equal variance not assumed $t(19)=12.93$, p<.0001) and those in which they did not match (M=69.3, SD=10.17; $t(19)=8.41$, $p<.0001$). A repeated measures ANOVA with condition (Match, Mismatch) and type of prosodic contour (Intonational and Phonological Phrase) was performed on the mean percentage. The ANOVA only revealed a significant main effect for condition ($F(1,19)=12.81$, $p=.002$, $\acute{\eta}^2 = 0.4$), but neither for type of prosodic contour ($F(1,19)=1.20$, $p=.287$, $\acute{\eta}^2 = 0.06$) nor for an interaction of type and condition ($F(1,19)=3.52$, $p=.076$, $\acute{\eta}^2 = 0.16$). The results show that adult listeners detect the congruency between hand gestures and the acoustic speech signal even when only the prosodic cues are preserved in the acoustic signal. The spontaneous gestures that accompany speech must therefore be aligned with the speech signal, suggesting a tight link between the motor-programs responsible for producing both speech and the spontaneous gestures that accompany it. The results of Experiment 1 thus also show that adult listeners are sensitive to the temporal alignment of speech and the gestures that speakers spontaneously produce when they speak. We thus asked whether the prosodic cues that adult listeners use for understanding spoken language may automatically be mapped to gestures.

3. Discussion

Our findings show that when presented with acoustic stimuli that contain only prosodic information (i.e., low-pass filtered speech), participants are highly proficient in detecting whether speech sounds and gestures match. The prosodic information of spoken language must therefore be tightly connected to gestures in speech production that are exploited in speech perception. The syntactic structure and the meaning of utterances are therefore not necessary for the perceiver to align gestures and prosody. As opposed to the visual perception of speech in the speakers' face, where the movements of the mouth, the lips, but also the eyebrows [14] are unavoidable in the production of spoken language, the gestures that accompany speech belong to a different category that is avoidable in speech production. Our results suggest that prosody is a domain-specific phenomenon (i.e., characteristic of language) that extends from the auditory modality to the visual one in speech perception. This link between speech and gestures is congruent with neuropsychological evidence for a strong correlation between the severity of aphasia and the severity of impairment in gesturing [15]. While further studies are clearly needed to identify the specific aspects of spontaneous gestures that are coordinated with speech acts, our results demonstrate that part of speech perception includes the anticipation that bodily behaviors, such as gestures, be coordinated with speech acts.

4. Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° 269502 (PASCAL), and the Fyssen Foundation.

5. References

[1] McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746-748.

[2] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

[3] Graf, P. H., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington D.C.

[4] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & VatikiotisBateson, E. (2004).Visual prosody and speech intelligibility. *Psychological Science, 15*, 133-137.

[5] Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General, 140*, 102-115.

[6] de Ruiter, J. P., Bangerter, A., & Dings, P. (2012), The Interplay Between Gesture and Speech in the Production of Referring Expressions: Investigating the Tradeoff Hypothesis. *Topics in Cognitive Science, 4*, 232-248.

[7] Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes, 15*, 593-613.

[8] Pine, K. J, Bird, H., & Kirk, E. (2007). The effects of prohibiting gestures on children's lexical retrieval ability. *Developmental Science, 10*, 747-754.

[9] Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken
language: A literature review. *Language and Speech, 40*, 141-201.

[10] Nespor, M., & Vogel, I. (2007). *Prosodic Phonology*. Berlin. Mouton De Gruyter.

[11] Nespor, M., & Sandler, W. (1999). Prosody in Israeli Sign Language. *Language and Speech. 42*, 143-176.

[12] Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language, 51*, 523-547.

[13] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago:
University of Chicago Press.

[14] Krahmer, E., & Swerts, M. (2004). More about brows: A cross-linguistic analysis-by-synthesis study, In: C. Pelachaud & Zs. Ruttkay (Eds.) *From Brows to Trust: Evaluating Embodied Conversational Agents*. Kluwer Academic Publishers.

[15] Cocks , N., Dipper, L., Pritchard, M., & Morgan, G. (2013): The impact of impaired semantic knowledge on spontaneous iconic gesture production. *Aphasiology*.

Rising pitch and quoted speech in everyday American English

Joseph Tyler[1]

[1] Department of English Literature and Linguistics, Qatar University, Doha, Qatar
josephctyler@gmail.com

Abstract

Phonetic variation in rising pitch has been analyzed for how it correlates with contextual factors like speaker gender, utterance type (questions vs. statements) and turn position (turn-medial vs. turn-final). This paper analyzes variation in terminal rising pitch between quoted and non-quoted speech, using data from the Santa Barbara Corpus of Spoken American English. Results show rises in quoted speech start and end higher, rise more overall, have steeper slopes but are no different in duration. These results are gender-dependent, however, for while women produce 65% of all rises in the corpus sample, they produce 100% (n=23) of the quoted speech rises.

Index Terms: rising pitch, intonation, quoted speech, reported speech, gender, uptalk, HRT, corpus analysis

1. Introduction

Terminal rising pitch, also known as uptalk or the High Rising Terminal, has received attention for its semantic and pragmatic meanings [1, 2], its social meanings [3, 4], its perception effects [5, 6] and its production patterns [4, 7-10]. Production studies have used corpora to explore phonetic variation in rises and how such variation correlates with contextual factors like gender, utterance type (questions vs. statements) and turn position (turn-medial vs. turn-final) [11, 12]. Most of these studies used speech from a corpus of map task data [13]. This study makes two novel contributions: first, it uses everyday conversational American English; second, it looks at how rises in quoted speech differ from rises in other contexts.

1.1. Literature Review

Corpus studies of rising pitch in speech production have shown systematic correlations between phonetic dimensions of those rises and contextual factors. The rises have been analyzed phonetically for their starting pitch, ending pitch, rise onset position, rise span, and pitch dynamism [4, 11, 12]. Some studies code phonological representations of a rising pitch contour [14], using a transcription system like ToBI [15], and other studies combine both phonetic and phonological analyses.

This variation in rise form has been correlated with a range of contextual factors, e.g. speaker gender, utterance type (statement vs. question) and turn position (turn-medial vs. turn-final). Results show systematic differences in rise production between men and women: women have been found to produce more rises [4, 9], to start their rises later [3, 4], and to use more pitch dynamism [12]. In addition, results for speakers in Southern California, London and New Zealand suggest that women produce bigger rises than men [3, 4, 9, 12]. This result persists even when using the ERB scale, a scale designed to control for physiologically-determined differences [12].

In addition to gender effects, rises have been found to vary by utterance type, turn position, dialect, age, and a rise's position in the discourse [4]. Other corpus work using different measures of intonation have examined pitch variation by speech register [14] or speech situation [16]. What has not received as much attention is how rises vary by the stylistic implementation of the speech by the speaker. Rampton [17] describes stylization as "reflexive communicative action in which speakers produce specially marked and often exaggerated representations of languages, dialects, and styles that lie outside their own habitual repertoire (at least as this is perceived within the situation at hand)". One such type of stylization may be the performance of quoted speech, especially in the recounting of a narrative.

Quoted speech, also known as direct reported speech, has been studied from a number of perspectives. The intonation of quoted speech was analyzed many years ago by Bolinger [18], who argued that the goal of the speaker in such situations is to "re-enact the original intonation," though the achievement of this goal may be more or less successful. Klewitz and Couper-Kuhlen [19] discuss how quoted speech is sometimes set apart intonationally from the discourse around it, e.g. with prosodic changes at the boundaries between quoted and non-quoted speech. But more often, they report, intonation may mark reported speech but not exactly at the boundaries, serving as a more indirect cue to what is quoted. In addition, a variety of discourse types and data sources have been used to show quoted speech is produced with wider pitch range than non-quoted speech [20-22].

From this work on intonation and quoted speech, it remains unclear how terminal rises are produced in quoted speech compared to rises in other contexts, though the fact that quoted speech often has wider pitch range suggests rises may be larger in quoted speech than elsewhere. The data for many of the corpus studies discussed above have been elicited through map tasks [4, 9, 11, 23, 24], leaving open the question of how such behavior generalizes to more everyday speech contexts. In this study, phonetic features of rises in everyday conversational English are examined and compared in both quoted and non-quoted speech. The results can tell us something about how the quoted speech is realized in rises and how those rises compare to rises more generally.

2. A Corpus Analysis

The spoken data analyzed in this study come from Disc 1 of the Santa Barbara Corpus of Spoken American Spoken English (henceforth SBC). The SBC includes conversations from different regions of the United States, discussing a variety of topics in everyday settings. Some conversations are dyadic and some are multi-party (up to 9 speakers). This corpus sample contains 51 unique speakers across 14 different conversations; except for two speakers who were in two conversations, all speakers were in only one conversation. Recordings lasted an average of 67 minutes, with the shortest and longest being 28 minutes and 120 minutes respectively. They took place in a range of settings, including private settings (living room, kitchen, bedroom, dining room) but also professional contexts (an office, a classroom or a bank).

Speakers' home states were around the US, with more common states being California, Indiana and Illinois, but also Alabama, Montana and Texas. Thirty-one speakers lived in their home state, ten were listed as living in a different state than their home state, and ten had this information missing.

In addition to the naturalness of the data in this corpus, a second important benefit of the SBC is the transcription system used. All conversations have been transcribed into time-stamped intonation units with terminal pitch marked as either rising (marked with a "?"), flat (marked with a ",") or falling (marked with a "."). A simple search for intonation units marked with a "?" reveals hundreds of rises. In addition to the relative ease of identifying the rises, it also supplies a sample of *perceived* rises, regardless of their specific phonetic characteristics.

All rises that the transcript suggested were phonetically analyzable were annotated to a TextGrid in Praat [25]. Cases where the transcript indicated that environmental noise or overlapping speech masked the rising pitch were excluded. This resulted in a set of 636 tokens from Disc 1 of the SBC that were then subjected to phonetic analysis.

The rises were analyzed phonetically using a pitch window of 50-500Hz. The rise domain was annotated from the start of the rise to the end. The rise start was determined impressionistically, though guided by the phonetic trough and with a slight buffer after the syllable onset to reduce segmental effects. The rise start usually occurred on the last stressed syllable of the intonation unit. The rise end was the phonetic peak, i.e. the point at the end of the rise with the highest Hz value.

All tokens were analyzed with a Praat script that extracted the rise-start pitch (in Hz) and the rise-end pitch (in Hz), as well as rise duration information. During the original TextGrid annotation, some tokens did not exhibit a pitch value or showed an unreliable pitch value (e.g. apparent halving or doubling errors). Because these instances were going to result in unreliable output from the Praat script, they were marked for follow-up manual analysis.

The tokens marked for manual measurement were then analyzed individually. During manual measurement, the pitch settings were adjusted according to the author's discretion to control for halving or doubling errors. For some rises, the script outputted Hz values of zero; these tokens were analyzed individually. Outliers were also examined manually; outliers were identified as values at the extremes (<75Hz, >400Hz, and with a span that was >200Hz or negative) or those that were more than two standard deviations away from a speaker's mean. Following the manual analyses, (5%) of the 636 total tokens ended up being excluded because they were unmeasurable. The remaining 603 tokens served as the basis of the corpus analysis. These tokens were produced by 39 unique speakers (see Table 1 for the distributions across speakers). In addition to rise-start and rise-end, variables were also defined for rise-span, defined as rise-end minus rise-start, and rise-slope, defined as rise-span(Hz) divided by duration(seconds).

While the original phonetic measurements were taken on the linear Hz scale, they were converted in two ways for analysis: 1) log-transformation, and 2) conversion to the ERB (Equivalent Rectangular Bandwidth) scale. The log transformations resulted in a more normal distribution. The ERB scale was developed using psychoacoustic tests to track changes in Hz to perceived changes in prominence, and has been argued to be more appropriate for the analysis of intonation [26]. It has been argued to better correspond to listeners' perceived pitch scale. It has also been used in other research on rising pitch [3]. Duration measures were log-transformed to increase the normality of the distribution.

In addition to the phonetic measurements, each rise was coded for various contextual factors, including quoted speech, speaker gender, utterance type (question vs. statement) and turn position (turn-medial vs. turn-final). While part of a larger project, this paper focuses on rises in quoted and non-quoted speech. For this study, quoted speech was coded exclusively as direct reported speech, i.e. where the speech was presented as said exactly as it was said in the past. In each of the following examples, the reported speech ("oh you mean adults" and "oh so you're going to host them are you") is presented as a word-for-word representation of what was said in the original speech context. The original transcription format of the SBC is maintained.

(1) PAMELA: (H) she said,
oh you mean,
... adults=?

(2) ALINA: And I said,
oh.
.. So you're going to host them are you?

As is visible in these two examples, the quoted speech is introduced with the quotative "to say". In this corpus sample, all 23 tokens of quoted speech were introduced with a quotative, 17 with the verb to say (e.g. (1) and (2)), 6 with "to go" (e.g. (3)).

(3) CAROLYN: And they look at] you and they go,
... <Q the what Q>?

While it is not the focus of this study, quotative use has been examined for its systematic variability, e.g. in the context of narrating a life story [27] as well as in language change [28].

Table 1: *Distribution of rises by speaker*

Speaker	Gender	#Rises	#Quoted	Speaker	Gender	#Rises	#Quoted
1	f	65		22	f	2	
2	f	9		23	f	1	
3	f	1		24	m	51	
4	f	19		25	f	13	
5	m	1		26	m	6	
6	m	14		27	f	12	
7	m	12		28	f	12	1
8	f	22	1	29	f	4	
9	m	13		30	m	62	
10	m	17		31	f	1	
11	f	9		32	f	2	
12	f	16	2	33	f	18	1
13	m	1		34	f	12	
14	m	20		35	f	13	
15	f	27	3	36	m	4	
16	f	39	12	37	m	8	
17	f	8		38	m	11	
18	f	33	1	39	m	4	
19	f	35	1	40	m	1	
20	f	19		41	m	3	
21	f	16					

Table 1 lists the number of rises and quoted rises by speaker, showing that all quoted speech tokens were produced by women. Therefore, all results are gender-dependent, for while women produce 65% of all rises in the corpus sample, they produce 100% (n=23) of the quoted speech rises. Moreover, while eight different women in the corpus sample produced measurable rises in quoted speech contexts, a single woman (speaker 16) produced a majority (52%).

2.1. Results

Table 2 shows the distribution of untransformed values for rise-start, rise-end, rise-span, rise-duration, and rise-slope (risespan/duration) in quoted and non-quoted speech.

Table 2. *Descriptive statistics for rise measurements.*

Quoted	N	Rise-start (Hz)	Rise-end (Hz)	Rise-span (Hz)	Rise-dur (sec)	Rise-slope
Min	23	121	160	6	0.105	27
Max	23	361	745	425	0.634	1139
Mean	23	215	320	105	0.290	404
SD	23	56	127	101	0.145	358
Non-Quoted						
Min	580	79	91	-24	0.047	-113
Max	580	350	636	352	1.716	1724
Mean	580	174	224	50	0.333	189
SD	580	48	79	51	0.228	204

The distribution suggests quoted speech rises tend to start higher, end higher, rise more and be steeper. In this table, minimum and maximum rise-span may not correspond to the difference between rise-end and rise-start because they could be the values for different tokens.

In order to test whether quoted speech rises are produced differently from non-quoted rises, mixed models were used, which have been found to have advantages over other models [29]. Random effects for speaker and conversation (the sound file from which each speaker's data are drawn) were included in order to control for variation due to speaker- or conversation-specific differences. The predictor variable was a binary variable for quoted vs. non-quoted speech. The dependent variables were the rise-start, rise-end, rise-span, and rise-slope values on both log(Hz) and ERB scales. Also included was a dependent variable for rise duration (in seconds, log-transformed). These models were implemented using the `lmer` function in R [30]. Because the `lmer` function does not output a p-value, p-values were retrieved with the `pvals.fnc` function.

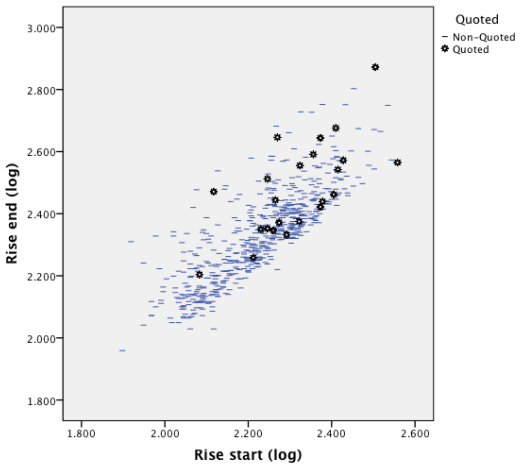
Results in Table 3 show that quoted speech has higher rise-start pitch, rise-end pitch, rise-span, and rise-slope in both Hz and ERBs. No difference was found for rise durations. These results indicate that rises in quoted speech start higher, end higher, rise more and more steeply in this corpus sample. They do not, however, rise for a longer duration. In order to test whether the pitch results depend on duration differences, the same models were run for rise-start, rise-end, and rise-span but with log(rise-duration) as a covariate. In these models, quoted speech remained a significant predictor for all pitch measures, with rise-duration also significantly predicting the rise-end and rise-span measures, but not the rise-start measures. And when data from male speakers are excluded, significant effects for the quoted speech variable remain.

Table 3. *Quoted vs. non-quoted speech rises.*

	Quoted vs. non-quoted speech
Rise-start (log(Hz))	ß=.074, SE=.016, t=4.56, pMCMC<.001
Rise-start (ERB)	ß=.003, SE=.001, t=4.56, pMCMC<.001
Rise-end (log(Hz))	ß=.117, SE=.024, t=4.93, pMCMC<.001
Rise-end (ERB)	ß=.005, SE=.001, t=4.93, pMCMC<.001
Rise-span (log(Hz))	ß=.087, SE=.026, t=3.33, pMCMC=.001
Rise-span (ERB)	ß=.002, SE=.001, t=2.382, pMCMC=.021
Rise-slope (risespan(log) /duration)	ß=.234, SE=.097, t=2.43, pMCMC=.014
Rise-slope (risespanERB /duration)	ß=.007, SE=.003, t=2.073, pMCMC=.042
Rise duration (log(seconds))	ß=-.024, SE=.059, t=-.40, pMCMC=.690

Figure 1 presents results for quoted and non-quoted rise-starts and rise-ends.

Figure 1: *Scatterplot for quoted and non-quoted rises*



While there is overlap in the distributions, this plot shows that quoted speech rises tends to start and end higher in pitch.

Among the rises in quoted speech, some tokens reach the extremes of pitch excursion. For example, on the word "balcony" in the excerpt in (4), the speaker reaches almost 800Hz. The pitch contour for this rise is plotted in Figure 2.

(4) ALINA: of course I said,
.. <VOX ~Cassandra,
... you wanna play one [bounce off the] balcony VOX>?

The domain of the rise is marked in Figure 2 as the portion between the two vertical dotted lines and corresponds to the nucleus of the first syllable to the nucleus of the last syllable

of the word "balcony". This remarkable rise is in the context of narrating a story about someone who was jumping. The narrator is annoyed, and quotes her own speech in a dramatic and excited manner. The speaker's excitement in the story may have resulted in extremely high pitch on the final rise.

Figure 2: *Pitch contour for the utterance "wanna play one bounce off the balcony?" produced by speaker 16 (Alina), with the rise domain indicated by vertical dotted lines.*

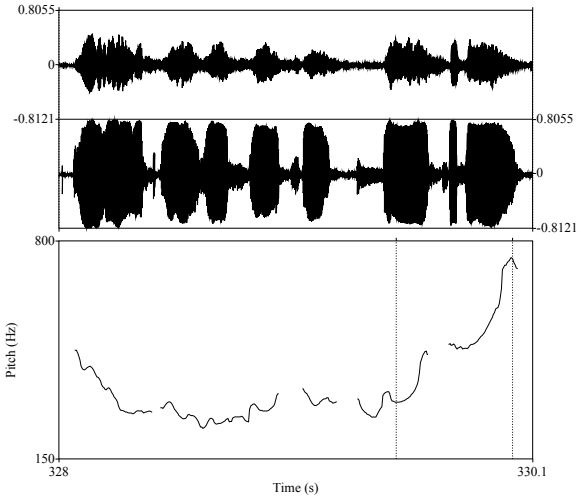
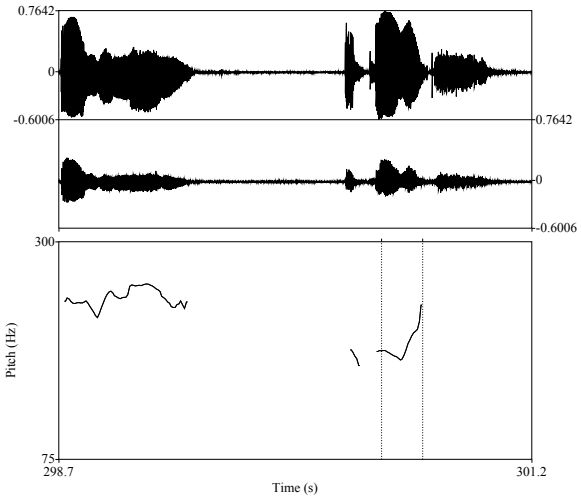


Figure 3: *Pitch contour for the utterance "oh you mean… adults?" produced by speaker 15 (Pamela), with the rise domain indicated by vertical dotted lines.*



Another speaker seemed to exploit pre-rise pauses as part of her repertoire. For example,

(5) PAMELA: (H) she said,
oh you mean,
... adults=?

The domain of the rise covers the onset through the end of the word "adults". As is visible in Figure 3, there is a sizeable pause (794ms) between the words "mean" and "adults". The pitch contour also shows that the speaker drops in pitch before starting the rise. This study has not explored the relationship between the pitch of terminal rises and pre-rise pitch, but (5) suggests that that relationship may display meaningful variation.

While this paper has not specifically addressed distinctions between types of rises other than being quoted or not, one popular contrast is between rises in polar questions and uptalk (often defined as rises on declarative syntax that are not questioning [31]). Uptalk has been derided in popular media [32] and studied in scholarly work [5]. Statistical models were run again with variables for rise duration, questioning speech act and interrogative syntax as covariates (to control for the factors of uptalk). In all of these models, the effects of quoted speech on the pitch measures remain significant. Therefore, the findings reported here for quoted speech are not reducible to an uptalk effect.

3. Discussion

The results discussed above suggest that in the context of quoted speech, (female) speakers produce rises that start higher, end higher, rise more overall and have a steeper rise. These results complement existing research that has determined quoted speech in general to be characterized by higher than average pitch [20-22]. One possible explanation for this is that the act of quoting, i.e. of representing speech from another time, is generally dramatic. Rendering speech from outside the immediate context is a kind of performance. The nature of that performance includes higher, larger and steeper than average rises.

The findings that quoted speech rises are produced with different intonation from non-quoted speech rises may contribute to ongoing discussions on the nature of speech stylization. Given Rampton's [17] definition of stylization referenced above, quoted speech may itself be a case of stylization, where the speech is marked and exaggerated, lying outside the speaker's habitual repertoire. Notably, the speaker could be quoting themselves, as in (4) above. While the original production that the speaker is quoting is not available for comparison, it seems likely the original did not end in a rise that reached almost 800 Hz (if, in fact, she said these words at all). Therefore, her goal may not necessarily be to most accurately reproduce the original speech, as Bolinger [18] had claimed. Instead, the special circumstances of quoting, even when the one being quoted is oneself, can lead a speaker to mark the quoted speech as other. One of the mechanisms by which speakers can mark quoted speech as other is through distinctive intonation. The results discussed above suggest that the size and slope of a rise, in addition to wider pitch range [20-22], can help stylize quoted speech as marked speech.

4. Acknowledgements

The author would like to thank Paul Warren for sharing references and Irene Theodoropoulou for input.

5. References

[1] Lai, C., "Rises all the way up: The interpretation of prosody, discourse attitudes and dialogue structure," Linguistics, University of Pennsylvania, 2012.

[2] Gunlogson, C., "A Question of Commitment," *Belgian Journal of Linguistics,* vol. 22, pp. 101-136, 2008.

[3] Warren, P. and Daly, N., "Sex as a factor in rises in New Zealand English," in *Gendered speech in social context: Perspectives from gown and town*, J. Holmes, Ed., ed Wellington: Victoria University Press, 2000, pp. 99-115.

[4] Ritchart, A. and Arvaniti, A., "Do We All Speak Like Valley Girls? Uptalk in Southern Californian English," presented at the Acoustical Society of America, San Francisco, 2013.

[5] Tomlinson Jr, J. M. and Fox Tree, J. E., "Listeners' comprehension of uptalk in spontaneous speech," *Cognition,* vol. 119, pp. 58-69, 2011.

[6] Tyler, J., "Discourse Prosody in Production and Perception," PhD, Linguistics, University of Michigan, Ann Arbor, 2012.

[7] Fletcher, J., Grabe, E., and Warren, P., "Intonational variation in four dialects of {E}nglish: the high rising tune," in *Prosodic typology: The Phonology of Intonation and Phrasing*, S. Jun, Ed., ed: Oxford University Press, 2005.

[8] Warren, P., "Patterns of late rising in New Zealand English," *Language Variation and Change,* vol. 17, 2005.

[9] Barry, A. S. and Arvaniti, A., ""Uptalk" in Southern Californian and London English," in *BAAP 2006 Colloquium*, ed. Queen Margaret University College, Edinburgh, 2006.

[10] Guy, G., Horvath, B., Vonwiller, J., Daisley, E., and Rogers, I., "An Intonational Change in Progress in Australian English," *Language in Society,* vol. 15, pp. 23-51, 1986.

[11] Fletcher, J. and Harrington, J., "High-Rising Terminals and Fall-Rise Tunes in Australian English," *Phonetica,* vol. 58, pp. 215-229, 2001.

[12] Daly, N. and Warren, P., "Pitching it differently in New Zealand English: Speaker sex and intonation patterns," *Journal of Sociolinguistics,* vol. 5, pp. 85-96, 2001.

[13] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., *et al.*, "The Hcrc Map Task Corpus," *Language and Speech,* vol. 34, pp. 351-366, 10 1991.

[14] Carmichael, L. M., "Situation-Based Intonation Pattern Distribution in a Corpus of American English," PhD, Linguistics, University of Washington, 2005.

[15] Silverman, K. E. A., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C. W. S., and Price, P., "ToBI: A standard scheme for labeling prosody," in *Proceedings of the 2nd International Conference on Spoken Language Processing*, Banff, Canada, 1992, pp. 867-879.

[16] Cheng, W. and Warren, M., "// / CAN i help you //: The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English," *International Journal of Corpus Linguistics,* vol. 10, pp. 85-107, 2005.

[17] Rampton, B., "Interaction ritual and not just artful performance in crossing and stylization," *Language in Society,* p. 149, 2009.

[18] Bolinger, D. L., "The intonation of quoted questions," *Quarterly Journal of Speech,* vol. 32, pp. 197-202, 1946/04/01 1946.

[19] Klewitz, G. and Couper-Kuhlen, E., "Quote-unquote. The role of prosody in the contextualization of reported speech sequences," *Pragmatics,* vol. 9, pp. 459-485, 1999.

[20] Wennerstrom, A., *The Music of Everyday Speech:Prosody and Discourse Analysis*: Oxford University Press, 2001.

[21] Jansen, W., Gregory, M. L., and Brenier, J. M., "Prosodic correlates of directly reported speech: Evidence from conversational speech," presented at the ITRW on Prosody in Speech Recognition and Understanding, Molly Pitcher Inn, Red Bank, NJ, USA, 2001.

[22] Hirschberg, J. and Grosz, B., "Intonational Features of Local and Global Discourse Structure," presented at the Proceedings of the Speech and Natural Language Workshop, 1992.

[23] Fletcher, J., Wales, R. J., Stirling, L. F., and Mushin, I. M., "A dialogue act analysis of rises in Australian English map task dialogues," in *Speech prosody 2002: Proceedings of the 1st international conference on speech prosody*, ed Aix-en-Provence: Universite de Provence, 2002, pp. 299-302.

[24] Shobbrook, K. and House, J., "High Rising Tones in Southern British English," in *15th International Congress of Phonetic Sciences*, Barcelona, 2003, pp. 1273 - 1276.

[25] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer," ed, 2009.

[26] Hermes, D. J. and van Gestel, J. C., "The frequency scale of speech intonation," *J Acoust Soc Am,* vol. 90, pp. 97-102, Jul 1991.

[27] Keller-Cohen, D. and Gordon, C., ""On Trial": Metaphor in Telling the Life Story," *Narrative Inquiry,* vol. 13, pp. 1-40, 2003.

[28] Cukor-Avila, P., "She say, She go, She be like: Verbs of Quotation over Time in African American Vernacular English," *American Speech,* vol. 77, pp. 3-31, 2002.

[29] Baayen, R. H., Davidson, D. J., and Bates, D. M., "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language,* vol. 59, pp. 390-412, 2008.

[30] R Development Core Team, "R: A language and environment for statistical computing," ed. Vienna, Austria: R Foundation for Statistical Computing, 2012.

[31] Cruttenden, A., *Intonation*. Cambridge [U.K.] ; New York, NY, USA: Cambridge University Press, 1997.

[32] Davis, H. (2010, October 6) The Uptalk Epidemic: Can you say something without turning it into a question? *Psychology Today*. Available: http://www.psychologytoday.com/blog/caveman-logic/201010/the-uptalk-epidemic

Fine temporal structure of Finnish sign language

Daniel Aalto[1], *Stina Ojala*[2]

[1]University of Helsinki, Finland
[2]University of Turku, Finland
daniel.aalto@helsinki.fi, stina.ojala@utu.fi

Abstract

Signs can be divided to syllables and further into transitions and nuclei based on the sign flow of the handshapes. Here, a mixed effects linear regression model is used to describe the variation in the duration of the syllable nuclei in a data set of 341 signs (474 syllables) produced by five native FinSL signers during a map task. The phonetic fixed variables are the duration of the adjacent transitions and syllable nuclei; phonological fixed variables are the syllabic length of the sign, the syllable position within the sign, and the sign type (functional or content bearing). Both preceding and following nucleus had a significant effect on the nucleus duration, while an asymmetric effect was found for the transitions: only the postnuclear transition had a significant effect. The syllable structure had no effect. However, the nuclei were shorter in function signs. These results suggest that signs are produced in two stages where the first stage, preparatory transition, is merged with the production of the previous syllable, and the second stage consists of executing the sign.

Index Terms: Finnish sign language, spontaneous signing, mixed effects model

1. Introduction

Introduction is divided into three subsections, which address different aspects of sign language research respectively. Signs in a sign language correspond to words in a spoken language, and thus they are the basic carriers of meaning in the particular language. The sign flow divides into individual signs and further into syllables. However, the notion of a syllable is under a debate within sign linguistics community. Nevertheless, it is accepted that signs divide temporo-spatially to segments/gestures. In the current work we study the impact of lexical and segmental factors to the duration of signed syllable nuclei.

1.1. Temporal organization of signed languages and prosody

Previous studies have shown that facial expressions can switch the function of a signed utterance from a statement to a question [1]. Facial expressions are markers for linguistic stress as well [2], but as Wilbur et Nolen discovered, it does not increase the duration of the signs as such but manifests in faster movement and larger displacement instead [3].

Temporal structure and prosodic features of signed language have been studied mainly in relation to the linguistic scope [4], [5] while this study concentrates on the articulatory issues related to sign flow *per se*.

Kröger et al. [6] compare the principles of syllable structures in relation to articulatory phonology. This study concentrates on the precise syllable-internal structures and their durations. The current project compiles the approach by Kröger et al. in comparing sign data to articulatory phonology principles, and the approach by Wilbur et Nolen [3] in studying syllable durations.

Signs within a sign language are further divided into articulatory gestures. These alternate temporally within sign stream, so that the stream has salient movements and motionless moments called holds. Holds and movements represent the sequential organisation of a sign flow. It also forms the basis for our definition of a signed syllable.

Recent sign language research has concentrated on rhythm in connection with signed poetry, but some of the phenomena found are not present in ordinary signed discourse [7]. Studies on sign coarticulation have suggested that the index finger has a dominant role in governing the rate and speed of both coarticulatory and interarticulatory phenomena [8]. The intertwining rhythms in sign are found within manual and facial elements. One of the key facial elements in signed discourse are the headnods [9]. Sign language prosody is also expressed in facial expressions, however, they are not in scope in this study. The twofold structure of prosody is similar to spoken language prosody division between segmental and suprasegmental features.

In a previous study [10] we showed that syllable durations change in terms of sentence position and sign type. The data showed a reduction of duration in polysyllabic signs and it would suggest that spatial relations have a special role in timing of signed languages. Now we want to take the study questions one step further and look if there are differences according to the syllable type (nucleus vs. transition), and what type of interaction between sign types (function vs. content sign) can be found. Function signs, such as pronouns and conjunctions, denote grammatical functions whereas content signs account for the meaning in the sentence. The position of the sign in a signed utterance affects its duration but not its linguistic stress.

1.2. Coarticulation, signing rate, and segmental factors

Just like isolated words differ from the words spoken in their context because of coarticulatory patterns and phenomena, the signs produced in isolation, i.e. in their dictionary form, are also different from those produced within their context. In signed languages, the coarticulatory patterns differ in several dimensions, as not only the signs in succession within one hand influence each other, but the hands also interact together to form another layer of coarticulatory patterns. This is sometimes called interarticulation. For example, the hands slow down when they are nearer to each other [8], [11]: p. 23.

The two main models of sign language phonology share the goal of segmenting an on-going sign stream, but approach it from different perspectives. The segmental models, includ-

ing HM (hold and movement) model, [3], [12], [4] and the latest contemporary model [13], derive from a more linguistic perspective, while the gesture model takes a more physiology based approach to the problem, and bases the segmentation principles on the general principles of human movement: the alternation of more and less active segments during the signed utterance [14] . This results in a different type of segments within the sign stream. In speech production theories the derivatives of articulatory phonology [15] and motor theory of speech production [16] most resemble the above-mentioned. In this study we reflect on our data in the light of these two types of theories, segmental and gestural theories.

Human action is often a combination of many intertwining rhythms, be it basic human actions, such as walking or eating [17], or something more complex, such as speaking or playing an instrument. The rhythms we use are highly individual. Speech rhythm is one of the key features we observe when trying to identify a speaker [18] – the individual "figure of speech".

Lindblom, Mauk et Moon [19] study both sign and speech in equal terms based on the dynamics of the production in a sequence. Here too, the motor equilibrum theory with the human movement principles forms the basis for the analysis. According to Lindblom et al. we should not concentrate on the distinct categories *per se*, but on the interpolation between categories. They ask: if perception likes change, why is phonetic specification built mainly around steady-state attributes?

1.3. Mental lexicon and dictionary forms

The articulatory gestures in spontaneous signing are formed by hands (elbows, wrists, metacarpals and phalanges) in relation to the body (face, chest and shoulders). The particular geometric form taken by bones distal to wrists is called the handshape. Often, the handshapes are static, but fingers might move to create a dynamic handshape. Signs in a sign dictionary are categorised by handshapes, even though handshapes are one of the most difficult parts of the sign, and one of the last elements to be learned by sign acquiring children [18].

As stated earlier, individual signs are the basic building blocks of sign language and as such, they are also the meaning-carriers. In addition, signs are collected into sign language dictionaries, similar to dictionaries of spoken languages. However, unlike spoken languages, signed languages do not have written forms, so a sign language dictionary is based on static pictures or videos. Dictionary forms present normative versions of signs and they also guide the language teachers in their teaching, similarly to how spoken language teachers use dictionaries in their teaching [20].

These dictionary forms are thought to be the basic, prototypical representations in our mental lexicon as well. These mental images or soundscapes of words, signs and sentences are adjusted according to the person we are communicating with. This is done based on so-called perceptual anchors, such as greeting words. In English the word that is often used is "Hello". This gives the acoustic guidelines to the speaker's vocal tract. Similarly in signed language a hand wave is used to greet. This gives many clues to the use of physical space anchors, such as size of hand, positioning of hand in relation to face etc. [21].

For the reasons stated above, the dictionary forms found in Finnish Sign Language online dictionary Suvi (http://suvi.viittomat.net) were used as the main reference guide where applicable to the syllabification processing of the spontaneous signed data. After the syllabification, the signs were labelled as monosyllabic or polysyllabic, and as function or content signs. Syllables were further categorised as nuclei or transitions.

In this study, sign language prosody is in focus, in particular the temporal variation within the constituents of the signs in spontaneous Finnish Sign Language (FinSL). Both phonetic (duration of adjacent constituent) and linguistic factors will be considered. We concentrate on the fine temporal structure of Finnish Sign Language based on the movement patterns of the dominant hand. The purpose of the article is to investigate if lexical factors can affect the duration of the syllables.

2. Methods

2.1. Subjects and data recording

The recording was done in a well-lit office. No spotlights were used but lighting was kept constant by shutting the blinds and curtains to prevent possible glare from outside. The camera was on a tripod 2 metres from the subject facing the subject directly. The subject was facing the camera and was sitting on a sofa to prevent the obstruction of signing space by armrests on one hand and to make the situation more comfortable for the subject at the same time. The equal positioning of the camera and the subjects was ensured by using tape markings on the floor for the camera tripod and the subject's feet position. The background was a light brick wall with solid colouring. Five native signers of FinSL were recorded with 24 Hz frame rate during a map task. The task elicited free signing from a map with a starting point, a path, and an end-point. A more detailed description of the gathering procedure and the data can be found in [18]: pp. 77-78, 85-86. Here, a subset of the data is further analysed.

2.2. Sign flow segmentation

The sign flow of the dominant hand was segmented and labelled. The hand a signer uses to sign one-handed signs is called the dominant hand. Most people use their right hand but there are left-handed signers as well. A labelled sign was compared to the dictionary form. Thereafter, the sign was classified either as a function or a content sign. The data contained a total of 335 signs out of which 95 were function signs (28%).

The signs were divided into smaller constituents. We used a phonetic definition of a syllable although there is no generally accepted definition for a signed syllable. The signs were divided into syllables according to the following rules:

1. Each sign consists of at least one syllable.
2. If the sign can be decomposed to temporal segments each constituting an independent sign, then the sign is called a compound sign with each component being composed of at least one syllable; otherwise the sign is simple.
3. If the movement in the handshape flow is cyclic inside a simple sign or compound sign component then each cycle contains at least one syllable.
4. Every signed segment of an utterance belongs to at least one syllable.

The syllable boundaries were defined following the definition of a syllable. The boundaries of the signs were always syllable boundaries and the beginning of the new cycle (even if it remained incomplete) defined a syllable boundary. Only when two consecutive signs overlapped, the syllables could overlap. This could be a result of the two hands articulating different signs simultaneously or when the hand dominance switched temporarily to the other hand.

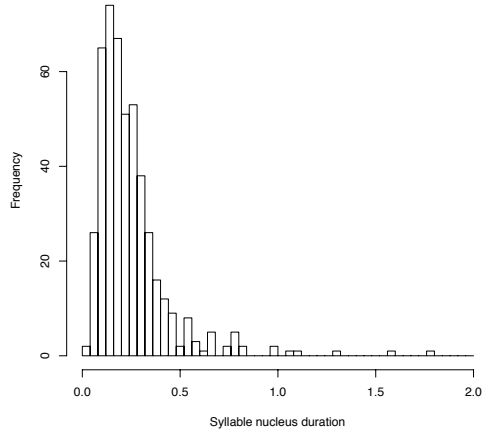


Figure 1: *A histogram of the syllable nucleus durations. The distribution is unimodal with mean 0.24 s, standard deviation 0.20 s, and median 0.21 s.*

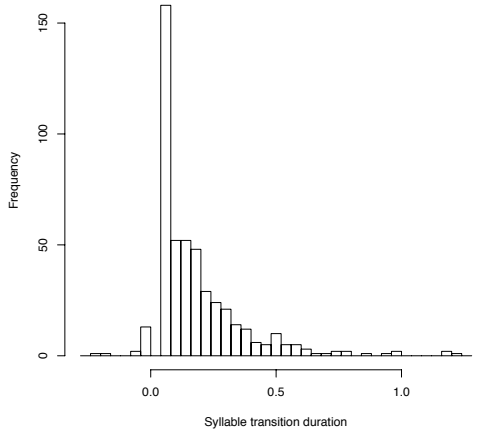


Figure 2: *A histogram of the syllable transition durations. The distribution is unimodal with mean 0.17 s, standard deviation 0.19 s, and median 0.13 s.*

The monosyllabic signs were most common (235 signs; 70% of the signs). In addition, there were 70 disyllabic signs (21%), 22 signs with three syllables (7%), 7 signs with four syllables (2%), and 1 sign with five syllables (<1%) The mean duration of a syllable was 0.41 s with standard deviation 0.30 s.

Every (phonetic) syllable was then further divided into a nucleus and a transition. The onset of the syllable nucleus was defined as the beginning of the syllable. The end point of the syllable nucleus, i.e. the beginning of the transition, was defined as the best match with the end handshape or secondarily the end position of the hand compared to the dictionary form of the syllable, and thirdly, as the first moment that clearly belonged to the movement towards the next syllable, i.e. the transition. The syllable nucleus durations are shown in Figure 1. The mean duration was 0.24 s and the standard deviation was 0.20 s.

The segments between consecutive nuclei were transitions (with one transition occurring between the nuclei – possibly with zero duration). Sometimes no transition took place (13 cases in the dataset). This was the case in polysyllabic signs consisting of a reduplicated simple sign like *sauna*. The transition could be even negative (4 cases in the dataset) when the end of a sign was reached (e.g. a hand-body contact ending a sign) after the handshape already had transformed to the beginning of the following sign. A histogram of the transition durations is shown in Figure 2. The mean duration of a transition was 0.17 s with standard deviation 0.19 s. The transition were in average somewhat shorter than the nuclei.

2.3. Statistical analyses

A mixed effects linear regression model was used to describe the variation in the duration of the syllable nuclei in the data set. The phonetic fixed variables were the duration of the adjacent transients and syllable nuclei; phonological fixed variables were the syllabic length of the sign, the syllable position within the sign (1 for the first syllable, 2 for the second, etc.), and the sign type (function sign or content bearing). The subjects were treated as random variables in the model.

3. Results

The minimal mixed effects model used to explain the syllable nucleus durations, which did not significantly differ from the full model with all the main effects (ANOVA), consisted of the adjacent syllable nuclei durations, the duration of the post-nuclear transition, and the type of the sign (function vs. content sign). The minimal model is described in Table 1. Syllable nuclei within a function sign were produced in average 56 ms faster in comparison to the syllable nuclei within a content sign. The overall signing rate was reflected by the strong impact of the previous and following syllable nucleus durations. Somewhat surprisingly then, only the post-nuclear transition duration had an impact on the nucleus duration (and not the pre-nuclear transition).

Although the syllable structure was taken into account by including the number of the syllables and position within a sign as fixed factors, this did not have any significant effect in explaining the nucleus durations. The phonological factors, syllable position and the syllabic length of the sign, did not explain the variation in the data, and were dropped from the minimal model.

Table 1: *The minimal mixed effects model fitted to the data to explain the variation in syllable nucleus durations.*

Fixed effect	Estimate	t value
Intercept (s)	0.12	7.1
Type (function sign)	-0.056	-2.9
Duration of the post-nuclear transition	0.21	4.6
Duration of the following nucleus	0.19	4.3
Duration of the previous nucleus	0.20	4.7

4. Discussion

Lexical factors affected the duration of the syllable nuclei of the signs even when the duration of the transitions were controlled for. The function signs were shorter possibly because of

faster retrieval from the mental lexicon or more fluent production. This could be either because of the special syntactic role function words have in signing or it could be a consequence of function signs being more frequent in average than the content signs. As sign frequency is also a property of the sign, the results suggest that lexical properties influence the sign durations. In the current data, no attempt was made to control the sign frequency. Most likely the function signs in the current data set were more frequency than the content signs, which could have led to a confound. It is fully possible that the sign frequency *alone* would be sufficient to explain the sign type effect were it so that more frequent signs are produced faster *and* the function signs in the data set were more frequent. Unfortunately, the sign frequencies are not readily available for FinSL, yet.

The impact of adjacent constituents to the syllable nucleus durations is highly expected and it reflects to the local signing rate. In the results, there was an asymmetry in the influence of the transitions: only the post-nuclear transition had an significant effect on the nucleus duration while the effect of prenuclear transition was only half in size and did not reach significance (ANOVA for dropping the term, $p > 0.1$). The statistical model fitted to the current data set suggests the following asymmetry: for transitions that last longer, the preceding nucleus would be lengthened but not the nucleus following the transition. This could be a consequence of a trade-off between available time and necessary articulatory effort: for a longer distance between two consecutive nuclei more time could be allocated which could be compensated for by reduced articulation. Hence, the salience of the sign would be intact. Moreover, if a long distance within a transition is performed faster this would result in hyperarticulation. This trade-off then would move the boundary between the sign nucleus and the post-nuclear transition earlier during hyperarticulation. Importantly, the boundary is defined through the handshape which can be controlled rather independently compared to the position of the hand [22], [23].

Compared to an earlier analysis of the *syllable durations* by the current authors [10], the syllable nucleus durations show somewhat different behaviour. The syllable durations alone were shorter in polysyllabic signs than in monosyllables but here the syllable nuclei did not significantly differ as a function of the number of syllables. Also, function sign syllables were not significantly shorter than content signs although here the nuclei durations varied according to the sign type.

These opposing results can be reconciled. The bisyllabic signs often consist of reduplicating a monosyllabic sign leading to short (or non-existant) transitions. Indeed, in the current data, the sign internal transition are much shorter (mean 0.054 s) than the transitions between the signs (mean 0.22 s; Wilcoxon signed rank test, $p < 10^{-15}$). The transition durations themselves probably depend primarily on the spatial distance between the articulatory locations of the syllables, which is smaller sign internally. This remarkably shortens the duration of the syllables in polysyllabic signs. In addition to the evidence from the analysis of the last section, a direct comparison of the first syllable nucleus durations of mono vs. polysyllabic signs reveals no significant difference (Wilcoxon, $p > 0.1$).

In the current analysis of syllable nucleus durations, a relatively small effect of the sign type emerged although such an effect was not visible in the earlier syllable duration analyses. That the effect was not earlier visible could be simply due to larger variances in the syllable durations as opposed to the nucleus durations. In addition, it could be that only the nucleus durations are affected by the sign type and the duration of the transitions is solely determined by the spatial relations of the adjacent syllables.

The distributions of the durations for syllables, syllable nuclei, and syllable transitions were all non-normal. The logarithms of the respective durations show less non-normal shapes but are still clearly (QQ-plots, normality tests) different from normal. To diminish the possibility that the results would be an artifact of the non-normality of the data, we built a minimal mixed effect model for the logarithmically transformed durations. The resulting minimal model has exactly the same terms and the statistical significance of the individual factors is improved. This hints that the non-normality is not the main source of the results.

During spoken word production, frequent words tend to be produced faster when controlled for phonological length and the frequency of the word [24]. This is true for signs as well (at least what comes to the phonological length) as the syllables in the polysyllabic signs are faster [10]. The current work suggests that this phenomenon might be completely explained in the sign language context by the spatial relations of the sign components (syllables). An alternative explanation that the phonologically longer signs are produced faster because of repetition induced optimization of the syllable sequences within a sign as opposed to across sign boundaries, is not supported by the current sign language data.

5. Conclusion

In the current work we demonstrated using Finnish Sign Language that even in a sign language lexical factors affect the duration of signed segments. In addition, the syllable nucleus durations are affected on the phonetic level by the durations of adjacent nuclei and the postnuclear transition but not by the prenuclear transition. Hence, the current data analysis suggests that the transitions are more strongly attached to the signs (syllables) preceding the transitions. In addition, function signs tended to have faster syllable nuclei in the data set similar to the spoken language results. This points to a sign production model where the spatial relation of the signs (and sign syllables) determines the transition durations which in turn affect the nucleus durations alongside with the lexical properties of the sign.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement n° 287678. The authors would like to thank the participants of the study.

7. References

[1] M. Nespor and W. Sandler, "Prosody in Israeli sign language," *Language and Speech*, vol. 42, no. 2-3, pp. 143–176, 1999.

[2] R. B. Wilbur, "Stress in ASL: Empirical Evidence and Linguistic Issues," *Language and Speech*, vol. 42, no. 2-3, pp. 229–250, 1999.

[3] R. B. Wilbur and S. B. Nolen, "The duration of syllables in American Sign Language," *Language and speech*, vol. 29, no. 3, pp. 263–280, 1986.

[4] D. Brentari, *A prosodic model of sign language phonology*. The MIT Press, 1998.

[5] S. K. Liddell and R. E. Johnson, "American Sign Language: The Phonological Base." *Sign language studies*, vol. 64, pp. 195–278, 1989.

[6] B. J. Kröger, P. Birkholz, J. Kannampuzha, E. Kaufmann, and I. Mittelberg, "Movements and holds in fluent sentence production of Aerican Sign Language: The action-based approach," *Cognitive computation*, vol. 3, no. 3, pp. 449–465, 2011.

[7] M. Blondel and C. Miller, "Movement and rhythm in nursery rhymes in LSF," *Sign Language Studies*, vol. 2, no. 1, pp. 24–61, 2001.

[8] S. Ojala, "Rytmin ja koartikulaation vaikutus viittomiin ja puheeseen. The effect of rhythm and coarticulation on sign and speech." *Jantunen, Tommi (ed.) Näkökulmia viittomaan ja viittomistoon. Jyväskylä: Jyväskylän yliopisto*, pp. 29–43, 2010.

[9] R. Sutton-Spence and B. Woll, *The linguistics of British Sign Language: an introduction*. Cambridge University Press, 1999.

[10] D. Aalto and S. Ojala, "The duration of syllables in spontaneous Finnish Sign Language," in *Nordic Prosody, Proceedings of the XIth Conference, Tartu, 2012*, E.-L. Asu and P. Lippus, Eds. Peter Lang, 2012, pp. 39–47.

[11] S. Ojala, T. Salakoski, and O. Aaltonen, "Coarticulation in sign and speech," in *actes de NODALIDA 2009 workshop Multimodal Communication, from Human Behaviour to Computational Models, Costanza Navarretta, Patrizia Paggio, Jens Allwood, Elisabeth Alsén and Yasuhiro Katagiri (Eds). NEALT Proceedings Series*, vol. 6, 2009, pp. 21–24.

[12] S. K. Liddell, *American sign language syntax*. Mouton The Hague, 1980.

[13] R. E. Johnson and S. K. Liddell, "A segmental framework for representing signs phonetically," *Sign Language Studies*, vol. 11, no. 3, pp. 408–463, 2011.

[14] S. Kita, I. Van Gijn, and H. Van der Hulst, "Movement phases in signs and co-speech gestures, and their transcription by human coders," in *Gesture and sign language in human-computer interaction*. Springer, 1998, pp. 23–35.

[15] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[16] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[17] E. R. Kandel, J. H. Schwartz, T. M. Jessell *et al.*, *Principles of neural science*. McGraw-Hill New York, 2000, vol. 4.

[18] S. Ojala, "Towards an Integrative Information Society: Studies on Individuality in Speech and Sign," Ph.D. dissertation, University of Turku, 2011.

[19] B. Lindblom, C. Mauk, and S.-J. Moon, "Dynamic specification in the production of speech and sign," *Dynamics of Speech Production and Perception*, vol. 374, p. 7, 2006.

[20] B. Fuchs, "Phonetische Aspekte einer Didaktik der Finnischen Gebärdensprache als Fremdsprache," Ph.D. dissertation, University of Jyväskylä, 2004.

[21] S. Ojala, "Rytmi puheessa ja viittomisessa," *XXVI Fonetiikan päivät 2010*, pp. 47–49, 2010.

[22] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*. Springer, 1990, pp. 403–439.

[23] ——, *Economy of speech gestures*. Springer, 1983.

[24] A. Bell, J. M. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

Pause insertion prediction using evaluation model of perceptual pause insertion naturalness

Hiroko Muto, Yusuke Ijima, Noboru Miyazaki, Hideyuki Mizuno

NTT Media Intelligence Laboratories, NTT Corporation, Japan

Abstract

This paper describes a pause insertion prediction approach for generating more natural synthesized speech for Text-to-Speech (TTS) synthesis systems. A novel point of the proposed approach is the use of an evaluation model of perceptual pause insertion naturalness in addition to a prediction model based on machine learning. The evaluation model represents the relationship between several features related to pause insertion and the perceptual pause insertion naturalness obtained in a subjective evaluation. First, using a prediction model based on machine learning, we obtain the N-best sequences that indicate whether or not a pause is present at each phrase boundary. We then estimate pause insertion naturalness scores for each N-best sequence using the evaluation model and select the sequence with the highest naturalness score. Objective and subjective evaluation results show that the proposed approach gives better results than a conventional approach.

Index Terms: pause insertion prediction, text-to-speech synthesis, perceptual naturalness, machine learning

1. Introduction

A key to synthesizing speech with improved naturalness is (silent) pause insertion prediction at essential steps in the text-to-speech synthesis process. This is because pause insertions are important for understanding speech. Pause insertions are also important for expressing the characteristics of specific speaking styles. Wang [1] reported that the pause locations of read, expressively read, and spontaneous expressively read speech are different from each other and that irregular pause insertions can improve the expressiveness of synthesized speech. Parlikar [2] reported that natural synthetic speech for the target speaking style can be generated by using a speaking style-dependent pause insertion model. These reports indicate that one needs to be able to predict pause insertions which are suitable for the target speaking style in order to generate expressive synthetic speech.

Machine learning is a suitable way to generate expressive synthetic speech, because it can automatically train the characteristics of pause insertions from training data in the target speaking style. Various machine learning-based approaches have been proposed for predicting pause insertions, such as decision trees [3, 4], Hidden Markov Models (HMMs) [5], Maximum entropy models [6], and Conditional Random Fields (CRFs) [7, 8]. In these studies, various linguistic features, which are extracted from training data, are used for prediction and contribute the prediction performance to improve. On the other hand, it is known that the characteristics of pause insertions have wide variances [9]. Even if the linguistic features are similar, the pause locations have many variations. This indicates it may be difficult to predict more accurate pause insertions by the prediction models trained using only linguistic features.

One useful approach to alleviating this problem would be utilizing not only machine learning-based decisions but also general measures that are independent of specified training data. For example, Kim [10] utilized grammatical rules for statistical pause prediction and showed their effectiveness in reading speaking style. However, this approach would be difficult to apply to a specific speaking style that does not strictly adhere to grammatical rules. To generate expressive synthetic speech, a new measure is needed that can express the characteristics of the target speaking style.

In this paper, we focus on naturalness scores obtained from subjective evaluation. Humans can intuitively judge the naturalness of speech regardless of speaking style. By revealing the features that influence to perceptual pause insertion naturalness and usages of pause insertion prediction, it would be improving the pause insertion performance for specific speaking style. A related study [11] indicates that a perceptual accent naturalness measure obtained from a subjective evaluation is effective at speech segment selection for concatenative speech synthesis. The task of pause insertion prediction would be able to predict more accurate pauses by using perceptual pause insertion naturalness as well as the speech segment selection.

To investigate the characteristics of pause insertions that influence the naturalness of speech, we performed multiple regression analysis to analyze the relationship between several features that related to pause insertions and the naturalness scores obtained by a subjective evaluation using synthesized speech with different pauses. Then, based on the analysis, we propose a pause insertion prediction approach that utilizes the obtained characteristics for machine learning-based pause insertion prediction. We constructed a multiple regression model on the basis of the analysis and used it to evaluate the naturalness of the pause insertions predicted by machine learning. In the proposed approach, we first predict N-best hypotheses of the pause insertion prediction result by using a CRF-based pause insertion prediction. Then, we calculate the naturalness score using the multiple regression model for each N-best hypothesis and rerank them. Finally, the hypothesis with the highest naturalness score is selected from the N-best hypotheses. We conducted a preliminary evaluation of our approach on a reading speaking style corpus and obtained encouraging results.

2. Evaluation of perceptual pause insertion naturalness

In this section, we describe the details of the subjective evaluation using synthesized speech with different pauses and multiple regression analysis to analyze the relationship between several features that related to pause insertions and the naturalness scores obtained by subjective evaluation.

2.1. Subjective evaluation

2.1.1. Speech stimuli

For the subjective evaluation, we used 400 pattern synthesized speech stimuli with different sentences and pause insertion patterns. As a sentence set, 50 sentences were selected from 503 phonetically balanced sentences in the ATR Japanese speech database (Set B). Eight pause insertion patterns were used for each sentence. We did not evaluate all possible pause insertion patterns for each sentence, because it is useless to evaluate patterns humans rarely use. To evaluate patterns that humans use, we utilized patterns from the utterances of several speakers. To evaluate various kinds of pause insertion patterns, we selected speakers whose pause locations are widely distributed. In addition to them, we used the patterns predicted by a rule-based pause insertion predictor [12] to compare with the other patterns. For these purposes, we selected seven speakers and one pause insertion predictor as follows.

- One professional speaker : the pause insertion pattern was used as a general pattern.
- Six non-professional speakers : the average number of their pauses differed significantly. We used them to keep the variation of the pause insertion pattern.
- One Japanese pause insertion predictor : pause insertion pattern was predicted by a manually designed rule.

To evaluate the effects of pause insertion on naturalness, it is desirable to use speech stimuli with the same voice quality and prosody (F0 contour, phoneme duration, accent phrase boundary, and accent type). For this reason, all speech stimuli were generated by a speech synthesizer [13].

2.1.2. Experimental conditions

We had 16 subjects (seven males, nine females) listen to the 400 speech stimuli in random order and rate their naturalness of pause insertion using a 5 point scale (from 1: very unnatural, to 5: very natural). The subjects evaluated each stimulus twice. For each stimulus, the average value of the 16 subjects' scores was used as the naturalness score.

2.1.3. Experimental results

Figure 1 shows the histogram of the naturalness score for each stimulus. 96.5% of the sample scores are over 3 point (fair) because most of speech stimuli employed pause insertion patterns extracted from real utterances, while, the naturalness scores of all sample are not always more than 4 point (natural) and broadly distributed from 2.5 to 4.5. In comparison of the average score for each speaker, the patterns of the professional narrator was the highest and that of the pause insertion predictor was the lowest.

2.2. Multiple Regression analysis

2.2.1. pause distribution features used in analysis

In the multiple regression analysis, we used the following 10 features related to pause insertion, which are called "pause distribution features".

(a) The average and variance values of the number of mora in pause phrases

(b) The average and variance values of the number of content words in pause phrases

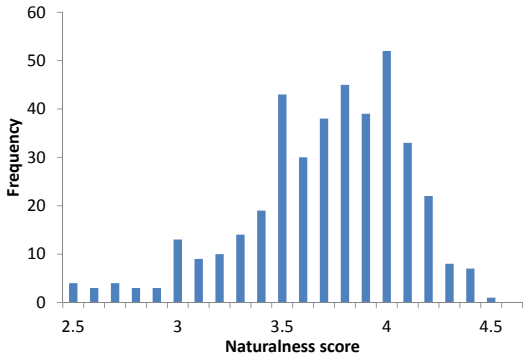


Figure 1: *Naturalness score histogram.*

(c) The existence of four outliers in the number of mora in pause phrases $(\mu - \sigma,\ \mu - 2\sigma,\ \mu + \sigma,\ \mu + 2\sigma)$

(d) The number of pauses inserted between phrases having a dependency relation

(e) The number of non-pauses inserted between phrases not having a dependency relation

These features are calculated for each sentence. A mora is a syllablesized unit in Japanese. "Pause phrase" is one or more consecutive phrases delimited by pauses. (a) and (b) are basic features related to the length and amount of information of each pause phrase. (c) is four features related to outliers. They indicate whether pause phrases that are too long or too short exist. Generally, if such pause phrases exist, the naturalness would become worse. These features can be set to a value of 1 or 0. It is set to 1 if a pause phrase exists in which the number of mora is outside the standard value. In other cases, it is set to 0. The average value μ and the standard deviation value σ, used to determine the standard value, were calculated for all 400 sentences. (d) and (e) are the features related to the relationship of modification between phrases. A previous study [14] showed that these two features affect the naturalness of pause insertions.

2.2.2. Multiple regression analysis results

We performed multiple regression analysis to model the relationship between the pause distribution features given above and the naturalness scores obtained in the subjective evaluation. To investigate the relationship of these features and the naturalness scores, we first calculated a multiple correlation coefficient. We confirmed that the naturalness scores and the estimated ones were correlated; the multiple correlation coefficient was 0.61. We also investigate the effect of each pause distribution feature on the naturalness scores. Partial correlation coefficient for each feature is shown in Table 1. From this table we can see that the features (a) (especially average value), (c) (especially $\mu - 2\sigma$), (d), and (e) were highly correlated for the naturalness score.

3. Proposed approach

By the analysis, we obtain the features that influence to the perceptual pause insertion naturalness. To evaluate the effectiveness of the features for pause insertion prediction, we developed a pause insertion prediction approach that combines the information obtained by the analysis and machine learning-based

Table 1: *Partial correlation coefficients (PCC) for each pause distribution feature.*

pause distribution feature	PCC
Average value of #mora	-0.21
Variance value of #mora	-0.10
Average value of #contents word	-0.03
Variance value of #contents word	-0.01
Outlier ($\mu - \sigma$)	-0.08
Outlier ($\mu - 2\sigma$)	-0.23
Outlier ($\mu + \sigma$)	-0.02
Outlier ($\mu + 2\sigma$)	-0.06
#pauses inserted between dependency relation phrases	-0.42
#non-pauses inserted between non dependency relation phrases	-0.24

pause insertion prediction. We constructed a multiple regression model on the basis of the analysis and used it to evaluate the naturalness of the pause insertions predicted by machine learning. We call it the "evaluation model". A block diagram of the proposed approach is shown in Fig. 2. It consists of two stages: a prediction stage and an evaluation stage.

In the prediction stage, we predict a Boolean sequence that indicates whether a pause has to be inserted after each phrase within the input phrase sequence. (We call this Boolean sequence a " pause assignment sequence"). A CRF-based pause insertion prediction model is used for prediction and N-best pause assignment sequences are output.

In the evaluation stage, for each N-best sequence, we extract pause distribution features. We then estimate the naturalness score using the extracted pause distribution features and an evaluation model of perceptual pause insertion naturalness. Finally, we select the pause assignment sequence with the highest naturalness score among the N-best sequences.

Although pause insertions were mainly predicted by word units in previous studies, however, in this study, we predict them by phrase units (also called " bunsetu" [15]) because pauses are usually inserted into phrase boundaries.

4. Pause insertion prediction model

4.1. Pause insertion modeling with CRF

In this subsection we describe the pause insertion prediction model. Whether a pause should be inserted after a phrase greatly depends not only on the various linguistic features of the phrase but also on what comes before and after the pause. Machine learning approaches may be effective for modeling such comprehensive and complex features. In this study we used the CRF, which is an effective machine learning approach for sequential labeling. Given a sequence of vectors of the linguistic features $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_M)$ with M phrases, the goal is to find a highest probability sequence of pause assignment labels $\boldsymbol{y} = (y_1, \cdots, y_M)$ for M junctures after every phrase. The linguistic features of sequence $\boldsymbol{x}$ are described in Sect. 4.2. Each y_i, the i-th pause assignment label in sequence $\boldsymbol{y}$, can be 1 or 0. If a pause is inserted immediately after the i-th phrase, y_i is 1. In other case, y_i is 0.

In this study we used the linear chain CRF, a special form of CRF. The conditional probability $P(\boldsymbol{y}|\boldsymbol{x})$ is calculated as

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{y}} \exp(\boldsymbol{w}, \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y}))}$$

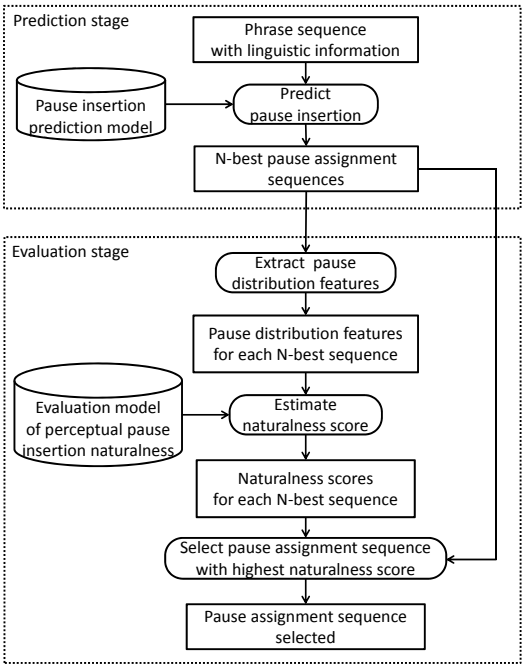


Figure 2: *Overview of proposed approach.*

where $\boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{y})$ is a feature function and $\boldsymbol{w}$ is a parameter to be estimated from training data. CRF are usually trained by maximizing the log-likelihood over a given training set. In this study we obtained the N-best sequences in order of the probability of $P(\boldsymbol{y}|\boldsymbol{x})$.

4.2. Linguistic features for pause insertion modeling

These linguistic features were used for pause insertion modeling .

(1) Text of current phrase
(2) Syntactic category of current phrase
(3) Flag of whether current phrase modifies next phrase
(4) Text of last word of current phrase
(5) Part of speech (POS) of last word of current phrase
(6) Pronunciation of last word of current phrase
(7) Text of first word of next phrase
(8) POS of first word of next phrase
(9) Pronunciation of first word of next phrase
(10) Above features' quin-context and combinations of them

In this study, we modified the word-based features used in conventional approaches [7, 8] to suitable ones for the prediction by phrase units. We also used the features that indicate the dependency relationships between phrases.

5. Evaluation

5.1. Experimental conditions

In the following experiment, a Japanese news corpus was used. It contained 1000 Japanese news sentences uttered by one professional male narrator, with an average of 70.0 words per sentence, 23.4 phrases and 11.7 pauses. The POS, pronunciation,

Table 2: *The results of pause insertion prediction.*

	Recall	Precision	F-measure
Proposed	93.3	72.4	81.5
Conventional	93.6	68.4	79.0

and dependency relation were labeled automatically using a Japanese morphological analyzer [16] and a Japanese dependency parser [17]. Pause locations were labeled manually according to the speech data.

As a conventional approach, the 1-best result predicted by CRF-based pause insertion prediction model was used. CRF++ toolkit [18] was used for pause insertion prediction model training. We set the parameter N (the number of N-best sequences) as 10 from the preliminary experiment results.

5.2. Objective evaluation

We first compared the prediction performance of the proposed approach with that of the conventional approach. From the corpus, 100 sentences were used as the evaluation data and the rest were used as training data for the pause insertion prediction model. We performed a 10-fold cross validation test. The prediction performance was evaluated by precision, recall, and F-measure. Where, precision means the ratio of correctly inserted pauses to the total number of inserted pauses, and recall means the ratio of correctly inserted pauses to the total number of pauses in the evaluation data.

The results are shown in Table 2; the proposed approach had slightly lower recall but higher precision and F-measure than the conventional approach.

5.3. Subjective evaluation

We also compared the naturalness of synthesized speech generated from the results predicted by each approach by a pair comparison test. Since the sentences in the corpus were too long to evaluate subjectively, we truncated them to short sentences that can be listened to easily. From the corpus we selected 30 sentences for evaluation and truncated them to 148 short sentences, averaging eight seconds in length. The rest were used for training the pause insertion prediction model. We randomly selected 30 short sentences from the short sentences in which pause location differences were found between results predicted by the proposed and the conventional approaches. The synthesized speech was generated by the same speech synthesizer used for the subjective evaluation described in Sect. 2.1. The pause durations in the synthesized speech were consistently 0.5 seconds. Subjects were 24 persons (12 males, 12 females), and presented a pair of synthesized speech samples in random order and then asked which samples had better naturalness.

The preference scores obtained for the experiment (Fig. 3) show the proposed approach outperformed the conventional approach. This shows that using an evaluation model of perceptual pause insertion naturalness is an effective way to improve pause insertion naturalness in synthesized speech.

We conducted the objective evaluation using the short sentences used for the subjective evaluation. The prediction performance was shown in Table 3. From this table we can see that the prediction performance was almost the same as those using the original sentences.

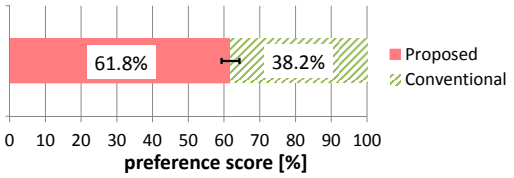


Figure 3: *Preference score from the subjective evaluation. (Error bars show the 95% confidence intervals.)*

Table 3: *The results of pause insertion prediction. (Using short sentences used for subjective evaluation)*

	Recall	Precision	F-measure
Proposed	94.1	74.4	83.1
Conventional	95.2	72.5	82.3

6. Conclusion

In this paper, we proposed a pause insertion prediction approach involving the use of an evaluation model of perceptual pause insertion naturalness for generating natural synthesized speech. Objective and subjective evaluations demonstrated the proposed approach provides better results than the conventional approach. This shows that consideration of perceptual pause insertion naturalness to pause insertion prediction is an effective way to improve naturalness in synthesized speech.

Currently the proposed approach's effectiveness has been demonstrated only for the reading speaking style. A future task is to explore its effectiveness for other speaking styles. We also intend to analyze pause duration naturalness and attempt to develop a pause duration estimation approach for generating more natural synthesized speech.

7. References

[1] X. Wang, A. LI, C. Yuan, " A Preliminary Study on Silent Pauses in Mandarin Expressive Speech," Speech Prosody, pp. 673–676, 2008.

[2] Parlikar. A and A.W. Black, " A Grammar Based Approach to Style Specific Phrase Prediction," INTERSPEECH, pp. 2149–2152, 2011.

[3] M. Q. Wang and J. Hirschberg, " Automatic classification of intonational phrase boundaries," Computer Speech and Language, vol. 6, pp. 175–196, 1992.

[4] P. Koehn, S. Abney, J. Hirschberg and M. Collins, " Improving intonational phrasing with syntactic information," ICASSP, pp. 1289–1290, 2000.

[5] P. Taylor and A. W. Black, " Assigning phrase breaks from part of speech sequences," Computer Speech and Language, Vol. 12, pp. 99–117, 1998.

[6] J. Li, G. Hu and R. Wang, " Chinese prosody phrase break prediction based on maximum entropy model," INTERSPEECH, pp. 729–732, 2004.

[7] Y. Qian, Z. Wu, X. Ma and F. Soong, " Automatic prosody prediction and detection with Conditional Random Field (CRF) models," ISCSLP, pp. 135–138, 2010.

[8] J. Sun, J. Yang, J. Zhang and Y. Yan, " Chinese prosody structure prediction based on Conditional Random Fields," 5th International Conference on Natural Computation, pp. 602–606, 2009.

[9] H. Fujisaki, S. Ohno and S. Yamada, " Analysis of occurrence of pauses and their durations in Japanese text reading," ICSLP, Vol. 4, pp. 1387–1390, 1998.

[10] B. Kim and G. Lee. ”Statistical/Rulebased Hybrid Phrase Break Detection,” ICSP, 1999.

[11] A. Yoshida, H. Mizuno and K. Mano, “Segment selection method based on tonal validity evaluation using machine learning for concatenative speech synthesis,” ICASSP, pp. 4617–4620, 2008.

[12] K. Matsuoka, E. Takeishi and H. Asano, “ AUDIOTEX. A Text-To-Speech System for Japanese Text”, 52nd National Convention of IPSJ, Vol. 2, pp. 409–410, 1996. (in Japanese)

[13] K. Mano, H. Mizuno, H. Nakajima, N. Miyazaki and A. Yoshida, “ Cralinet—Text-To-Speech System Providing Natural Voice Responses to Customers,” NTT Technical Review, Vol. 18, No. 11, pp. 19–22, 2006.

[14] N. Kaiki and Y. Sagisaka, “ Study of pause insertion rules based on local phrase dependency structure,” IEICE, Vol. J79-D-II, No. 9, pp. 1455–1463, 1996.

[15] Y. Zhang and K. Ozeki, “ Automatic bunsetsu segmentation of Japanese sentences using a classification tree,” Language Information and Computation, pp. 230-235, 1998.

[16] K. Imamura, K. Saito and H. Asano, “ Basic Japanese text analysis technology as a platform for Knowledge extraction,” NTT Technical Review, Vol. 6, No. 9, 2008.

[17] K. Imamura, G. Kikui and N. Yasuda, “ Japanese Dependency Parsing Using Sequential Labeling for Semi-spoken Language,” ACL, pp. 225–228, 2007.

[18] CRF++: Yet Another CRF toolkit , http://crfpp.sourceforge.net/.

Non-native perception of final boundary tones in French interrogatives

Fabián Santiago[1], *Paolo Mairano*[2], *Elisabeth Delais-Roussarie*[1]

[1] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris Diderot, France
[2] University of Turin, Italy

rotinet@hotmail.com, paolomairano@gmail.com, elisabeth.roussarie@wanadoo.fr

Abstract

The aim of the paper is to present the results of a perception experiment in which native and non-native listeners were asked to rate the appropriateness of resynthesized questions varying in respect to two aspects: their morphosyntactic structure (presence/absence of an interrogative marker) and the form of their final intonational contour (falling, rising and extra-rising). The goal of the experiment was to examine how non-native listeners of French did perceive the extra-rising final contour that was observed in learners' productions. Do they consider it as appropriate even if it did not occur often in the native speakers' productions? By and large, the results of the experiment show that native listeners preferred rising contours over falling ones in all question types, whereas non-native listeners rated the extra rising contours higher than French natives in stimuli having a morphosyntactic structure that differs from the one used in their L1. These results may suggest that rising contours represent a default tonal form associated with the interrogative modality not only at the beginning of the L2 acquisition process, but also in speakers' mental representation, irrespective of their L1.

Index Terms: L2 intonation, L2 acquisition,

1. Introduction

It is often assumed that the prosodic patterns observed in the productions of second/foreign language learners are mainly influenced by their L1 prosodic structure. L1 transfer is thus considered as playing a crucial role (see [1] and [2] for an overview). However, several studies have shown that, in some cases, the L1 transfer cannot be invoked to account for some observed patterns. Other factors have thus to be taken into consideration to explain singular prosodic forms of the L2 learners' grammar [3]. In previous work, for instance, we have shown that Mexican Spanish learners of L2 French produce systematically an extra-rising contour at the end of information-seeking questions (see [4] and [5]). Even if the occurrence of this form can be considered as resulting from a L1 transfer in yes-no questions, it is not possible in the case of wh-questions (see section 2.3 for more information). The overuse of this contour could thus be seen as resulting from the L2 acquisition process itself. Such explanation may have consequences on the perception, this form being not necessarily considered as accurate in French. Do the learners consider it as appropriate at various early stages of the L2 acquisition process, even if the form is not often heard?

This article is organized as follows. In section 2, we provide background information on the prosody of French and Spanish interrogatives. Section 3 focuses on the experimental protocol of the perception experiment. In section 4, we present the results obtained. Finally, the implications of the results for the theories focusing on the acquisition of intonation in an L2 are summarized in the last section.

2. Intonation in questions in French and Spanish: L1 vs. L2

In descriptions of French and Spanish intonation, yes-no questions are usually characterized by ending with a rising contours (H%), while wh-questions are associated with falling contours (L%). Nevertheless, different tonal configurations can be observed in both interrogative sentence types: falling tones could appear in French yes-no questions when the lexical content and/or the morphosyntactic structure already indicates the modality of the utterance, whereas rising patterns can appear in both Spanish and French wh-questions. In the following sub-sections, we summarize the distribution of these contours with respect to the morphosyntactic structure and to the canonical patterns usually reported in the literature.

2.1. The intonation of yes-no questions

Information-seeking yes-no questions do differ in Spanish and French with respect to their morphosyntactic forms. In Spanish yes-no questions, two morphosyntactic structures are observed with nominal subjects: either the subject precedes the verb as in an assertive sentence (1), or the subject and the verb are inverted as in (2). In contrast, French yes-no questions can be produced with a larger range of linguistic forms: in addition to the declarative structure (3), a subject-verb inversion can occur, be the subject nominal or pronominal (4), and the interrogative marker 'est-ce que' can be inserted at the beginning of the utterance (5).

(1) *¿Juan estudia?* 'Juan studies?'
(2) *¿Estudia Juan?* 'Studies Juan ?'
(3) *Vous étudiez ?* 'You study?'
(4) a. *Étudiez-vous ?* 'Study you?'
b. *Pierre étudie-t-il ?* 'Does Pierre study?'
(5) *Est-ce que vous étudiez ?* 'You study?'

As for intonational patterns, Mexican Spanish yes-no questions are produced with a rising contour regardless of the syntactic structure: a final rise, which covers approximately an octave (12 semitones), is realized from the penultimate syllable to the end of the utterance, and the final pitch target generally reaches the top of the speakers' range. This tonal pattern is perceived as an extra high rising pitch movement, and is more frequently observed in Mexico City Spanish (see [6], [7], and [8]).

In contrast, French speakers may use a larger inventory of melodic movements at the end of questions without changing the meaning of the utterance. In yes-no declarative questions (as in (3)), a rising pitch movement realized on the last syllable of utterance (or on the last accentual phrase) and covering in average 8 semitones is usually observed (see [9] and [10] among others). By contrast, in yes-no questions in which the interrogative modality (as in (4) or (5)) is expressed by a marker or the morphosyntactic form, the final contour can

be either falling L% or rising H%, the H% vs. L% contrast being in a certain extent neutralized.

In the remaining of the paper, the differences between the various tonal configurations will be noted as follows: the label H% is used for rising contours, the label HH% for extra rising one, and the label L% for falling one (see the representation in Fig. 1 in section 2.3).

2.2. The intonation of Wh-questions

The morphosyntax of wh-questions differs in Spanish and French with respect to the possible positions of the interrogative expression. Information-seeking wh-questions are always constructed with the interrogative word in utterance-initial position in Spanish (6a). By contrast, it is possible in French to utter a wh-question with the wh-expression in sentence-initial position (wh-'fronted') as in (6b), or with the expression in the position where the answered complement should occur (wh-'in-situ') as in (7):

(6) a. *¿Qué estudias?* 'What study (you)?'

b. *Qu'est-ce que tu études ?*

(7) c. *Tu études quoi ?* 'You study what?'

A cross-linguistic comparison of the tonal configurations occurring at the end of wh-questions in both French and Spanish shows that there are more similarities between both languages than in the case of yes-no questions: the final contour is often falling L% (see [6], [7], [8], [10], [11], [12] and [13]), but rising one could also be used. Note also that the extra rising contour HH% is almost inexistent in natives' oral productions in both Spanish and French in this question type (see [5], [8] and [10] among others).

2.3. Intonation of interrogatives in L2 French

In previous studies dedicated to the intonation of questions in L2 French by Mexican Spanish speakers, it has been shown that learners often use the extra rising contours HH% in yes-no and wh-questions (se [4], [5]). The fig. 1 gives a schematic representation of the contours observed at the end of the two question types –'Do you study?' for the yes-no question, and 'Where are you going?' for the wh-question– in the productions of Native French speakers (FL1), Spanish learners of French (FL2) and natives Spanish (SL1). The horizontal dotted line in the figure represents the top of the speaker's range.

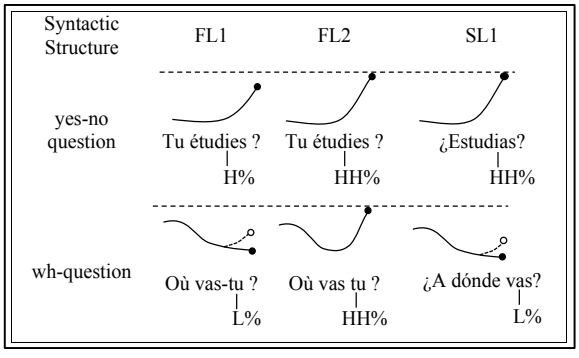


Figure 1: *Stylization of the three final boundary tones*.

From the data analyzed, [4] and [5] argued that the extra-rising tonal form HH% could be attributed to an L1 transfer in the cases of yes-no questions, since this form is quite common in the Mexican Spanish variety. Yet, for wh-questions, an analysis in terms of transfer cannot be argued for, since this question type is generally produced with a falling contour L% in Spanish. Hence, different explanations have to be found to account for the occurrence of this pattern. The first proposal we could argue for is that at an early stage of the L2 acquisition process speakers have at their disposal a reduced inventory of tonal forms, in which rising tunes (be they H% or HH%) are employed by default for expressing any kind of interrogative modality (see also [3]). The second one relies on a certain difficulty for learners to distinguish the difference between rising and extra-rising contours, since this difference is more phonetic in nature. The third one, which will not really be investigated here, assumes that learners, by being less confident, express linguistic insecurity about their performances by using such a form. The extra rising contour HH% would thus encode a lack of confidence in the L2 (see also [14]).

3. Perception experiment

In order to get some information on what could motivate the occurrence of the extra-rising contours, and to evaluate which of the first two hypotheses mentioned in section 2.3 is the most relevant, we set up a perception test. The learners had to evaluate which of the various contours (between HH%, H% and L%) is the more appropriate at the end of the different question types. If the use of extra rising contours at the end of questions is related to the speakers' L1 or to the acquisition process itself, this contour should be significantly perceived as appropriate, at least by learners at an early stage, in all settings. However, the more advanced learners should not rate extra-rising contours as accurate as beginners. If the difference between the two forms is phonetic in nature, the choice of one contour over the other should not necessarily be significant, at least for the learners. Results obtained in this experiment should thus help us to evaluate whether the preference for HH% observed in learners' oral productions is related to the L2 acquisition process, or not.

In addition, the other assumptions can be partly verified by referring to the use of the forms. Indeed, native and non-native listeners should display differences according to their L1 in the way of evaluating the different final boundary tones in relation to the morphosyntactic structure of the question. As final rises are usually HH% for Mexican native speakers (be they learners of French or not), they should rate HH% better than H% in yes-no declarative questions, whereas French native participants should rate H% better than HH%. In the case of yes-no questions, in which the interrogative modality is signaled by the morphosyntactic forms (as in (4) and (5)) as well as in wh-questions (as (6) and (7)), native listeners show a preference for H% and L% over HH%, which has rarely been observed in the data (see [4] and [5]). In the case of the L2 learners, we could expect to obtain the same results as for SL1 listeners, despite the fact that they use mostly the HH% contour. This expectation is based on the hypothesis that perception in an L2 is influenced by the participants' L1.

3.1. Experimental procedures and materials

Three classes of listeners were asked to participate to the perception test: native French speakers, native Mexican speakers, and Mexican learners of L2 French. The stimuli consisted of 96 resynthesized information-seeking questions (66 in French and 30 in Spanish). They were classified in four sets: (i) yes-no questions without any interrogative marker in

declarative form (as in (1) and (3)); (ii) yes-no questions with a syntactic marker indicating the modality of the utterance (French only, as in (5)); (iii) wh-'fronted' questions (as in (6)); and (iv) wh-'in-situ' questions (French only, as in (7)). The questions contained in set (i) displayed two different final rising contours: HH% and H%. Interrogatives in sets (ii), (iii) and (iv) had three different final tones: HH%, H% and L%.

Two native phoneticians (one for either language) recorded stimuli at the Laboratory of Linguistics from the University of Paris Diderot. At a first step, we obtained a stylization of the entire F0 trace associated with the various questions. At a second one, the final contours were manipulated in *Praat* in order to obtain perfectly coherent realizations. The manipulation was achieved along the following guidelines:

1. The HH% contour was obtained from the H% final rise in the natural stimuli, which were manipulated so that the rise would span >11 semitones.
2. The H% contour was also manipulated with *Praat* in order to enhance comparability between stimuli (we wanted to avoid mixing natural and manipulated stimuli): rises of 8 semitones were decreased to 6, while rises of 6 semitones were increased to 8 st.
3. The L% contours were extracted from phoneticians' realizations and manipulated into low flat plateaus.

In figure 2, we illustrate the manipulations that were carried out in the experiment:

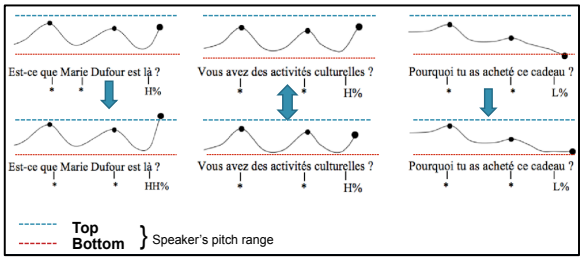


Figure 2: *Stimuli manipulation*.

The final syllable of each question (on which the tonal contour is realized) was composed exclusively of sonorants and mid oral vowels. Spanish questions were stressed either on the penultimate or on the final syllable. As for French wh-'in-situ' questions, a noun followed the interrogative word in order for the rise related to the pitch accent to be anchored on the wh-word.

3.2. Participants and task methodology

As mentioned three categories of participants were set up, the native speakers being control groups and the non-native speakers the experimental group. This latter group (hence FL2) was composed of 23 native Mexican Spanish speakers aged 24.7 years on average (*SD* 5.59) who attended a French course at the university of Mexico. These participants were classified into two sub-groups according to their level of proficiency in L2 French: 14 and 9 students were positioned at A2 and B1 level respectively (according to the CEFR). One of the two control groups was composed of 17 native speakers of French aged of 29.5 years (*SD* 11.14), and the other control group (SL1) consisted of 16 native speakers of Mexican Spanish aged 32.5 years on average (*SD* 8.14). The FL1 and FL2 groups were tested on French stimuli, while SL1 participants were tested on Spanish stimuli.

Since the tree contours tested here are actually acceptable in both French and Spanish, without implying a specific meaning in the investigated questions, we choose a methodological procedure that allows evaluating how these differences are perceived as gradient by native and non native listeners. Therefore, we opted for a rating task methodology.

Participants were asked to read on a computer screen different discursive contexts presenting a scenario for each question. They were instructed to listen to the resynthesized questions inserted in each scenario, and to evaluate their melody within a 1 to 5 scale (1= melody is inappropriate, 5= melody is appropriate). 66 fillers were included in the test, stimuli were randomized, and the test was presented in 4 batches to each group of participants (batches were also randomized). The test was carried out in a quiet room with an ordinary computer and high-quality headphones.

4. Results

Listeners' ratings were separated in four sets for the analysis, each set being associated with a question type: yes-no declarative questions, yes-no questions with an interrogative marker, wh-'fronted' questions, and wh-'in-situ' questions. In order to analyze the ratings attributed to the stimuli by the various groups of participants, and to test which differences were statistically significant, we constructed linear mixed effect models that took into account the predictor variables Contour (HH%, H%, L%), Group (FL1, FL2, SL1) and random intercepts and slopes for Subjects, for each of the structures which were tested separately. The contribution of each predictor variable was assessed using model reduction and likelihood ratio tests (χ^2).

With regard to yes-no declaratives questions, results showed significant main effect of Contour for ratings: listeners rated HH% higher than H% ($p<.01$). There was, however, no main interaction between Group and Contour for the ratings, i.e., all participants prefer extra rising tunes independently of their L1 in this question type (FL1 included). It was expected that, on the basis of the canonical rising patterns pointed in section 2, Spanish listeners would evaluate better HH% than H%, whereas French native listeners would prefer H%. So, the results obtained for this question set cannot help validating the hypothesis of a difference between L2 and natives speakers: FL1 participants prefer extra-rising contours to rising ones, even if this tonal contour is not frequent in their oral productions, as well as in the canonical description of French intonation.

As for yes-no questions with a lexical or morphosyntactic marker, we found a main effect of Group interacting with the Contour for the ratings: FL2 rated rising tunes (H% and HH% grouped) higher than the falling ones, while FL1 did not show a clear preference ($p<.0001$). In other words, the distribution of ratings for this question set depends on the native language of listeners. In this case, the two Groups did not show a preference for either H% or HH% ($p>.05$). As illustrated in Fig 3, ratings for L% in yes-no questions with a marker were extremely different across the two groups. These results could be interpreted as showing that non-native listeners prefer rising tunes, even when the modality of the sentence is marked by others means.

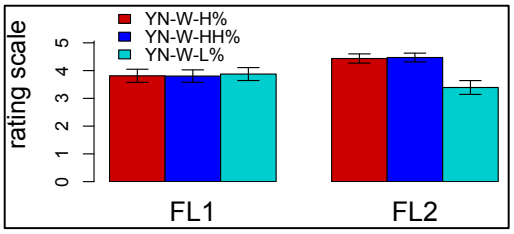


Figure 3: *Ratings given to yes-no marked questions*

In the set containing wh-questions with fronted wh-words, the analysis revealed that all listeners preferred rising contours (HH% and H% grouped) to falling ones ($p<.0001$). When comparing ratings of Groups concerning only the H% and HH%, it appears that there are significant differences as well: all participants rated higher H% than HH% ($p<.01$). Contrary to our expectations, there is no a main effect of interaction between Group and Contour in ratings. These results revealed that all groups, independently of their linguistic background, show a preference for H% in this structure. The results obtained cannot allow supporting any hypothesis. Both native and non-native speakers do show preference for tunes that they do not use so much.

An analysis carried out for wh-'in-situ' questions showed that there is a main effect of Contour for scores: all participants prefered rising tunes (H% and HH% grouped) to falling ones ($p<.001$). However, an analysis modeling the interaction between Group and Contour for ratings (rising vs. falling) did not reach significance: all participants, be they French or Spanish natives, prefer rising contours than the falling ones for this question set. When we excluded the falling contours, we found a significant interaction between the Group and Rising Contours (H% vs HH%) for ratings: FL2 evaluated the extra-rising contour better than H%; whereas FL1 did not show any preference ($p<.05$). These differences are illustrated in figure 4:

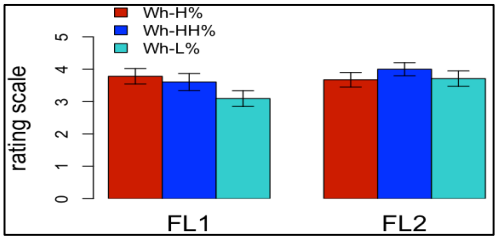


Figure 4: *Ratings given to wh-'in-situ' questions.*

Finally, in order to test whether the proficiency level had an effect on the choice of contour, we carried out an analysis including all question types. Main effects were found for ratings of Contour: rising tunes (H% and HH% grouped) received higher scores than falling ones. When comparing only ratings attributed to rising contours (excluding L%), results showed that all learners evaluated HH% higher than H% ($p<.05$). Nevertheless, no significant interaction between proficiency level and Contour on the rating choices was found. In other words, all FL2 listeners attributed ratings in the same way, despite their proficiency in L2 French.

5. Discussion and conclusion

In a global perspective, this study shows that all listeners, independently of their L1, rated rising contours better than falling contours when listening to yes-no and wh-questions, showing that there is a discrepancy between the patterns observed in oral productions and the perception. Regarding the scores of native listeners, the study revealed unexpected results. On the one hand, native listeners prefered rising tunes over falling tunes in questions they usually produced with falling contours (wh-questions). On the other hand, expected differences in rating concerning the form of the rising tunes (H% vs. HH%) were not observed in yes-no questions: FL1 listeners showed the same tendency as the other groups in rating the HH% contour higher, even though they do not employ it in their oral productions. These results suggest that rising tunes are considered as a universal prototypical form for declarative questions, regardless of their phonetic form (H% or HH%). In addition, rising forms are associated with all types of questions in the mind of both French and Spanish listeners, even if the contours that surface in the productions may be falling for some question types. In other words, listeners preferred rising contours because it is the form they associate with questions in their mind, independently of the fact that it does not correspond to what they produce. This explanation goes along the lines of Ohala's Frequency Code and Gussenhoven's Effort Code (see [15]).

In wh-questions, native listeners did not prefer the extra rising contour HH%. This could possibly result from the fact that this tune is unfamiliar to them for these question types. In addition, learners provide higher ratings for rising tunes (H% and HH%) associated with yes-no questions with a marker and with 'in-situ' questions than native listeners do. Interestingly, in the case of wh-'in-situ' questions, learners display a preference for the HH% contour in comparison to native speakers. By and large results show that native listeners prefer rising tunes displaying a familiar or natural increase in pitch for wh-questions (H% covering approx. 8 semitones like in their L1), and non-native listeners actually prefer uncommon extra-rising contours (HH% increasing more than 11 semitones) when listening to L2 stimuli. This is even more so in the case of questions with a morphosyntactic form that does not exist in the learners' L1. This interesting result suggests that the use of the HH% contour may be is related to the occurrence of a linguistic form that does not exist in their L1. This form could thus be seen as a default form in the case of linguistic insecurity. Further research on this issue is necessary to confirm that hypothesis and guaranty that results may not result from the task itself that asks for a kind of metalinguistic analysis.

6. Acknowledgements

This study was supported by a doctoral grant from CONACYT (Mexico) and by the ANR Labex EFL "Empirical Foundations in Linguistics" (workpackage PPC 4: The acquisition of phonetics, phonology and prosody in French and English as an L2).

7. References

[1] Mennen, I. "Bi-directional interference in the intonation of Dutch speakers of Greek", Journal of Phonetics 32, 543-563, 2004.

[2] Mennen, I. "Phonological and phonetic influences in non-native intonation", .in J. Trouvain, & U. Gut [Eds.], Non-native Prosody: Phonetic Descriptions and Teaching Practice (Nicht-muttersprachliche Prosodie: phonetische Beschreibungen und didaktische Praxis), Mouton De Gruyter, 53-76, 2007.

[3] Jilka, M, “Different manifestations and perceptions of foreing accent in intonation”, in J. Trouvain, and U. Gut [Eds], Non-Native Prosody. Phonetic Description and Teaching Practice, 77-96, Mouton de Gruyter, 2007.
[4] Santiago, F. and Delais-Roussarie, E., “The acquisition of Question Intonation by Mexican Spanish Learners of French”, in Prosody and Language in contact: L2 acquisition, attrition, languages in multilingual situations, Springer Verlag, to appear.
[5] Santiago, F. & Delais-Roussarie, E. “La prosodie des énoncés interrogatifs en français L2”, in L. Besacier, B. Lecouteux and G. Sérasset [Eds], Proceedings of Journée d’Études sur la Parole, JEP 2012, 265-272, 2012.
[6] De la Mota, C., Butragueño, P. and Prieto, P., “Mexican Spanish Intonation”, in P. Prieto, and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, 319-350, Lincom Europa, 2010.
[7] Sosa, J., La entonación del español: su estructura fónica, variabilidad y dialectología, Cátedra, 1999.
[8] Sosa, J., “Wh-questions in Spanish: Meanings and Configuration Variability”, Catalan Journal of Linguistics, 2: 229-247, 2003.
[9] Grundstrom, A. “L’intonation des questions en français standard”, in A Grundstrom and P. Léon, Interrogation et Intonation, 19-51, Klincksieck, 1973.
[10] Di Cristo, A., “Intonation in French”, in H. Daniel J, and D. Albert [Eds], Intonation Systems: A Survey of twenty languages, 195-218, Cambridge University Press, 1998.
[11] Beyssade, C., Delais-Roussarie, E., and Marandin, J.-M., “The prosody of interrogatives in French”, Nouveaux cahiers de linguistique française, 28: 163-175, 2007.
[12] Déprez, V., Syrett, K., and Kawahara, S., “The interaction of syntax, prosody, and discourse in licensing French wh-in-situ questions”, Lingua, 124: 4-19, 2012.
[13] Quilis, A., Tratado de fonología y fonética españolas, Gredos, 1993.
[14] Horgues, C. Prosodie de l'accent français en anglais et perception par des auditeurs anglophones. Unpublished PhD Thesis, Université Paris Diderot, 2010.
[15] Gussenhoven, C, The Phonology of Tone and Intonation. Cambridge, Cambridge University Press, 2004.

Where do questions begin? – phrase-initial boundary tones in Hungarian polar questions

Katalin Mády[1], *Ádám Szalontai*[1,2]

[1]Research Institute for Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[2]Eötvös Loránd University, Budapest, Hungary
{mady|szalontai}@nytud.hu

Abstract

Hungarian prosody is left-headed, as suggested by the placement of the accent on the initial syllable on the level of prosodic words and the placement of the strongest pitch accent on the first accented word of the prosodic phrase. Earlier studies have pointed out that the left edge of the intonational phrase can bear a phrase-initial boundary tone that distinguishes between string-identical *wh*-interrogatives and *wh*-exclamatives. In this paper, two other string-identical sentence types, polar questions and declaratives, are investigated with respect to their prosodic features. Polar questions were characterised by a higher f0 maximum and a lower sentence-initial f0 than declaratives. The only pitch accent within the sentence was low, whereas declaratives had falling pitch accents. Sentence-final f0 and the pitch level of the accented syllable did not show a consistent pattern across speakers. It is concluded that low sentence-initial f0 together with the high t! one on the penultimate syllable is a relevant marker of polar questions in Hungarian.
Index Terms: intonation, phrase-initial boundary tone, polar question, declarative, sentence type.

1. Introduction

1.1. Left-headedness in Hungarian prosody and syntax

Hungarian prosody is left-headed, as suggested by the placement of the accent on the initial syllable on the level of the prosodic words and the placement of the pitch accent on the first accented syllable of the prosodic phrase [1, 2]. [3] who builds on earlier observations by [4] and [5] suggest a mapping rule between prosody and syntax that aligns the left edge of the syntactic phrase with the left edge of the phonological phrase. [6] makes a similar claim, namely that the nuclear stress rule of [7] operates in a direction opposite to that in English in that the primary phrasal stress falls on the left edge. This correlates well with the structure of the Hungarian clause, which can be divided into two basic units, first the topic (optional) which is linearly followed by the predicate (obligatory) at the left edge of which is the focus position, bearing the primary a! ccent. If a focussed constituent is present, it leads to the deaccentuation of the verb and the post-verbal elements within the same prosodic unit.

(1) [$_{Top}$ Tegnap] [$_{Pred}$ [$_{Foc}$ János] ette meg a levest]
yesterday John ate PV the soup
'Yesterday it was John who ate the soup.'

Due to the obligatory nature of the predicate, it is this syntactic structure that best correlates with higher level prosodic units, namely the intonational phrase. The word order of the Hungarian clause is determined by information structure: the constituent which is focussed is placed in the immediate pre-verbal position, thus receiving the main pitch accent in the clause [6]. In neutral sentences (those with broad focus) this position is usually occupied by the verbal prefix (PV in (1) above), which bears the main pitch accent in these cases. However if there is a focussed constituent and a verbal prefix, the prefix occurs immediately after the verb.

1.2. Sentence types and prosody in Hungarian

In theoretical classifications of clause or sentence types, the set of basic sentence types includes declaratives, interrogatives and imperatives, whereas exclamatives are considered to be outside of this set [8, 9]. The basis for this distinction is that while basic sentence types are definable with the help of a small number of necessary and sufficient formal criteria in the languages where they occur, there do not seem to be equally available unambiguous formal criteria for setting apart structures that express the meaning attributed to exclamatives. Instead, exclamatives can be only characterised by their intonational pattern.

Exclamatives were described as having a "high tone followed by a slow descent" by [10] previously. In a recent production study, string-identical *wh*-interrogatives and a particular type of *wh*-exclamatives were compared with respect to their tonal differences [11]. The analysis was based on tonal categories such as phrase-initial and phrase-final boundary tones and pitch accent patterns along with a parametric analysis of f0. Although it was expected that *wh*-exclamatives contained a higher f0 maximum than *wh*-interrogatives and a higher phrase-final boundary tone, they had actually lower f0 maxima than *wh*-interrogatives, and there was no difference between the sentence-final f0 values. The main distinction was the shape of the (only) pitch accent on the *wh*-word: exclamatives had rising accents, interrogatives falling ones. It was not clear whether the accent patterns were the primary cues for the into! national distinction, or whether they were a consequence of different phrase-initial boundary tones as had been proposed by [12] for *wh*-interrogatives. A follow-up perception experiment [13] with deaccented particles before the pitch-accented *wh*-word showed that sentence-initial f0 enhanced sentence-type identification, whereas the pitch accent pattern and sentence-final f0 did not.

Another example of sentence types that are only distinguished by their intonational patterns are yes/no interrogatives and declaratives. [14] and [15] suggest that there is an L* accent on the verb and an H L% boundary tone spread over the penultimate and final syllables. According to [16], the perception as a question is enhanced both by a higher f0 and a relatively late

timing of the peak within the penultimate syllable. It is not clear whether the L* pitch accent and a phrase-initial boundary tone contribute to the intonational distinction between these sentence types.

For the prosodic description of the two sentence types, the following questions are of interest: (1) Is there a sentence-initial tonal distinction? (2) Does the rising-falling sentence-final tone in interrogatives result in higher f0 due to truncation? (3) Are the L* accent and the H tone in questions indeed lower and higher in terms of absolute values than the corresponding values in declaratives?

2. Material and methods

2.1. Material

The study used five pairs of string-identical sentences with the structure in (2):

(2) De most végül lámpa is van náluk ./?
But now in the end lamp also be.3SG with them
'But in the end they have a lamp with them.'
'But in the end do they have a lamp with them?'

The first three items of the string *De*, *most*, and *végül* are unaccented discourse particles. These were included to provide enough phonological space for the sentence-initial boundary tones to manifest independently of the main pitch accent which in this sentence falls on *lámpa* 'lamp', the constituent in the focus position of the predicate. The combined length of these discourse markers was 4 syllables in all target sentences. The target sentences were preceded by disambiguating contexts.

The expected prosodic pattern for both questions and declaratives was to contain one single intonational phrase, 2–3 deaccented particles (4 syllables altogether) and one pitch accent on the focussed word with post-focal deaccentuation. The last pitch accent preceded the sentence-final boundary by 4–8 syllables, which typically resulted in an overall rising f0 between the accented and the penultimate syllable for questions, and a flat or falling contour for declaratives. Sentences that contained more than one prosodic phrases or several pitch accents were excluded from further analysis. 296 utterances in total were analysed.

There were 7 subjects (all female) with a mean age of 21 years. Their task was to read aloud sentences that were presented to them on a screen using the experimental software SpeechRecorder [17]. Each pair of sentences was presented five times in individually randomized order.

2.2. Methods

The following parameters were investigated:

- sentence-initial f0 (on the first vowel),
- sentence-final f0 (on the last vowel),
- lowest f0 on the accented syllable,
- f0 maximum within the sentence.

F0 was measured using Praat's standard autocorrelation method with a window length of 30 ms and a window shift by 5 ms. Tones with an f0 minimum below 130 Hz were regarded as creaky-voiced.

Additionally, pitch accents were labelled with respect to their pattern. Categorisation relied on the actual f0 movements rather than on phonological labels.

Mixed models with the random effects subject and sentence were carried out for each parameter. Significance was tested by the `Anova()` function of the `car` package in R, and the significance level was set to $p < 0.05$. Hertz values were transformed into semitones.

3. Results

Figure 1 shows typical examples of a declarative and an interrogative sentence. The declarative contains one pitch accent that falls on the first syllable of the word *LÁMpa* 'lamp'. The falling accent is typically found in pre-verbal elements within the predicate [18]. The polar question contains one pitch accent on the same syllable which has a low tone. The f0 maximum is associated with the penultimate syllable.

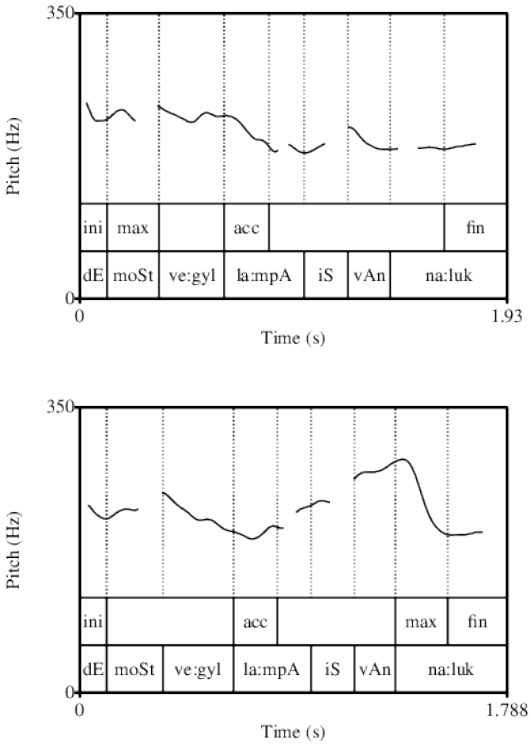


Figure 1: *Pitch contour of the declarative sentence 'But in the end they have a lamp with them' (top) and the polar question 'But in the end do they have a lamp with them?' (bottom). Measurements described in Section 2 were based on the syllables marked as* ini, acc, max, fin.

3.1. Sentence-initial f0

There was an overall tendency for sentence-initial f0 to be lower in interrogatives (mean = 206 Hz) than in declaratives (mean = 224 Hz, difference: 1.48 semitones, $p < 0.005$). The tendency was present in 6 out of 7 speakers (see Fig. 2). Intra-speaker differences ranged from -0.03 to 5.88 semitones. Occurrences of creaky voice were rather low in this position (3% of all cases), and they were evenly distributed between declaratives and polar questions.

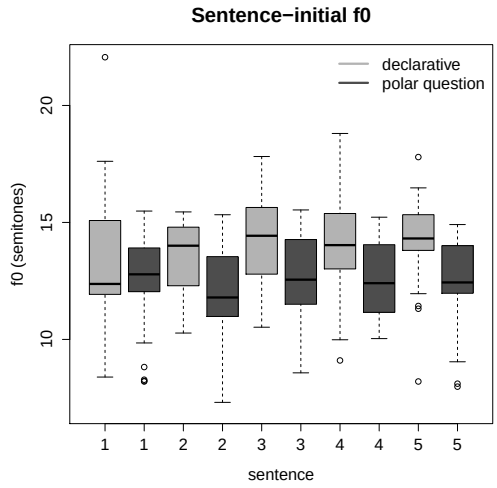


Figure 2: *Sentence-initial f0 in semitones. Pairwise comparison of sentences (top) and speakers (bottom).*

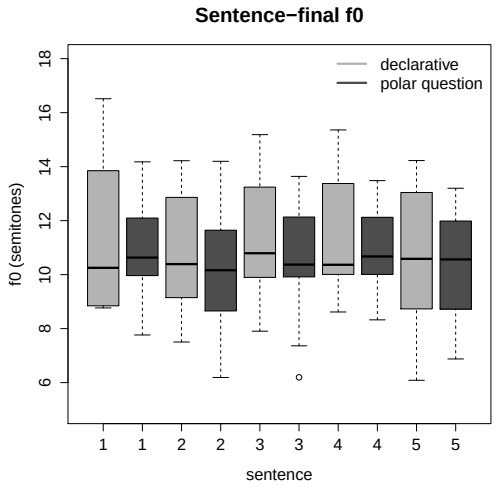


Figure 3: *Sentence-final f0 in semitones. Pairwise comparison of sentences (top) and speakers (bottom).*

3.2. Sentence-final f0

As is shown in Fig. 3, sentence-final f0 in interrogatives was realised somewhat lower (mean = 185 Hz) than in declaratives (190 Hz, difference: 0.48 semitones, $p < 0.005$). This tendency was present in 4 out of 7 speakers. Intra-speaker differences ranged from -1.82 to 4.42 semitones. 34% of all sentence-final syllables were produced with creaky voice, and their occurrences in the two sentence types were equal.

3.3. F0 maximum

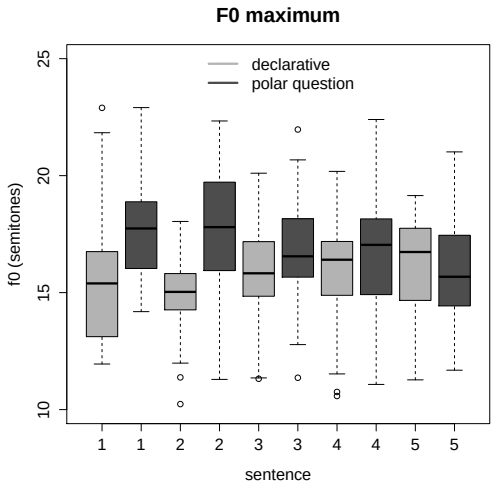
The maximal f0 was significantly higher in questions than in declaratives (269 Hz vs. 248 Hz, 1.39 semitones, $p < 0.005$, see Fig. 4). One out of 7 speakers showed an opposite pattern. Intra-speaker differences ranged from -3.13 to 1.88 semitones, and the voicing was always modal.

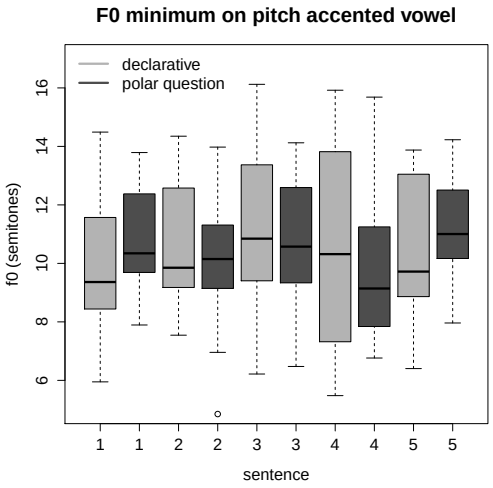
The maximal pitch in polar questions is presumably perceived even higher than the actual f0 suggests, since the maximum appears later in the sentence than in declaratives, and it is relatively higher compared to the declination line.

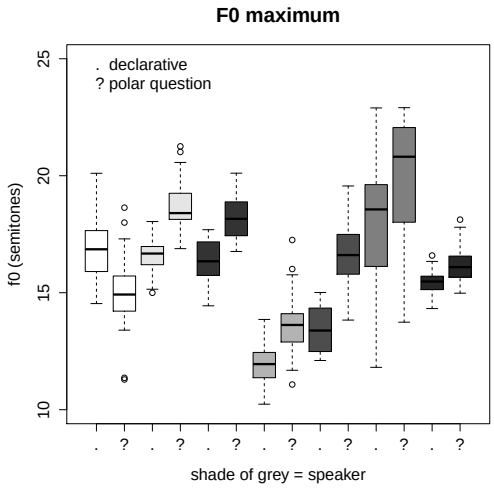
3.4. Pitch accent

Declaratives and polar questions are characterised by different pitch accents. The default pitch accent for focussed words in declaratives is a falling one, typically H*+L, see [18, 19]. Polar questions are realised with low or rising accents [16, 14]. In order to test whether the lowness of the f0 of the pitch accent is relevant for the distiction, the lowest f0 on the accented syllable was measured.

As demonstrated in Fig. 5, there was no clear tendency across speakers with respect to local f0 minima. The mean f0 was slightly higher for declaratives (187 Hz) than for questions (184 Hz, $p = 0.02$), but this is reflected only in 4 out of 7 speakers, and the mean difference was only 0.2 semitones which is only slightly above the just noticeable difference. The intra-speaker difference ranged from -0.42 to 1.72. Creaky voicing occurred in 7% of all questions, whereas it was half as frequent in declaratives.







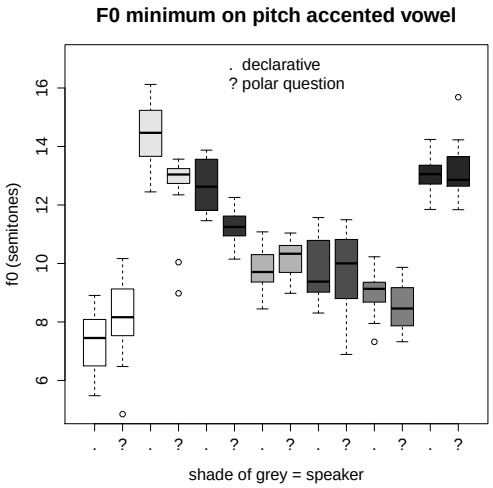


Figure 4: *F0 maximum in semitones. Pairwise comparison of sentences (top) and speakers (bottom).*

Figure 5: *Local f0 minimum in pitch-accented vowels. Pairwise comparison of sentences (top) and speakers (bottom).*

4. Discussion and conclusions

The above results show that string-identical polar questions and declaratives differ both in terms of their tonal pattern and their f0 parameters. A consistent intra-speaker pattern was only observed for sentence-initial f0 that was lower and the overall f0 maximum that was higher in questions. Speakers did not show a homogeneous pattern with respect to the utilisation of sentence-final f0 and the local f0 minimum on the pitch-accented vowel.

As was said in the Introduction, string-identical *wh*-interrogatives and exclamatives are distinguished by higher sentence-initial f0 and higher f0 maxima for interrogatives, whereas sentence-final f0 does not play a role. The present data also show a higher f0 maximum for interrogatives, but a *lower* sentence-initial f0. Thus, it cannot be generalised that interrogatives were marked by higher sentence-initial f0 altogether, but it seems that sentence-initial f0 is indeed relevant for the sentence type distinction. Sentence-final f0, however, is rather inhomogeneous and is thus not expected to be a crucial marker for polar questions and declaratives. These results further corroborate the importance of the the left edge of prosodic units in Hungarian.

Two questions need further investigation in the close future: (1) Can the perceptual relevance of phrase-initial boundary tones for polar questions and declaratives be stated? (2) Is a high f0 maximum generally characteristic for interrogatives, and if yes, is it a linguistic or a pragmatic feature?

5. Acknowledgements

The study was supported by the Hungarian Scientific Research Fund (101050, 100804) and the Momentum program of the Hungarian Academy of Sciences, project title *Interaction of linguistic subsystems in the production and perception of scope*. Thanks to Zsuzsanna Bárkányi and Beáta Gyuris for their support with the creation and the collection of the data.

6. References

[1] L. Hunyadi, *Hungarian sentence prosody and universal grammar: on the phonology-syntax interface*. Frankfurt/Main: Lang, 2002.

[2] L. Varga, *Intonation and Stress: evidence from Hungarian*. Basingstoke and New York: Palgrave Macmillan, 2002.

[3] K. Szendrői, "A stress-based approach to the syntax of hungarian focus," *The Linguistic Review*, vol. 20, pp. 37–78, 2003.

[4] I. Vogel and I. Kenesei, "The interface between phonology and other components of the grammar: The case of hungarian." *Phonology Yearbook*, vol. 4, pp. 243–263, 1987.

[5] ——, "Syntax and semantics in phonology." in *The Phonology-Syntax Connection*, S. Inkelas and D. Zec, Eds. Chicago: The University of Chicago Press, 1990.

[6] K. É. Kiss, *Hungarian Syntax*. Cambridge: Cambridge University Press, 2002.

[7] N. Chomsky and M. Halle, *The Sound Pattern of English*. New York: Harper and Row, 1968.

[8] J. M. Sadock and A. M. Zwicky, "Speech act distinctions in syntax," in *Language Typology and Syntactic Description*, T. Shopen, Ed. Cambridge University Press, Cambridge, 1985, pp. 155–196.

[9] E. König and P. Siemund, "Speech act distinctions in syntax," in *Language Typology and Syntactic Description, Vol. 1*, T. Shopen, Ed. Cambridge University Press, Cambridge, 2007, pp. 276–324.

[10] L. Kálmán, *Magyar leíró nyelvtan 1. Mondattan*. Budapest: Tinta Könyvkiadó, 2001, [Hungarian descriptive grammar 1. Syntax].

[11] B. Gyuris and K. Mády, "Approaching the prosody of Hungarian wh-exclamatives," in *VLLXX: Papers presented to László Varga on his 70th birthday*, P. Szigetvári, Ed., 2013, http://seas3.elte.hu/tmp/vlfs/gyuris-mady.html.

[12] L. Mycock, "Prominence in Hungarian: the prosody–syntax connection," *Transactions of the Philological Society*, vol. 108, no. 3, pp. 265–297, 2010.

[13] K. Mády, B. Gyuris, and A. Szalontai, "Phrase-initial boundary tones in hungarian interrogatives and exclamatives," in *Proceedings of the Prosody-Discourse Interface Conference (IDP-2013)*, 2013, pp. 69–73.

[14] D. R. Ladd, *Intonational phonology, 2nd ed.* Cambridge: Cambridge University Press, 2008.

[15] l. Varga, "Boundary tones and the lack of intermediate phrase in hungarian (revisiting the hungarian calling contour)," *The Even Yearbook*, vol. 9, pp. 1–27, 2010.

[16] M. Gósy and J. Terken, "Question marking in hungarian: Timing and height of pitch peaks," *Journal of Phonetics*, vol. 22, pp. 269–281, 1994.

[17] C. Draxler and K. Jänsch, "SpeechRecorder – a universal platform independent multi-channel audio recording software," in *Proc. International Conference on Language Resources and Evaluation*, Lissabon, 2004, pp. 559–562.

[18] K. Mády and F. Kleber, "Variation of pitch accent patterns in Hungarian," in *Proc. 5th Speech Prosody Conference, Chicago*, 2010, pp. 100 924:1–4.

[19] K. Mády, "A fókusz prozódiai jelölése felolvasásban és spontán beszédben [prosodic marking of focus in read and spontaneous speech]," in *Beszéd, adatbázis, kutatások*, M. Gósy, Ed. Budapest: Akadémiai Kiadó, 2012, pp. 91–107.

Encoding and decoding Confidence information in speech

Xiaoming Jiang [1], *Marc D. Pell* [1]

[1] School of Communication Sciences and Disorders and Center for Research on Brain, Language and Music, McGill University, Canada

xmjiang1983@gmail.com, marc.pell@mcgill.ca

Abstract

This study aims to investigate the perceptual-acoustic correlates of vocal confidence. Statements with different communicative functions (e.g., stating facts, making judgments) were spoken in confident, close-to-confident, unconfident and neutral voices. Statements with preceding linguistic cues (e.g. I'm positive, Most likely, Maybe, etc.) or no linguistic cues were presented to sixty listeners in a perceptual study. The listeners were asked to judge whether statements conveyed some level of confidence, and if so, they were asked to evaluate the level of confidence of the speaker. The results demonstrated that the intended levels of confidence varied in a graded manner in the perceptual rating score; the more confident the statement intended to be, the higher the rating. In general, the neutral voice was judged to be more confident than the close-to-confident voice, but less than the confident voice. The presence of a linguistic cue tended to increase ratings of confident voices but decrease ratings of voices in the less confident voice conditions. To evaluate how specific prosodic cues are used to encode and decode confidence information, acoustic analyses were performed on the stimuli without the linguistic cue based on the mean perceptual rating of speaker confidence for each item. Results showed that statements rated as confident versus unconfident differed in the mean and the variance of fundamental frequency (f0) as well as speech rate, with confident statements exhibiting lower mean f0, smaller f0 variance, and faster speaking rate than unconfident statements. The perceived level of confidence was differentiated in the mean fundamental frequency in a parametric way, the lower the level of confidence, the higher the mean f0. Confident voices were also distinct from the other three conditions in terms of mean and range of amplitude (i.e., loudness). These findings shed light on how linguistic and paralinguistic cues reveal confidence-related information to listeners during speech.

1. Introduction

Humans have the ability to encode different types of emotive and social meanings in speech communication, which are often understood by listeners through an inferential process that weighs evidence from available linguistic and paralinguistic cues. Among the different emotive states that can be expressed are emotive devices that serve to foreground speaker-content (*evidentiality* devices). Within this category, *confidence* refers to cues that provide evidence of the reliability, correctness, or truth value of a speaker's statement; in social interactions, listeners make inferences about the confidence of other speakers to make appropriate decisions based on the perceived reliability of what is said, and they may also use this information to associate specific social or personality traits to the speaker (e.g. perceived confidence is usually associated with persuasiveness, [1]). It is suggested that evidentiality/confidence is communicated through the choice of specific linguistic structures (e.g. modal adverbs) and/or the speaker's prosody (e.g. changes in pitch / intonation contour and other acoustic parameters to make speech sound doubtful, certain, authoritative, submissive, etc.). However, little empirical work has been conducted on how different levels of confidence are realized by a speaker in terms of prosodic variation, nor is it well known how speakers use prosodic and linguistic cues in the decoding of speaker confidence. The goal of the present study was to supply preliminary data on perceptual and acoustic features of utterances that convey varying levels of confidence in English, as a first step for advancing knowledge of how evidentiality devices operate in speech communication.

2. Methods and results

2.1 Emotion elicitation study

2.1.1 Participants

Six native Canadian English speakers (Mean age in years = 22.8, three females) were recruited to produce statements expressing different levels of confidence in their native language. Speakers were selected for having lay experience in acting (e.g. in community theatre) or in public speaking (e.g. radio) and were compensated $35 after the recording session.

2.1.2 Materials

One hundred and fifty-one sentences were constructed by a native speaker of Canadian English (one of the authors, MDP). Sentences were of three types: 67 were statements describing a fact, 40 were descriptions of one's intention to initiate an action, and 44 were descriptions of one's judgment toward other people or things. Each sentence was intended to be produced in a neutral manner and to convey three levels of confidence about the subject matter: confident, close-to-confident, and unconfident. In order to facilitate production of sentences that were inflected to convey different levels of confidence in a way that was as natural as possible, lexical phrases consistent with each level of confidence were used as the beginning of the sentence during recording. For the confident level, speakers began with either "Definitely", "For sure", "I'm certain" or "I'm positive"; for the close-to-confident level, they produced "I think" "Most likely", "I'm pretty sure", "I'm almost certain"; and for the unconfident level, they began with "Maybe", "Perhaps", "It's possible", "There's a chance". Finally, 151 wh-questions were also created to facilitate naturalistic expressions of each sentence during the recording session by having speakers produce target sentences as part of a mini-dialogue (e.g. What will happen? – We will run out of gas). Care was taken that none of the questions led speakers to emphasize specific constituents in

the sentence, but rather, they could freely respond to the question with the appropriate level of confidence.

2.1.3 Elicitation and recording procedure

Each speaker was recorded separately in a sound-attenuated chamber. Sentences conveying neutral affect and each of the three levels of confidence were recorded in a separate block during the elicitation study. The neutral level always preceded the other levels. The order for recording specific confidence levels was randomized across speakers. Within each confidence condition, the three types of sentences (facts, intentions, judgments) were also recorded in a separate block. The order for recording specific type of sentences was randomized across confidence levels.

To facilitate expressions of confidence that were as naturalistic as possible, we created a dialogue setting in which the actor responded to questions posed by a female examiner, who was a native Canadian English speaker. For each confidence level, the examiner asked the question constructed for each target sentence and then the speaker produced the sentence with the appropriate level of confidence as if answering the speaker's question. The examiner provided clues to the speaker at the onset of each level that included descriptions of a scenario that would be likely to elicit the target level of confidence [2]. However, at no time did the examiner never model the vocal features that may be associated with specific impressions of confidence to participants. In the confident/close-to-confident/unconfident recording blocks, the speaker was instructed to produce the lexical phrase followed by the main sentence stem and to try to portray the level of confidence in their voice (prosody) throughout the sentence. They repeated each sentence twice. Recording blocks were separated by a short break to help the transition between modes of social expression. After producing each confidence level, speakers were asked to rate on a 7-point confidence scale (1=not at all confident, 7= very confident) how confident they were subjectively feeling when they produced sentences in each condition; the mean ratings for the six speakers were 6.5 in the confident voice condition, 4.7 in the close-to-confident condition, 1.8 in the unconfident condition, and 5.0 in the neutral condition.

All utterances were recorded onto digital media (Tascam Recorder) using a high-quality head-mounted microphone. The recordings were transferred to a computer and were saved as individual sound files in Praat. To reduce the number of exemplars included in the perceptual rating study, the two repetitions of each item produced by a given speaker were initially evaluated by two native English speakers to select the best (single) exemplar per item/speaker, based on a judgment of which item conveyed the intended target level of confidence, while excluding items that sounded unnatural (i.e., posed) and/or that had recording artifacts. This process yielded 3624 stimuli in total (6 speakers x 4 confidence levels x 151 items).

2.2 Perceptual study

2.2.1 Participants

A total of 60 listeners took part in the study (31 females and 29 males, with the mean age 25.2 years, mean education 16.6 years). All listeners were born and grew up in Canada and had English as their first language. None had lived outside of Canada for more than 1 year. All participants had normal hearing and no history of psychiatric or neurological disorders.

2.2.2 Materials and procedure

To test the perceptual recognition of confidence from utterances that contained lexical phrases or just prosodic cues, all recordings were further edited by removing the lexical phrase; this produced two versions of each statement with and without the lexical phrase ("with-cue" and "no-cue statement"). The resulting 6342 statements (6 speakers x 4 confidence levels x 151 items 2 cue types) were divided into six experimental lists (each with 1057 statements). The "with-cue statement" and the "no-cue statement" with the same sentence stem produced by the same speaker in the same prosody were never repeated in one list. Statements in each list were randomized for each listener and were separated in 20 short blocks. Each list was judged by 10 participants.

Listeners were tested separately or in pairs in a quiet experimental lab and were asked to perform two tasks sequentially. After presented with a statement, they first judged whether the statement conveyed some level of confidence by clicking YES or NO printed in two squares on the screen; they were informed that the level of confidence could be signaled by a lack of confidence or much confidence. If they answered YES, a 5-point scale was presented on the screen and they were asked to rate the speaker's level of confidence by choosing a number that best fit their impression of the last statement. A same number scale was also shown if they answered NO, however, they were asked to select any number to continue to the next trial.

2.2.3 Data analysis

A univariate analysis of variances were performed on the mean rating score of speaker's confidence for all statements of the two speakers, taking intended level of confidence (confident vs. close-to-confident vs. not confident), cue type (with lexical cue vs. no lexical cue), and type of utterance function (fact vs. intention vs. judgment) as three independent factors. Further analysis was planned when interaction between the factors were significant.

2.2.4 Results

Table 1 and Figure 1 demonstrate the mean perceptual ratings of with-cue and no-cue statements of each utterance type and in each level of confidence. The univariate ANOVA on all perceptual data revealed a significant main effect of level of confidence, $F(2, 1794)=2948.76$, $p<0.001$. Post-hoc Tukey's comparison confirmed that the speaker's confidence was rated the highest in the confident condition, followed by close-to-confident stimuli, followed by the unconfident stimuli. There was significant effect of cue type, $F(1, 1794)=215.78$, $p<0.001$, suggesting that with-cue statements were in general rated lower than the no-cue statements. The effect of type of utterance function was also significant, $F(2, 1794)=13.63$, $p<0.001$, suggesting that statements of intention were rated lower overall than statements of fact and judgment, with the

latter two not different from each other. Moreover, there was a significant interaction between level of confidence and cue type, $F(2, 1794)=136.20$, $p<0.001$ and a significant interaction between level of confidence and type of utterance function, $F(4, 1794)=2.40$, $p<0.05$. Separate analysis on each level of confidence revealed a significant effect of cue type for all levels of confidence: for confident, $F(1, 598)=51.80$, $p<0.001$, for close-to-confident, $F(1, 598)=257.28$, $p<0.001$, and for unconfident, $F(1, 598)=129.94$, $p<0.001$. Post-hoc comparison revealed a higher rating for the with-cue than the no-cue statement in the confident condition, whereas *lower* ratings were observed for the with-cue than no-cue statement in the close-to-confident and unconfident conditions.

Separate analysis also revealed an effect of type of utterance function for confident, $F(2, 598)=4.74$, $p<0.01$, for close-to-confident level, $F(2, 598)=12.98$, $p<0.001$, but not for unconfident level, $F(2, 598)=1.70$, $p>0.1$. These findings suggested that the lower ratings for intentions than for the other two functional types were most prominent in the confident and the close-to-confident conditions.

In order to examine the difference between the intended-neutral and the other levels of confidence, an additional ANOVA on the no-cue statements only revealed a significant effect of level of confidence, $F(3, 1195)=670.58$, $p<0.001$. Post-hoc comparison revealed that the intended-neutral utterances were rated as comparable to the close-to-confident utterances, lower than confident and higher than the unconfident utterances.

Table 1. Mean and standard deviation of the rating score of speaker's level of confidence for with-cue and no-cue statements in each intended confidence level.

	Confident				Close-to-confident			
	With Cue		No Cue		With Cue		No Cue	
Utterance Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fact	4.54	0.29	4.36	0.30	3.31	0.38	4.08	0.36
Intention	4.50	0.29	4.28	0.31	3.28	0.47	3.75	0.61
Judgment	4.56	0.27	4.42	0.29	3.48	0.37	4.00	0.48

	Unconfident				Neutral	
	With Cue		No Cue		No Cue	
Utterance Type	Mean	SD	Mean	SD	Mean	SD
Fact	2.29	0.36	2.84	0.66	4.14	0.33
Intention	2.24	0.33	2.70	0.59	3.91	0.32
Judgment	2.29	0.37	2.76	0.65	4.05	0.31

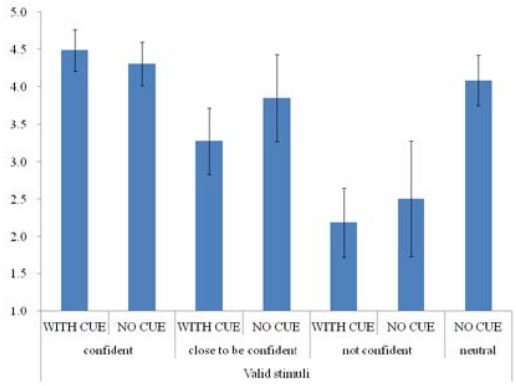


Figure 1. Perceptual rating scores of speaker's confidence for both with-cue and no-cue statements in each intended level of confidence

2.3 Acoustic study

2.3.1 Data analysis

In order to evaluate how different levels of confidence were acoustically differentiated in statements perceived as conveying confidence without lexical cues, acoustic measures were derived and analyzed only for "no-cue" statements. Due to the large number of intended neutral exemplars perceived as conveying some level of confidence, the statements with the frequency of "yes" response below 6 out 10 in the binary task were assigned as neutral, the statements with the frequency of "yes" response above 8 out of 10 were assigned as with-confidence. Among the with-confidence statements, the target level of confidence was re-assigned based on the perceptual results (see below). Those with the mean perceptual score above 4.2 in the 5-point rating task were designated as "confident", those with a mean score between 3.2 and 3.8 were designated as sounding "close-to-confident", and statements with a mean score below 2.8 were assigned to the "unconfident" condition.

Five acoustic measures that frequently differentiate among vocal emotion categories [2] [3] [4] were analyzed, including the mean fundamental frequency (mean f0, in Hertz), the range of fundamental frequency range (f0 variance, in Hertz), the mean amplitude (mean amplitude), the range of amplitude (amplitude variance), and speaking rate (in syllables per second). A normalization procedure was applied to the first four measures before comparing between speakers [2] [4]. Acoustic analyses were performed using Praat speech analysis software; the results for statements of fact for one male and one female speaker are reported here.

A multivariate analysis of variance (MANOVA) was performed on all valid no-cue statements that fell in the perceptual range defined for each target level of confidence. The mean and range of f0, mean and range of amplitude and speech rate were taken as dependent factors and the target level of confidence was considered as independent factors. A series of univariate ANOVA were further performed on each acoustic measure. We report the perceptual data for no-cue

statements from two speakers (1 male and 1 female) in the following section.

2.3.2 Results

Based on preliminary analysis of the two speakers, a total of 192 recordings were subjected to analysis. Table 2 demonstrated the mean perceived level of confidence and the values for each of the five acoustic parameters computed for the no-cue statements that were reassigned based on the ranking of the confidence perception. The one-way MANOVA was performed on acoustic data with the four perceived levels of confidence as independent variables and seven acoustic features (normalized mean f0, normalized f0 range, normalized mean amplitude, normalized amplitude range and speech rate) as dependent variables. The MANOVA indicated that the effect of level of confidence on the linear combination of the five acoustic parameters was significant, Wilk's λ = 0.575, $F(21, 523)=5.30$, $p<0.001$. Subsequent univariate analysis on each acoustic parameter revealed that the effect of level of confidence was significant for mean f0, $F(3, 188)=36.93$, $p<0.001$, f0 range, $F(3, 188)=4.16$, $p<0.01$, mean amplitude, $F(3, 188)=3.34$, $p<0.05$, amplitude range, $F(3, 188)=2.96$, $p<0.05$, and speech rate, $F(3, 188)=2.49$, $0.05<p<0.1$. Post hoc (Tukey's) comparisons revealed that the normalized mean f0 increased in value over neutral, confident, close-to-confident and unconfident meanings; the neutral level revealed a higher normalized f0 range and a higher speech rate than the three with-confidence levels, which did not exhibit any differences. Statements perceived as confident displayed both a higher mean amplitude and a higher amplitude range than statements in the other three conditions, which did not show any differences in amplitude measures. Statements perceived as not conveying any confidence were spoken more quickly overall than statements conveying the three levels of confidence.

Table 2. Mean normalized acoustic values for the no-cue statements produced by two speakers to express three levels of confidence and to the neutral level based on the perceived level of confidence.

Perceived level of confidence	Confidence rating	Mean f0	f0 variation	Mean amplitude	Amplitude range	Speech rate
Confident	4.50	0.40	1.37	1.23	1.96	4.61
Close-to-confident	3.60	0.65	1.86	0.96	1.58	4.50
Unconfident	2.30	0.76	1.85	1.02	1.64	4.27
Neutral	3.90	0.28	1.09	1.14	1.80	5.05

3. Discussion

This study demonstrates that listeners make use of prosodic information in speech to perceive different levels of confidence of a speaker when expressing facts, intentions, and making judgments which occur frequently in discourse. In particular, statements rated as conveying different levels of confidence were acoustically distinct in their f0 characteristics, with the more confident the statement intended to be the higher the confidence rating, among other acoustic distinctions. The additional presence of a lexical phrase that signaled speaker confidence had varying effects on perceptual ratings depending on the level of speaker confidence communicated: the presence of the lexical cue in confident statements tended to amplify confidence ratings, whereas lexical phrases in statements that lacked full confidence (close-to-confident, unconfident) tended to attenuate impressions of confidence, yielding lower ratings for "with-cue" utterances in these conditions. Interestingly, different utterance types (facts, intentions, judgments) differed in subtle but significant ways in how listeners perceived speaker confidence from prosodic cues. These findings are broadly consistent with Pell [5] who found that the attitudinal and interpersonal significance of prosody can be perceptually differentiated by both healthy and right-hemisphere-damaged adults, extending these findings to a group of healthy young adults.

The acoustic analysis on the no-cue statements further demonstrated several important mechanisms underlying the acoustic realization of confidence-related information in speech: 1) the f0 variance and the speech rate differentiated the with-confidence and neutral statements; 2) the mean f0 predicted the levels of confidence in a parametric way; and 3) the mean amplitude and the amplitude change highlighted the confident-level specifically. Taken together, our findings indicate that prosodic information guides how listeners infer the confidence state of a speaker, and that these pragmatic inferences may be affected by both linguistic and paralinguistic cues in speech communication, and may vary for different speech acts.

4. References

[1] Scherer, K., London, H. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality, 7,* 31-44.

[2] Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S.A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37,* 417-435.

[3] Cheang, H.S. & Pell, M.D. (2008). The sound of sarcasm. *Speech Communication, 50,* 366-381.

[4] Liu, P., & Pell, M.D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal stimuli. *Behavior Research Methods, 44,* 1042-1051.

[5] Pell, M.D. (2007). Reduced sensitivity to prosodic attitudes in adults with focal right hemisphere brain damage. *Brain and Language, 101,* 64-79.

5. Acknowledgements

This study was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to Marc D. Pell. We are very grateful to Sonia Kroll, Lorraine Chuen, Nabitha Kanagaratnam, Mary Giffen, Caleb Harrison, and François Anderson for their assistance in the study.

Aspects of Prosodic Phrasing in Turkish

Jane Kühn

Department of Linguistics (SFB 632), University of Potsdam, Germany
jkuehn@uni-potsdam.de

Abstract

This pilot study investigates the prosodic marking of contrastive in-situ focus in monolingual Turkish. The results of the production study are based on a phonological and phonetic analysis of information structure modified target sentences. The prosodic analyses reveal (i) feature that derive properties of prosodic phrasing which are inherent to phrase languages [5]. It is shown that Turkish is a radical splitting language since each prosodic word (ω) [14] forms its own phonological phrase (φ) [14] indicated by a high phrase tone (H-) aligned to ω- final syllables. The language´s preference for radical splitting of simple SOV sentences is maintained in information structure modified targets by one speakers group, but modified by another group in favor of wrapping adjacent given constituents into one φ. The analyses reveal (ii) that prosodic cues are not crucial to mark in-situ focus in Turkish, but they may be used to contextualize information structure. If focused constituents are marked at all by prosodic means they do not show an increased pitch like most Germanic languages, but focused constituents are aligned to prosodic boundaries. The data motivate the claim that prosodic alignment is an adequate way to describe the prosodic realization of focus in Turkish.

Index Terms: Turkish, focus, alignment

1. Introduction

Turkish is claimed to have two distinct focus marking strategies: syntactic movement and prosodic focus marking in-situ [2, 6, 9], i.e. when the focused constituent remains in the default word order. Previous studies on prosodic focus marking in Turkish describe focus in the original meaning of prominence as an increase of acoustic parameters and the modulation of pitch accents [10]. Correspondingly, different tunes are described for different information structural parts: e.g. H*L- for focused and L*H- for given constituents [15]. H-boundary tones are assumed for pre-nuclear phrases, and an H* nuclear pitch accent is designated to align to the immediately pre-verbal position in syntactically un-marked sentences [11], as demonstrated in Figure 1.

Figure 1. *Broad focus intonation contour in SOV*

H-	H*	L- L%
(subject)xp	(object	verb)vp

All tones are usually aligned to the last syllable of a ω, since Turkish word stress is final [12]. However, a recent acoustic analysis [8] reveals that Turkish shows no straightforward pitch accent modification of the focused constituent for in-situ focus. Despite the lack of pitch range expansion a change in the f0 contour is observed for final and initial focus: A focused verb in SOV declaratives shows an *immediately pre-focal rise* on the preceding constituent and post-focal compression (PFC) is observed after initial focus.

Concerning the cross-linguistic realization of focus prominence, typological studies [1, 3] show that focus is not necessarily marked by an increase of acoustic parameters such as f0, duration and intensity, as in most Germanic languages. Focus can also be captured by means of prosodic alignment understood as the correspondence between the edge of a syntactic and/or phonological constituent and the focused part of a sentence [13, 15]. On the basis of mapping theory [16], [17] proposes WRAP which offers at least three strategies for mapping: (a) radical splitting where each syntactical phrase (xp) forms its own (φ), (b) moderate wrapping where each xp forms its own φ and non-phrasal elements are wrapped with the closest phrase, and (c) radical wrapping where every element of the biggest xp gets wrapped into one φ. In the prominence theory of focus [17] focus needs to be maximally prominent which can be achieved by a modification of the phrasing structure, either by swapping of pitch accents or by the introduction and/or deletion of prosodic phrase boundaries. In focus as alignment theory [3] focus alignment may be obtained even in the absence of prominence.

This paper presents data that support the classification of Turkish as a phrase language [7] showing that high tones aligned to ω- final syllables represent phrase boundaries. Furthermore, the optionality of prosodic marking of in-situ focus is demonstrated. A change in phrasing according to information structure states prosodic alignment as optional tool to mark focus in Turkish.

2. Hypotheses

With regard to previous acoustic measurements [8] which did not succeed in identifying pitch increase as a focus marker, but revealed a modification of f0 by means of PFC and *pre-focal pitch increase* two assumptions are examined in the pilot production study:

- High tones on word final syllables do not represent typical pitch accents aligned to word stressed syllables but are phrase delimiting boundary tones.
- In-situ focus is prosodically marked by a modification of the phrasing structure, i.e. focus is aligned to prosodic boundaries by means of boundary insertion/deletion.

In practice, the high tone (H*) on nuclear constituents described by [11] was expected to be a further boundary tone in accordance to the pre-focal high tones (H-) observed by [11]. Since [8] describes a rise on (H*) in the immediately pre-focal position it was expected to be the result of the introduction of a boundary tone and not an increased pitch accent aligned to a ω stressed syllable. The previously

observed *immediately pre-focal rise* and PFC [8] were expected to be the results of boundary insertion and/or deletion to satisfy focus alignment.

3. Method

This production experiment is a modified replication of [8]. Both studies adopt the methodology of [18] to elicit in-situ focus on different constituents.

3.1. Stimuli

Five target sentences with a simple SOV structure including an accusative object were designed. Each target contained the same number of comparable segments: a three syllabic subject, a four syllabic object, and a three syllabic verb. To elicit the introduction of boundary tones and the status of (H*) vs (H-), only subjects and objects with non-final lexical word stress were considered to avoid that pitch accents and boundary tones would fall on the same segment. Each sentence was elicited as a contrastive focus condition, i.e. when the interpretation of a linguistic expression is limited to a set of contrasting alternatives. Targets were preceded by a question eliciting contrastive focus on either the subject, the object, or the verb. As a baseline, a broad focus condition was elicited for all targets. The contrastive constituent of the targets was always presented as the first alternative in the preceding questions. The elicitation of contrastive in-situ focus was furthermore supported by the focus sensitive question particle -mı attaching to the contrastively focused alternatives in the preceding questions. The morphological marker should help to avoid the triggering of a default reading prosody in the following answer and reduce errors concerning the focused constituent. Additionally, the focused items were presented with underlining to avoid prosodic pattern repetitions since the question-answer pairs were presented without fillers. In (1) a list of the target sentences is provided. In (2) the question-answer pairs are exemplarily provided for target sentence (a).

(1) *Target sentences.*

(a) Fahıre Naca'sını seviyor.
Fahıre Nacı-POSS-ACC love-PRS(3SG)
Fahıre loves her Nacı.

(b) Nasrettin babasını üzüyor.
Nasrettin father-POSS-ACC sadden-PRS(3SG)
Nasrettin saddens his father.

(c) Macide kardeşini çiziyor.
Macide sibling-POSS-ACC draw-PRS(3SG)
Macide draws her sibling.

(d) Nadide ablasını özlüyor.
Nadide sister-POSS-ACC miss-PRS(3SG)
Nadide misses her sister.

(e) Yasemin aynacıyı dinliyor.
Yasemin mirror dealer-ACC listen-PRS(3SG)
Yasemin listens to the mirror dealer.

(2) *Contrastive focus questions preceding target (a).*

(broad) Ne oluyor ?
What be-PRS(3SG)
What happens?

(subject) Fahire mi Naci'sini seviyor, Meral mı?
Fahire Q Naci-POSS-ACC love-PRS(3SG) Meral Q
Does Fahire love Naci or Meral?

(object) Fahire Naci'sini mı seviyor, polisi mi?
Fahire Naci-POSS-ACC Q love-PRS(3SG) police man Q
Does Fahire love Naci or the police man?

(verb) Fahire Naci'sini seviyor mu, üzüyor mu?
Fahire Naci-POSS-ACC love-PRS(3SG) Q sadden-PRS(3SG) Q
Does Naci love or sadden Naci?

Each pair of target sentence and preceding question was presented subsequently in its four focus conditions. No repetitions were made. 140 utterances were recorded: 7 speakers x 5 sentences x 4 foci.

3.2. Subjects

Seven native Turkish speakers, four females and three males, from Ege Universitesi in Izmir participated in the experiment. At recording time they were aged between 20 and 27, had no previous specific linguistic knowledge and no speaking or hearing disabilities. All of them were monolingual speakers; most of them had some basic foreign language skills in a second language.

3.3. Recording process

The data were recorded in a translation laboratory. The target sentences and preceding questions were presented on a power point slide. The subjects were asked to read out aloud first the question and subsequently the answer. The interviewer was a native Turkish speaker who was schooled to ask the participant to repeat the question-answer pairs in cases of discrepancies concerning the understanding of the task.

4. Analyses and results

4.1. Phonologic analysis

A phonological annotation of 135 target sentences was done by the author and double checked by a native Turkish speaker. Five utterances had to be excluded from the analyses due to creeky voice, slip of the tongue or hesitation. The phonological annotation of the *f0* contour among the different focus conditions of the whole ιp was done for each target sentence and speaker. All target sentences were manually segmented in Praat on the syllable level. Supra-segmental labeling basically follows [11], but is adapted and modified with respect to the features observed in the data using general ToBi labeling advices, since Turkish lacks a conventionalized annotation system. Special attention was paid to boundary tone insertion/ deletion under the changing foci.

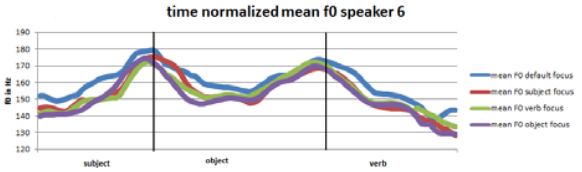
The descriptive phonological analysis revealed that all speakers of the production experiment implemented a high tone (H-) on the ultimate syllable of the subject and a further high tone (H-) on the following object in all broad focus and object focus sentences. In addition to the high tones on the last syllable of each non-final ω, some speakers arbitrarily implemented pitch accents on the word stressed syllables of the lexically accented words. A low final boundary tone (L%) was aligned with the ι-phrase final verb. Figure 2 demonstrates the broad focus intonation contour and phrasing of Turkish SOV declaratives for all speakers as observed in the data.

Figure 2. *Broad focus intonation contour in SOV*

(H*) H- (subject)φ (H*) H- (object)φ L% (verb)φ

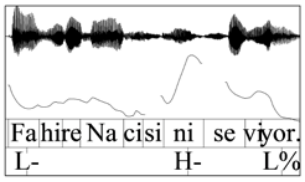
Concerning the tonal modification corresponding to different focus conditions, two groups were identified. For group A (speakers 1, 2, 3, 5, 6) the observed tonal implementation as shown in Figure 2 was not modified by information structure. Neither a categorical modification of the described high tones, nor the insertion or deletion of tonal boundaries was observed in the different focus conditions. Figure 3 displays the time normalized f0 of the realization of the four focus conditions of one male speaker averaged across all target sentences.

Figure 3. *Mean f0 contour in four focus conditions.*

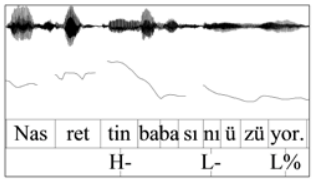


For group B (speakers 4 and 7) f0 modification was observed in different focus conditions. Like the remaining speakers they implemented a high tone (H-) on the rightmost syllable of each non-final ω. For subject focus both speakers continued implementing H- on the subject, but deleted the following H- of the object. De-accentuation continued until the ι-phrase final low boundary tone (L%). Speaker 7 used de-accentuation only in four of the five subject focus conditions. For verb focus they deleted or compressed the high tone on the subject, but implemented (H-) on the object. The verb was always aligned with (L%). Examples 1 and 2 show the modified f0 contour in verb and subject focus of group B.

Example 1. *F0 of verb focus in target sentence (1).*



Example 2. *F0 of subject focus in target sentence (2).*



4.2. Phonetic analysis

Due to the group dependent results of the phonological description, the analysis was amplified by a phonetic analysis. Maximum f0 of (H-) and break introduction were measured for each speaker and focus condition. Since group A showed no phonological modification of the f0 contour, it was tested, if the modification was phonetic in terms of a gradient modification. The test concerned the question whether break introduction and/ or a gradient modification of (H-) indicate a change in phrasing by establishing prosodic boundaries to align focused constituents, as stated in the hypothesis. The phonetic analysis was done time-normalized by the introduction of 10 measure points per syllable using a general purpose PRAAT script Prosody Pro [19]. The script enables comparison across all target sentences, focus conditions, and speakers. Each sentence was corrected manually with respect to spurious pitch values.

In a first step, the maximum f0 values were averaged across the five target sentences and four focus conditions for each speaker. The mean maximum f0 values on the high tones of the subject and the object of each focus condition and sentence were compared with each speaker to test if the downstep relation refers to focus alignment in the sense of a change in phrasing indicated by a reset. A reset by means of tonal upstep was considered in the case that the high tone of the second constituent was higher than the high tone of the preceding constituent. The maximum f0 analysis showed speaker-dependent variation. Downstep on the second high tone was not implemented in 34 sentences. A correlation between focus and upstep was only found for the speakers of group A. For verb focus, the maximum f0 on the objects excelled the maximum f0 of the preceding subject in all target sentences for both speakers corresponding to the tonal deletion/ compression observed in the phonological analyses. Nonetheless, upstep was also observed in other focus conditions for both speakers. For the remaining speakers, no clear relation between upstepped high tones and focus conditions was found and the tonal values of subsequent constituents were rather arbitrarily implemented and not obligatorily following downstep. Table (1) summarizes the upsteps implemented by all speakers.

Table 1. *Upsteps per speaker and focus condition.*

Upsteps according to focus condition								
speaker	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
amount	7(18)	7(19)	0(20)	6(20)	3(19)	2(19)	9(20)	34 (135)
default	1	2	0	1	0	1	2	**5**
subject focus	2	2	0	0	0	0	1	**4**
object focus	0	1	0	0	1	1	1	**3**
verb focus	**4**	2	0	**5**	2	0	**5**	**18**

In a second step, all target sentences were tested concerning break implementation. The test concerned the question whether prosodic breaks may serve as an indicator for focus prominence and focus induced change in phrasing in the sense of pre-focal break introduction like in French [4]. Break introduction showed speaker dependent variation. Only speaker 4 used considerable breaks of around 0,05s. The other speakers rather showed a tendency to final lengthening. Speaker 3, 5, and 6 used none to one break. The breaks implemented by speakers 1, 2 and 7 were mainly used after the subject, mapping syntactical phrasing. Sometimes speaker 1 and 2 used additional breaks after the object. There was no straightforward implementation of the breaks after the subject and/ or object and no correlation to focus for speaker 1, 2 and 7. Speaker 4 implemented a break after the subject in all target sentences, unless in verb focus. In verb focus she used a pre-

focal break after the object and deleted the preceding break after the subject. Figure 4 shows how prosodic breaks assisted syntactic or prosodic phrasing in the data of this study.

Figure 4. *Prosodic break implementation*

Syntactic phrasing:	(S)[break] (OV)
Prosodic phrasing:	(S) [break] (O) [break] (V)

5. Discussion and conclusion

The phonological description and the phonetic analysis of information structurally modified target sentences only partly fulfill the initial hypotheses.

The analyses of broad focus sentences offer a new contribution for the general description of the tonal structure in Turkish. The data revealed that each non-final ω in simple SOV sentences bears a high tone on the ultimate syllable. This high tone cannot be interpreted by means of general pitch accent implementation as in previous tonal descriptions, since it is not aligned to metrically strong syllables. The word stressed syllables were only additionally and arbitrarily aligned to less prominent pitch accents indicating that the prosodic structure in Turkish is primarily based on prosodic phrasing and less on the notion of pitch accents. From the present data the high tone on the last syllables of non-final constituents is interpreted as a high phrase tone (H-) aligned to the right boundary of each non-final ω. Consequently, Turkish is interpreted as a phrase language. It's prosodic structure is characterized by radical splitting, where each constituent forms its own φ and the function of H- is interpreted as delimiting φ on a supposed φ level. The concept of phrase language established by [5] is extended here to the inclusion of pitch accents aligned to word stressed syllables (Figure 2).

The analyses of information structure modified sentences revealed two groups in the prosodic realization of focus in Turkish. The phonological annotation and the acoustic measurements of maximum f0 and break implementation showed that group A did not change the tonal implementation and prosodic phrasing structure as observed for broad focus. For group B the study revealed that speakers change the prosodic structure according to focus. Whereas in the broad focus condition the same f0 contour was observed as for group A, group B implemented (H-) on the subject for subject focus and deleted the following high phrase tone to its right, approving the observation of PFC by [8]. For verb focus, the pre-focal (H-) on the subject was deleted or compressed and only the high tone on the object was implemented. The observed tonal deletion/ compression shows a change in phrasing according which is interpreted as an indicator for the prosodic alignment of focus. In subject and verb focus the broad focus phrasing structure as observed in the study and characterized by radical splitting changes. The focused constituent is separately wrapped into its own φ whereas the remaining given constituents are radically wrapped into a one further φ as long as they are adjacent. For object focus radical splitting was also maintained for subject and verb by group B, since the given elements were not adjacent and the focused constituent is claimed to be separately wrapped into an independent φ.

The variation of break implementation in the data show that prosodic breaks can be used for the individual organization of utterances on a supra-segmental level reflecting segmental and supra-segmental properties of the language. Hence, break implementation can assist phrasing, but has to be accompanied by tonal features to assist focus alignment and does not indicate focus on its own.

The acoustic analysis of maximum f0 on focused and un-focused subjects and objects revealed that a register change by means of upstep indicating a reset to establish a new φ is not related to information structure. Furthermore, the speaker-dependent variation in the data motivate the assumption that downstep is not systematic in Turkish.

Based on the group specific observations in the prosodic realization of contrastive in-situ focus in the present study it is assumed that prosodic focus marking in Turkish is optional. Turkish does not require a prosodic focus marking per se, but focused constituents are optionally aligned with prosodic boundaries. It is assumed that a preceding context which already sufficiently identifies the focused constituent overrides the need for a prosodic focus marking. Nonetheless, a prosodic representation of information structure was observed in the pilot, which can successfully be described by prosodic alignment. I thus interpret the results suggesting that Turkish can be classified as a boundary language in the framework of focus typology, and as a phrase language such as Hindi or French [4, 5] in the notions of general language classification. Figure 5 summarizes the general phrasing structure in Turkish as observed in the data and its modification according to the contrastive focus condition.

Figure 5. *Prosodic alignment in Turkish*.

Broad/object focus phrasing:	(S)φ ($O_{(F)}$)φ (V)φ.
Subject focus phrasing:	(**S_F**)φ(OV)φ.
Verb focus phrasing:	(SO)φ(**V_F**)φ.

Concluding, certain limitations concerning the data interpretation have to be made: (i) the alignment approach is based on the observations of a small data set and needs to be confirmed on the bases of more data. (ii) The observed phrasing structure is based on an analysis of simple SOV declaratives. Further research on more complex syntactic representations has to be done to confirm the implementation of the supposed phrase delimiting high tones. (iii) In order to access the reliability of the production data and the resulting claim of focus alignment, a perception test is necessary.

6. References

[1] Büring, D. 2010. "Towards a typology of focus realization", in M. Zimmermann and C. Féry [Eds], Information Structure. Theoretical, Typological and Experimental Perspectives. Oxford: Oxford University Press: 177-205.

[2] Erguvanlı, E. 1984. The function of word order in Turkish grammar. Berkeley: University of California Publications.
[3] Féry, C. 2013a. “Focus as prosodic alignment”, in Natural Language and Linguistic Theory 31(3) 683–734.
[4] Féry, C. 2013b. Final compression in French as a phrasal phenomenon. University of Frankfurt. Online: http://web.uni-frankfurt.de/fb10/fery/publications/French-PFC final % 20 copy .pdf
[5] Féry, C. 2010. “The intonation of Indian languages: an areal phenomenon”, in I. Hasnain and S. Chaudhury [Eds], Problematizing Language Studies. Festschrift for Rama Agnihotri. Delhi: Aakar Books: 288-312.
[6] Göksel, A. and Özsoy, S. 2000: “Is there a focus position in Turkish?”, in Studies on Turkish and Turkic Languages; Proceedings of the 9th International Conference on Turkish Linguistics. Wiesbaden: Harrassowitz : 219–228.
[7] Güneş, G. 2013. “Limit on syntax- prosody mapping in Turkish prosody of finite and non-finite clausal parentheticals”, in E. Erguvanli Taylan [Ed], Dilbilim Araştırmaları Dergisi. Istanbul: Boğaziçi University Press.
[8] İpek, C. 2011. “Phonetic realization of focus with no on-focus pitch range expansion in Turkish”, in Proceedings of ICPhS XVI. Hong Kong.
[9] İşsever, S. 2003. “Information Structure in Turkish: The word order-prosody interface”, in Lingua (113) 1025–1053.
[10] Jackendoff, R. 1972. Semantic interpretation in generative grammar. Cambridge Mass.: MIT Press.
[11] Kan, S. 2009. Prosodic domains and the syntax-prosody mapping in Turkish. Master’s thesis. Bogazici University Istanbul.
[12] Kornfilt, J. 1979. Turkish. New York: Routledge.
[13] McCarthy, J. and Prince, A. 1993. “Prosodic morphology”, in J. Goldsmith [Ed], Handbook of phonology. Oxford: Blackwell: 318-366.
[14] Nespor, M. & Vogel, I. 1986. Prosodic Phonology. Dordrecht: Foris.
[15] Özge, U. & Bozşahin, C. 2010. “Intonation in the grammar of Turkish”, in Lingua (120) 132-175.
[16] Selkirk, Elisabeth O. 1986. “On derived domains in sentence phonology”, in Phonology Yearbook 3: 371–405.
[17] Truckenbrodt, H. 1995. Phonological phrases: their relation to syntax, focus & prominence. Doctoral Dissertation. Cambridge, Mass.: MIT Press.
[18] Xu, Y. 1999. “Effects of Tone and Focus on the Formation and Alignment of F0 Contours”, in Journal of Phonetics (27) 55–105.
[19] Xu, Y. 2013. “ProsodyPro- A Tool for Large-scale Systematic Prosody Analysis.” in Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France. 7-10.

Perceptual evaluation of the effect of mismatched Fujisaki model commands and surface tone in Sesotho

Lehlohonolo Mohasi [1], *Hansjörg Mixdorff*[2], *Thomas Niesler* [1]

[1] Department of Electrical & Electronic Engineering, University of Stellenbosch, South Africa
[2] Department of Computer Science and Media, Beuth University Berlin, Germany
lmohasi@sun.ac.za, mixdorff@beuth-hochschule.de, trn@sun.ac.za

Abstract

Sesotho is a tonal Southern Bantu language which has so far received extremely little attention by the speech research community. We consider tone modelling for Sesotho using the Fujisaki model-based analysis with a view to the development of a text-to-speech (TTS) system. Fujisaki analysis can be used to indicate the tone associated with a syllable, but it often differs from the surface tone that would be available for TTS synthesis. We investigate instances in which the surface tone differs from the tone indicated by Fujisaki analysis, and determine the effect of these discrepancies on speech quality. The amplitude of Fujisaki tone commands is manipulated to match the surface tones, and the resulting resynthesized speech subsequently analysed by perceptual tests. We find that the effect of inserting tone commands at high surface tone syllables is more severe than matching the Fujisaki tone commands with low surface tone syllables, in terms of naturalness. Furthermore, some discrepancies can be attributed to errors in the surface tonal transcription. However, on average, all manipulations lead only to a mild degradation in speech quality. We conclude that the Fujisaki model is a feasible way to model tone in Sesotho even in the presence of limited and under-developed linguistic resources.

Index Terms: Fujisaki model, Sesotho, surface tonal transcription, text-to-speech synthesis

1. Introduction

Accurate prosodic modelling is crucial for natural-sounding text-to-speech (TTS) systems and can be achieved by correct modelling of pitch in tonal languages. For example, Ekpenyong et al. [1] found that the use of tone marking contributes significantly to the quality of synthetic Ibibio speech. In Sesotho, tone is not marked in the orthography, but must be deduced by the process of surface tonal transcription. This relies on morphological analysis, a tone-marked pronunciation dictionary, and a set of tonal rules. Each one of these three components can introduce errors, for instance, Schadeberg [2] and Roux [3] have pointed out the inconsistency of Sesotho tone-marked dictionaries.

The Fujisaki model is a tractable and powerful tool for prosody manipulation that has proven to be effective for modelling fundamental frequency (F0) contours. Its validity has been tested for several languages [4, 5, 6, 7] some of which are tonal languages such as Thai [8] and Mandarin [9]. It has been shown that the tone of a syllable can be represented in the Fujisaki model as a tone command, which is a pulse indicating where the rising and falling of F0 occurs.

Sesotho uses a register tone system with two tones, high and low. The Fujisaki model and surface tonal transcription have shown a strong agreement on the interpretation of tone in Sesotho sentences [10]. The purpose of this paper is to identify, investigate and perceptually evaluate those cases where the Fujisaki tone commands and surface tone labels do not agree. We want to determine if the mismatch between the actual F0 realisations and the surface tones seriously degrade the prosodic quality.

In modelling Sesotho prosody, certain phenomena must be taken into consideration. Tone sandhi is a phonological change occurring in tonal languages, in which tones occur in combination with other tones. It is present in Sesotho as pointed out by Demuth [11], and is modelled by the surface tonal rules HTD (spreading of a lexical high tone to the immediate right syllable) and IHTS (iterative spreading of a high tone to the end of a verb). Other tonal rules, however, do not model tone sandhi. Ekpenyong and Udoh [12] emphasise the importance of tone sandhi and its effect on the overall F0 contour in tone languages.

The Obligatory Contour Principle (OCP) is a phenomenon where adjacent identical tone elements are prohibited. According to Yip [13], OCP violations can be avoided in a variety of ways, such as tone deletion, blocking of spreading if it leads to adjacency, or fusion between tones. However, OCP in some cases is violable [13]. In Sesotho surface tonal transcription, OCP is observed via the RBD (dissociating the immediate right branch of a multiply-linked high tone syllable if, and only if, there is a high tone syllable immediately after the target of the HTS rule), LBD (delinking the immediate preceding left branch of a multiply-linked high tone syllable if it is preceded by a high tone syllable) and FR (exempting syllables at the end of a phonological phrase from the application of tonal rules) rules [14]. In contrast, OCP is not fully observed by the Fujisaki analysis and this is evident in the production of prolonged tone commands instead of exclusively single-syllable tone commands. Peak delay is observed when the F0 peak corresponding to a high-toned syllable occurs in the following syllable [15, 13], while anticipation is observed when a high tone is realised on the preceding syllable [16]. The former aspect is modelled in the surface tonal transcription via the HTS1, IHTS and GTI rules. The latter feature is not modelled in surface tone transcription. Both are observed in the Fujisaki analysis.

2. Data preparation

The following sections describe the compilation, preparation, annotation, and selection of the data on which our experiments are based.

2.1. Experimental data

Our data is drawn from a corpus that is based on a set of weather forecast bulletins obtained from the weather bureau in Lesotho. The corpus comprises a total of 256 sentences with an average of 23 words per sentence, and a total of 51 minutes of speech. The original audio data had a poor signal-to-noise ratio, making it unsuitable for use in TTS development. For this reason, the sentences were re-recorded by the first author, who is a female native speaker of Sesotho. Recording was performed in a quiet studio environment using a large membrane SHURE KSM32SL microphone. All recordings were made at a sampling rate of 48kHz.

For tone modelling in African tonal languages, in which tone is not indicated by the orthography, an algorithm that predicts the tonal labels of syllables in a word is a prerequisite [17]. As a starting point, we compiled a suitable domain dictionary for weather forecasts using two published tone-marked dictionaries – a Sesotho dictionary by Du Plessis et al. [18], and a Northern Sotho dictionary by Kriel and van Wyk [19]. The dictionaries contain lexical tones for each word.

Next, the sentences in our corpus were annotated with underlying tones from the dictionary. From this underlying tone transcription, a surface tone transcription was deduced by means of a morphological analysis as well as the tonal rules described in [14].

On completion of the surface tonal transcription, the sentences were annotated at word and syllable levels using Praat [20]. F0 values were extracted at a step of 10ms and inspected for errors. The F0 tracks were subsequently decomposed into their Fujisaki model components applying an automatic method originally developed for German [21]. Initial experiments in [22] had shown that the low tones in the critical words of the minimal pairs could be modelled with good accuracy without employing negative tone commands. Consequently, high tones were associated with Fujisaki tone commands with positive amplitude, while low tones had no associated tone command (i.e. zero amplitude).

2.2. Selection of mismatched syllables

First we selected from our data cases in which a Fujisaki tone command corresponds to one or more low surface tones. In general, the Fujisaki tone command spanned more than one syllable. Table 1 classifies the pattern of surface tones associated with each such tone command. Each low surface tone indicates a discrepancy. With this in mind, we selected cases in which a Fujisaki tone command coincided with one or more low surface tones. In total, 63 such cases were isolated from our data.

We also considered cases in which a high surface tone was not accompanied by a Fujisaki tone command. Examples of 1, 2, 3 and 4-syllable sequences with high surface tones but no corresponding Fujisaki tone command were identified in the data, and are listed in Table 2. In total, 64 such cases were isolated from our data.

Table 1: *Instances in which Fujisaki tone commands correspond to low surface tones (FHSL).*

Surface tone pattern	Number of syllables	Number of cases
Alternating, e.g. LHLHL	≥ 2	40
Low surface tone labels only, e.g. LLL	≥ 1	7
Any other combination, e.g. LHHL	≥ 2	16
TOTAL		63

Table 2: *Instances in which high surface tones do not correspond to Fujisaki tone commands (FLSH).*

Number of syllables	Number of cases
1	26
2	22
3	13
4	3
TOTAL	64

3. Pitch contour modifications

In the previous section, two types of mismatches between the Fujisaki model and the surface tone labels were identified. In this section, these mismatches are "resolved" by either inserting or removing the Fujisaki tone command in order to match the predicted surface tones.

3.1. Case FHSL: High Fujisaki tone associated with low surface tone

For cases in which the Fujisaki tone command coincided with a syllable with a low surface tone, the amplitude of the tone command was set to zero. Figure 1 illustrates this modification.

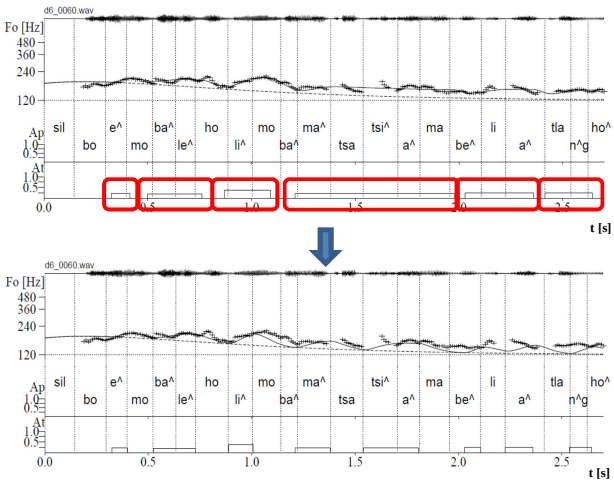


Figure 1: *Modifying Fujisaki tone commands to coincide with low surface tone syllables. The top panel shows an original utterance with surface tone patterns in the following order: HL, LHHL, LHL, HHLHHLH, HLH, and LHH. The bottom panel illustrates a change in pattern after modification. The phrase reads "Boemo ba leholimo ba matsatsi a mabeli a tlang ho …" – "The weather in the next two days …"*

When the Fujisaki tone command corresponded to a sequence of syllables with both high and low surface tones, the amplitude of the tone command was set to zero only for low surface tone syllables. The onset t_1 and/or offset t_2 times were unchanged for high surface tone syllables. In all other cases, t_1 and t_2 coincided with syllable boundaries. The optimal alignment of tone command onsets and offsets appears to be language-dependent [8, 23], and its determination for Sesotho remains the subject of ongoing work.

3.2. Case FLSH: High surface tone associated with low Fujisaki tone

In the case of high surface tone syllables with no Fujisaki tone command, tone commands with average amplitude were inserted and t_1 and t_2 were set to syllable boundaries. Sixty-four modified phrases were generated, in line with Table 2. Figure 2 illustrates an example of the modification, where the top panel shows the original utterance, and the bottom one displays the prosodic group generated (indicated by a rounded rectangle) to coincide with high surface tone markings.

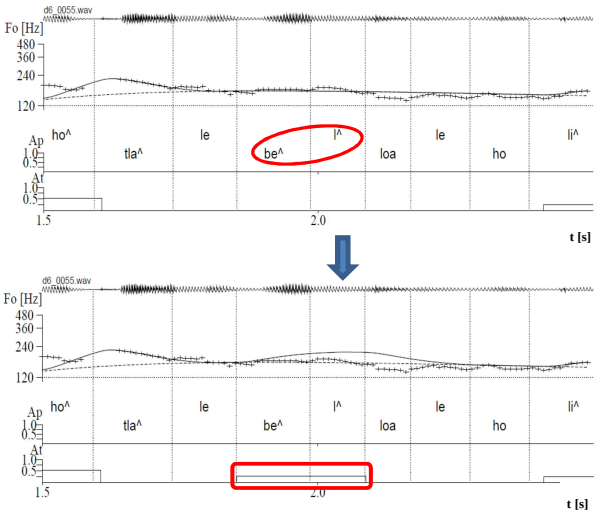


Figure 2: *Inserting a tone command. The top panel indicates a sequence of two high surface tone syllables, be^ and l^, with no Fujisaki tone command. In the bottom panel, a tone command is created for these syllables. The partial phrase reads "ho tla lebelloa leholi…" – "the sky will be expected …"*

4. Perceptual Evaluation

Perceptual evaluation of the modified phrases was carried out by twenty-one Sesotho speakers (7 females, 14 males). All subjects are native Sesotho speakers from Lesotho, and are students at the University of Stellenbosch. The classification of the data and the evaluation process are detailed below.

4.1. Data

For each group in Section 3.1, modifications were applied to the Fujisaki tone commands, after which the utterance was resynthesized using the PSOLA-based Praat ManipulationEditor (20) [19]. Tone commands which were inserted in Section 3.2 were also resynthesized in a similar manner. The resulting phrases were collected for perceptual testing.

Table 3: *Data used for perceptual evaluation.*

Modification	Modified phrases	Unmodified phrases	Total
FHSL: Fujisaki tone command removed	63	6	69
FLSH: Fujisaki tone command inserted	64	6	70

The data for evaluation included a number of unmodified phrases to serve as a baseline. Table 3 gives a summary of the data for the two modifications and groups.

4.2. The evaluation process

DMDX [24] was used to perform all perceptual evaluations. Subjects listened to the utterances in a quiet room using a headset. Evaluation of the data was based on a rating scale intended to reflect naturalness, as shown in Figure 3. Subjects were asked to rate each individual audio file according to this scale.

An informal SUS intelligibility test [25] of the modified phrases was also performed. Words from the modified and unmodified phrases were randomly selected to form semantically unpredictable sentences (Table 4). The sentences generated were each five words long. These were then resynthesised and the listener requested to indicate what they heard. Due to time constraints, this was performed by only one native Sesotho speaker.

1	Sounds like a mother–tongue speaker
2	
3	Sounds almost like a mother–tongue speaker
4	
5	Definitely not a mother–tongue speaker
6	
7	Disturbingly unnatural speech, hard to understand
8	Don't know

Figure 3: *The rating scale used for perceptual evaluation.*

Table 4: *Data used for the SUS intelligibility test.*

	FHSL	FLSH	Unmodified
Number of words	106	115	108
Number of sentences	21	23	22

5. Results

Phrases rated with an 8 ("Don't know") were excluded from the following analysis. In other cases, average scores according to the scale in Figure 3 have been considered.

Figure 4 illustrates the overall average perceptual scores resulting from the two types of modification described in Section 3. Unmodified phrases are perceived to be most natural, although interestingly they were usually not awarded a score of "1". The figure also shows that, on average, removal of a Fujisaki tone command is slightly less detrimental to perceived quality than the insertion of a tone command. However, even in the latter case, the perceived score is just above 2, neighbouring "almost mother tongue".

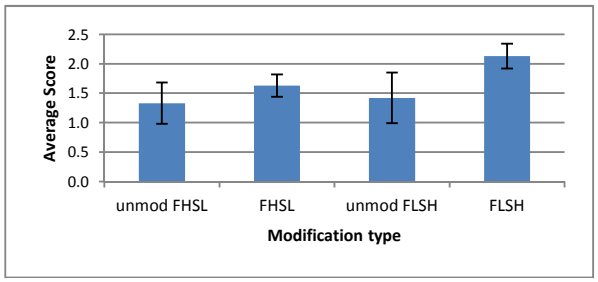


Figure 4: *Overall average perceptual scores for FHSL and FLSH modifications. Vertical bars denote 95% confidence intervals.*

From the FHSL data described in Table 1, we isolated a subset consisting of cases where a Fujisaki tone command was associated with a sequence of between 1 and 3 low surface tone syllables. This data is summarised in Table 5, while Figure 5 shows the corresponding results of the perceptual evaluation. For each case, the Fujisaki tone command amplitude was set to zero for one, two, or three consecutive syllables. The results show that 1 and 2 syllable modifications are rated similarly. Although 3-syllable modifications show larger degradation, the reliability of this average is low due to the small number of samples (4).

Table 5: *Instances in which Fujisaki tone commands correspond to consecutive low surface tone syllables.*

Consecutive low tone syllables	Number of cases
1	20
2	8
3	4
TOTAL	32

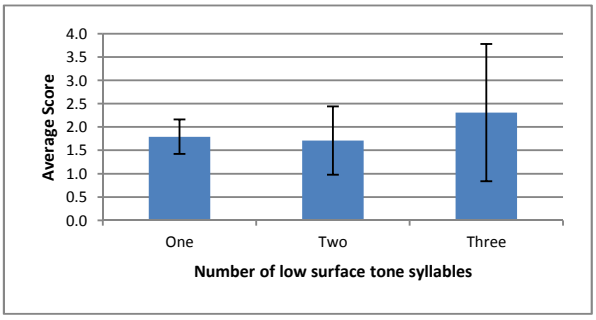


Figure 5: *Average perceptual scores when removing the Fujisaki tone command associated with 1, 2, and 3 consecutive low surface tones. Vertical bars denote 95% confidence intervals.*

Figure 6 shows a similar analysis for the FLSH data described in Table 2. The results are similar across number of syllables.

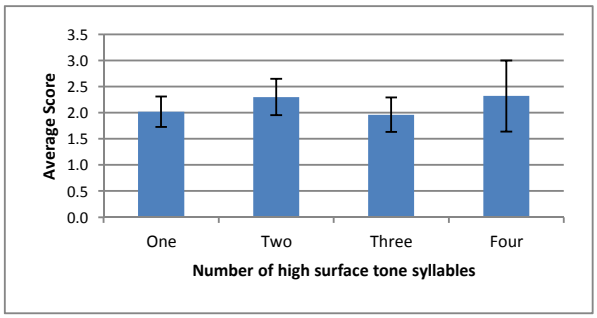


Figure 6: *Average perceptual scores when inserting Fujisaki tone commands for 1, 2, 3, and 4 consecutive high surface tone syllables. Vertical bars denote 95% confidence intervals.*

Table 6 shows the results from the informal intelligibility test described in Section 4.2. The intelligibility score is based on the number of words identified correctly in each sentence. Sentences composed of unmodified words have the highest intelligibility score, with the lowest score obtained for FHSL modifications. Removal of the Fujisaki tone command therefore appears to have a much higher detrimental effect on intelligibility than its insertion.

Table 6: *Intelligibility scores for FHSL, FLSH, and unmodified utterances.*

Modification	Intelligibility Score [%]
FHSL	38.1
FLSH	60.9
Unmodified	77.3

Finally, each of the 127 mismatches described in Tables 1 and 2 was considered individually in order to determine the source of the discrepancy. Tables 7 and 8 describe the results of this investigation. The totals exceed the values in Tables 1 and 2 because each mismatch can be due to more than one factor.

The tables show that tone sandhi, OCP and peak delay are major contributors of mismatches, while incorrect dictionary entries are less so. One phenomenon not yet taken into consideration in the deployment of both methods is downstep, which will be explored in future work.

Figure 7 illustrates the perceptual score due to the mismatch effect by each phenomenon on naturalness. Overall, the discrepancies where the Fujisaki tone command was inserted significantly affect naturalness than when the tone command was removed.

Table 7: *Mismatches in the FHSL case.*

Description	Number of instances
Tone sandhi is observed by the Fujisaki analysis but not by the relevant surface tonal rules.	33
Surface tone transcription observes OCP but Fujisaki analysis does not.	38
Peak delay is observed by the Fujisaki analysis but not by the relevant surface tone transcription rules. (The FR rule observes peak delay only for relative verbs and for ultimate syllables whose lexical tone is high.)	18
Anticipation is observed by the Fujisaki analysis but it is not modelled by the surface tone transcription rules.	19
Incorrect dictionary tone	10
Unresolved	5
TOTAL	123

Table 8: *Mismatches in the FLSH case.*

Description	Number of instances
Tone sandhi is observed by the surface tonal transcription but not by the Fujisaki analysis.	40
OCP is violated in the surface tonal transcription by adjacent high-tone syllables, but not violated by the Fujisaki analysis.	38
The Fujisaki analysis does not observe peak delay while the surface tonal transcription does.	23
Incorrect dictionary tone	18
Unresolved	5
TOTAL	124

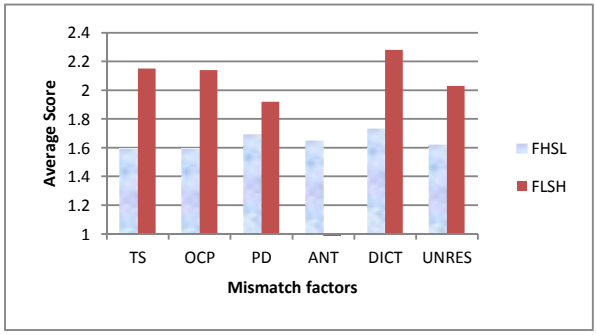


Figure 7: *Sources of mismatch between surface tone and Fujisaki tone commands, and their effect on naturalness. TS = tone sandhi, PD = peak delay, ANT = anticipation, DICT = incorrect tone in dictionary, and UNRES = unresolved cases.*

6. Conclusions

When considering TTS for a very poorly resourced language, such as Sesotho, the reliance of imperfect resources, including dictionaries, morphological analyses and tonal rules, is inescapable. In this paper we have performed experiments to determine the effect of these mismatches on the perceived quality or resynthesized speech. We find that, overall, the modifications applied to "rectify" the mismatch lead only to mild degradation in the perceived quality of the speech. From this we conclude that Sesotho TTS based on the Fujisaki model for tonal and prosodic modelling is feasible, even when based on imprecise resources. An analysis of the sources of the discrepancies indicates scores of error and that much can be gained by the improvement of the surface tone transcription process.

Acknowledgements - This work was supported in part by the National Research Foundation of South Africa (grant UID 71926), by a DFG International collaboration grant (Mi 625/16-1), and by Telkom South Africa. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the sponsors.

7. References

[1] Ekpenyong, M.E. et al. *Statisitical parametric speech synthesis for Ibibio.* Speech Communication, Vol. 56, pp. 243-251, 2014.

[2] Schadeberg, T. *Tone in South African Bantu languages.* Journal of African Languages and Linguistics, Vol. 3, pp. 175-180, 1981.

[3] Roux, J. On the perception and description of tone in the Sotho and Nguni languages. Kaji Shigeki [ed.]. *Proceedings of the symposium cross-linguistic studies of tonal phenomena. Historical Development, Phonetics of Tone and Descriptive Studies.* Tokyo : Tokyo University of Foreign Studies, ILCAA, 2003.

[4] Narusawa, et al. *A method for automatic extraction of model parameters from fundamental frequency contours of speech.* Orlando, Florida, USA : Proceedings of ICASSP, 2002.

[5] Rossi, P.S., Palmieri, F., and Cutugno, F. *Inversion of F0 model for natural-sounding speech synthesis.* Hong Kong, China : Proceedings of IEEE ICASSP, 2003.

[6] Moberg, M. & Parssnen, K. *Comparing CART and Fujisaki intonation models for synthesis of US-English names.* Nara, Japan : Proceedings of Speech Prosody, 2004.

[7] Aguero, P.D. *Automatic analysis and synthesis of Fujisaki's intonation model for TTS.* Nara, Japan : Proceedings of Speech Prosody, 2004.

[8] Mixdorff, H., Luksaneeyanawin, S., Fujisaki, H. & Charnvivit, P. *Perception of tone and vowel quantity in Thai.* Denver, Colorado, USA : Proceedings of ICSLP, 2002.

[9] Mixdorff, H., Hu, Y. & Chen, G. *Towards the automatic extraction of Fujisaki model parameters for Mandarin.* Geneva, Switzerland : Proceedings of Eurospeech, 2003.

[10] Mohasi, L., Mixdorff, H. & Niesler, T. *Characterisation of prosodic groups in Sesotho using the Fujisaki model.* Journal of Chinese Linguistics Monograph, 2013.

[11] Demuth, K. *Issues in the acquisition of the Sesotho tonal system.* Journal of Child Language, Vol. 20, pp. 275-301, 1993.

[12] Ekpenyong, M.E. and Udoh, E-O. *Intelligent prosody modelling: A framework for tone language synthesis.*

[13] Yip, M. *Tone.* Cambridge University Press, 2002.

[14] Khoali, B.T. A Sesotho Tonal Grammar. *PhD Thesis.* University of Illinois at Urbana-Champaign, 1991.

[15] Myers, S. *Tone association and F0 timing in Chichewa.* Studies in African Linguistics, Vol. 28 (2), pp. 215-239, 1999.

[16] Hyman, L.M. Universals of tone rules: 30 years later. Thomas and Gussenhoven, Carlos Riad [ed.]. *Typological Studies in Word and Sentence Prosody.* Mouton de Gruyter, Vol. 1, pp. 1-34, 2007.

[17] Louw, J.A., Davel, M. & Barnard, E. *A general-purpose isiZulu speech synthesizer.* South African Journal of African Languages, Vol. 25, pp. 92-100, 2005.

[18] Du Plessis, J.A. et al. *Tweetalige Woordeboek Afrikaans-Suid-Sotho.* Kaapstad : Via Afrika Bpk, 1974.

[19] Kriel, T.J. & van Wyk, E.B. *Pukuntsu Woordeboek Noord Sotho-Afrikaans.* 4th. Pretoria : Van Schaik, 1989.

[20] Boersma, P *Praat - A system for doing phonetics by computer*. 9/10, Glot International, Vol. 5, pp. 341-345, 2001.

[21] Mixdorff, H. Intonation Patterns of German - Model-based Quantitative Analysis and Synthesis of F0 Contours. *PhD Thesis.* TU Dresden, 1998.

[22] Mohasi, L., Mixdorff, H. & Niesler, T. *An acoustic analysis of tone in Sesotho.* Hong Kong, China : Proceedings of ICPhS XVII, 2011.

[23] Mixdorff, H. and Barbosa, P.A *Alignment of intonational events is German and Brazilian Portuguese - a quantitative study*. Shanghai, China : Proceedings of Speech Prosody, 2012.

[24] Foster, K.I. & Foster, J.C. *DMDX: A Windows display program with millisecond accuracy.* Behavior Research Methods, Instruments, and Computers: A Journal of the Psychonomic Society, Inc, Vol. 35 (1), pp. 116-124, 2003.

[25] Benoit, C., Grice, M. and Hazan, V. *The SUS test: A method for the assessment of text-to-speech intelligibility using Semantically Unpredicatable Sentences.* Speech Communication, Vol. 18, pp. 381-392, 1996.

MUSICAL INTERVALS OF TONES IN CANTONESE ENGLISH

Suki S.Y. Yiu

Linguistics Department, University of Hong Kong

syutji@hku.hk

Abstract

It has been shown that the relative pitch levels of Cantonese tones closely correspond to musical intervals (MIs) [1]. Given that an emerging tone language, Cantonese English, has developed tone under the substrate influence of Cantonese, this paper examines the correspondence between the newly emerged tones and MIs, and how the musical analogy relates to those established for Cantonese.

The fundamental frequencies of the tones produced by six speakers of Cantonese English were extracted with Praat, then time-normalized across rhymes. The mean values of the interval points of two tones were expressed in terms of ratio, then matched with the closest MI on the musical scale.

This paper demonstrates that the pitch levels of tones in Cantonese English correspond to MIs, given the converging ranges of MIs for different speakers and similar MIs of different tone pairs for different speakers. It also shows that the MIs of tones in Cantonese English are related to the corresponding tone pairs for Cantonese. The viability of MI as a means to understand the tonal system of non-tonal languages whose speakers' native language is tonal extends the link between the use of pitch in speech tones and music.

Index Terms: tone and music, musical interval, frequency ratio, Cantonese English, Hong Kong English

1. Introduction

To record the principal phonetic characteristic of lexical tone, i.e. fundamental frequency (F0), Chao developed the 5-level transcription of tonal pitch variation（五度標記法）with the use of sliding-pitchpipes [2]. When the pitch of the pitchpipe matches the pitch of the linguistic tone, the pitch values of a linguistic tone (the starting and ending points, also a turning point between the two if any) can be notated on a staff. The major advantage of this method is that the relativity of pitch, and therefore spatial relationship among lexical tones, could be recorded musically. The tones are represented phonologically on a tone scale with five numeric values from 1 to 5, a method usually adopted for Chinese languages [3]. Echoing Chao's method of notating the lexical tones via a musical means, this paper explores how linguistic tones can be understood in terms of musical intervals (MIs) based on the phonetic data obtained in Cantonese English.

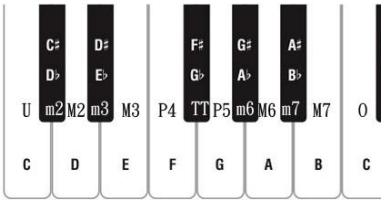


Figure 1: *MIs from C on piano keyboard [1]*

A MI is a perceptual distance of pitch, expressed in form of a ratio to show the distance between two adjacent notes on a scale. Figure 1 shows a chromatic scale on a piano keyboard. Different MIs are expressed in abbreviated forms: unison (U), minor (m), major (M), perfect (P), tritone (TT) and octave (O).

Let the leftmost C be an anchor key, one step up counts as one semitone. Moving from C to C♯ involves one semitone, and its MI is a minor second. Two semitones or one whole tone is involved when moving from C to D, and the corresponding MI is a major second. The longer the distance, the more the semitones involved, and the bigger the MI. Notably, the difference between two keys is relative in terms of MI. For example, the distance from C to E and from E to G♯ is the same: the MIs are both major thirds, with four semitones. As relativity of pitch is a core characteristic of phonemic tone, MI serves as a link to understand linguistic tones musically.

The MIs of linguistic tones can be obtained by calculating the frequency ratios of the tone pairs in a tone inventory, a method particularly useful for describing intervals in both Western and non-Western music. The absolute pitch of one tone is expressed in the form of a ratio to another tone. The ratios can then be matched with the closest MI on the musical scale. It has been shown that the pitch intervals of Cantonese tones closely correspond to the number of semitones derived from the MIs [1]. Given that Cantonese English has developed tone under the substrate influence of Cantonese, is there a musical basis or analogue to the tone inventory of this emerging tone language? If so, how does the musical basis or analogue relate to that for Cantonese?

Cantonese English is an emerging language which owns its distinctiveness in many linguistic aspects that are worth paying attention to for the sake of providing a neutral and unbiased description [4]. Many phonological aspects of Cantonese English have been studied so far [5-9]. Specifically, the use of pitch in Cantonese English, unlike tones in Cantonese, is perceptually distinct but carries a low functional load, with few if any minimal pairs in normal speech. This paper refers to this systematic use of pitch as tone, and examines the relationship among these tones, as described in recent works like [6, 10-13].

The goals of this paper are to find out the correspondence between the new tones and MIs, and examine how the musical analogy relates to those established for Cantonese. The first working hypothesis is that the pitch levels of tones in Cantonese English correspond to the MIs. This predicts that the ranges of MIs for different speakers should display a convergent pattern, and the MIs of different tone pairs for different speakers should be similar. The second working hypothesis is that the MIs of tones in Cantonese English are related to those for Cantonese. This predicts that there are corresponding tone pairs in Cantonese English and Cantonese.

2. Linguistic tone and musical tune

2.1. Relativity in tone and tune

The relationship between language and music has been shown to be multifold. In recent works on tone-melody mapping, the

conformity between lexical tone and melody in vocal songs has been the focus [14-22]. Both speaking and singing involve vocalization and manipulation of pitch, but composition of songs sometimes requires fixed pitch notation while pitch in speech tones is a relative concept in principle. Depending on the pitch of the adjacent, or even neighboring, tones in speech, the interpretation of tones may vary. If the pitch of the adjacent or neighboring tones is lower, the target tone will be perceived as a higher tone. Likewise, higher pitch of the adjacent or neighboring tones makes the target tone a lower tone [23]. Relativity is an important characteristic not only of speech tone, but also of musical tone because not everyone, even musicians, has absolute pitch, and many people sing songs with the right intervals but not at the fixed pitch transcribed in the musical notation. They need to attune their instruments and the pitch of their voice with the help of tuning devices.

2.2. Cantonese tones and musical intervals

Since relativity of pitch is crucial to the link between tone and tune, one would expect correspondence between the frequency ratios of tones in speech and tones in sung melody. [19] used a ratio called High-Low Quotient for each of the twenty Chinese languages by dividing the F0 of the highest pitch level with the F0 of the lowest pitch level articulated by a given speaker of each language. Instead of only using the highest and lowest F0, [22] identified six pitch intervals (1-2, 1-3, 1-5, 2-3, 2-5 and 3-5) in Cantonese and calculated the frequency ratios involved when transiting from one pitch level to another with single utterance data by five speakers from [24]. Though both studies claimed that the frequency ratios did not align with the musical intervals neatly, the values were close and seemed to show patterns.

[1] proposed an MI analysis for speech tones in Cantonese similar to the one adopted by this paper. It collected F0 data of Cantonese tones and calculated the frequency ratios for all of the seven tone pairs in Cantonese (T1:T3, T1:T6, T1:T4, T2:T5, T3:T6, T3:T4 and T6:T4) from multiple utterances produced by six speakers. Since the frequency ratio and the MI share the same way of calculation using the equation $\frac{F0x}{F0y}$, the frequency ratio for the speech tones is named as MI.

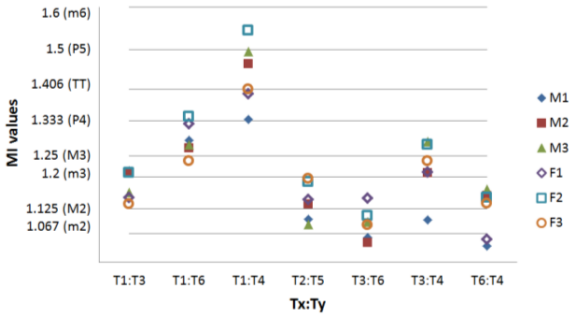


Figure 2: *Cantonese MIs of all subjects [1].*

Figure 2 shows the MIs identified for the six subjects. F/M1 to M/F3 represent fe/male speakers numbered 1 to 3 respectively. It was found that the minimal distance between two tones in Cantonese is a minor second for the tone pair of T3:T6, with pitch levels 3-2. The maximal distance between two tones in Cantonese is a perfect fifth, which is found for the tone pair of T1:T4, with pitch levels 5-1. This coincided with the findings in tone-melody mapping, where a tonal transition from 5-1 is found to be mapped onto a musical interval of a perfect fifth [20, 22]. Figure 2 demonstrates that MI is a possible scale to portray tones in a tone inventory produced by multiple speakers, and to display the spatial relationship between different tone pairs in each speaker's tonal space.

2.3. Musical intervals and semitones

MIs allow tonal analysis be done without committing to specific tuning systems which divide an octave slightly differently, thus avoiding small yet possibly significant effects of such differences on the interpretation of the spatial relationship of the tones in a tone inventory. Comparison of linguistic tones can be obtained through MI ratios ranging from 1 to 2 in an octave. Switching from one tuning system to another can be done by changing the ratios which the MIs are based on.

One can also divide the MI ratios into smaller logarithmic units like semitones by $\log_2 MI * 12$, and so as cents by $\log_2 MI * 12 * 100$. Adopting a logarithmic unit of measurement will fit the reported logarithmic characteristic in perception [25] and production [26] of pitch in speech.

This paper adopts the just intonation tuning system as provided by 5-limit tuning, where the main intervals and just intervals are found to sound pleasant and well-tuned to most people.

3. Methodology

Six Cantonese English speakers balanced for gender were chosen as subjects. They were born after 1980 and raised in Hong Kong. Their age ranged from 21 to 32 years old, representing young adult speakers of Cantonese English.

The data was elicited using a wordlist providing a comprehensive coverage of the syllable structure, vowel and consonant inventories, and surface tones of Cantonese English. The criteria for stimulus construction are listed in Table 1 below. 24 target items were randomized with 8 fillers during elicitation.

Table 1. *Criteria for stimulus construction.*

	Description
Syllable structure	(C)V(V/N/C)
Vowel	Syllables mainly contain but not limited to 3 cardinal vowels [i], [a], [u]
Consonant	[t], [b], [ʃ], [m], [k], [l], [s], [g], [n], [p], [ɹ]/[w], [f], [d]
Tone category	M(id), H(igh), Mf(alling), L(ow), Hf

The subjects were asked to produce each word three times in a row so that the basic position (initial, medial or final) of each target item in an utterance was controlled, for instance, *arm-arm-arm*. A total of 2376 target syllables (= 66 target syllables x 3 repetitions x 2 sets x 6 subjects) were recorded. Recordings were made with Praat (ver. 5.3.39) [27] with a sampling frequency of 22050Hz in a sound-proofed recording booth.

F0 tracks of the items in the utterance final position were extracted with Praat, then time-normalized at 10% interval points across the rhyme of each syllable with Praat script ProsodyPro (ver. 4.3) [28]. Based on the above categorization of tones, tone profiles of individual speakers were generated in the form of line graphs. The mean values of the ten interval points of each of the two tones were expressed as a ratio in

order to match with the equivalent ratio for the closest MI. The MI of each tone pair was then used to map the closest MI on the musical scale. For the sake of accuracy, the MIs were expressed in numbers, as in Table 2. Since the tones seemed to be restricted to certain syllable positions in a word, the selection of tones relevant for the calculation of MIs will be discussed in the following section after presenting the Cantonese English tone profiles.

4. Results and discussion

4.1. F0 profiles of subjects

The Cantonese English F0 profile of each subject was generated in form of line graphs like Figure 3. Recall that the words were repeated three times when recorded so as to control the position of the words in an utterance. Only the words in the utterance-final position are used because they display richer materials for tonal analysis. Also, since the placement of tones in Cantonese English seems to be restricted in certain syllable positions of a word, three graphs, from left to right, are used to present the tones occurring in the word-initial, word-medial and word-final positions respectively. H, M and L are each shown by solid, dashed and dotted lines.

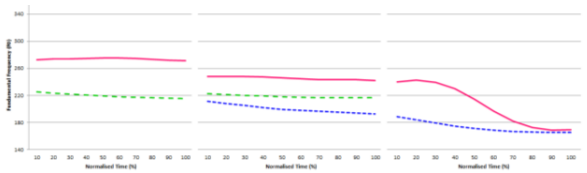


Figure 3: *Cantonese English tone profile of speaker F2.*

As observed above, not all of the five tones surface in all three positions in a word. H and M surface in both word-initial and word-medial positions. In the word-final position, the lower curve is regarded as a boundary L, and the upper curve with a distinguishable falling contour is a H transiting to a boundary L. Additional data shows that the more syllables between the rightmost H and the boundary L, the flatter the slope of each curve. The intervening tones from the rightmost H in a word to the boundary L are excluded from the calculation of MIs. To minimise declination effects, H and M in word-medial position and L in the final position are selected, given that the L only surfaces at the right boundary of a word. Also noticeable from the data is that M only occurs before H, and H only occurs before L. There are some variations concerning the details of the pitch contours, but the patterns described above generally hold across subjects.

4.2. Tones in Cantonese English on MI scale

The pitches of each pair of tones were expressed as a ratio and compared with the pitch ratios for the closest corresponding MIs. The data of speaker F2 is used as an example.

Table 2. *MIs of speaker F2.*

$\bar{T}x$:$\bar{T}y$	$\bar{T}x$ (Hz)	$\bar{T}y$ (Hz)	$\bar{T}x/\bar{T}y$	Closest MI	Closest MI ref. value	No. of semitones
H:M	245.2154	218.3493	1.123042	M2	1.125	2.008939
H:L	245.2154	172.9993	1.417436	TT	1.40625	6.039400
M:L	218.3493	172.9993	1.262140	M3	1.25	4.030460

Table 2 shows all possible combinations of ratios for the Cantonese English tones, H, M and L, of speaker F2. The first column ($\bar{T}x$:$\bar{T}y$) is filled with the ratio to which each row of data corresponds. The second and third columns ($\bar{T}x$ and $\bar{T}y$) are the mean F0 of the ten interval points on the tonal contour for each tone in Hz. The column $\bar{T}x/\bar{T}y$ shows the ratios in numbers. The ratios are then matched with the closest MI in the next column, followed by the actual ratio values for each MI. The last column shows the number of semitones calculated by the equation provided in Section 2.3. Let us take H:M as an example. The mean F0 of H and M are 245.2154 Hz and 218.3493Hz respectively. Their ratio is 1.123042 (= $\frac{245.2154\ \text{Hz}}{218.3493\ \text{Hz}}$),closest to a major second,having a ratio of 9:8, i.e. 1.125, and around two semitones. Figure4 shows the results using a similar method for each subject as in Section 2.2.

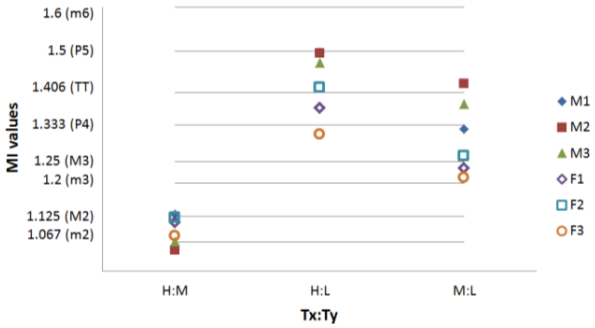


Figure 4: *Cantonese English MIs of all subjects.*

4.2.1. *Distance between linguistic tones and tone space*

The distance between two tones in a tone pair is reflected by the values of the MIs. The MIs of all subjects fall within the range of minor second to perfect fifth. A minor second is the shortest distance that two musical tones could have on a keyboard, with the distance of one semitone. A perfect fifth is the distance from C to G (cf. keyboard in Figure 1), with seven semitones in between. H:M is the tonal pair with the smallest MI (from minor to major second) whereas H:L is the pair with the biggest MI (from perfect fourth to fifth). M:L came within a minor third to a tritone.

Speaker F3 has a relatively narrow MI range – all tones are spaced within a perfect fourth, with five semitones between the highest and lowest tones. On the contrary, the male speakers have the widest range – their tones are spaced within a perfect fifth, with seven semitones between H and L. In other words, the tonal space of male speakers is wider than for the female speakers, especially for speaker F3. The tonal space of speakers F1 and F2 is wider than speaker F3 but still narrower than for the male speakers.

A wider tonal space allows a wider range of MIs. This can be checked by the number of semitones between the highest and lowest MIs of the subjects. Speaker F3 uses MIs ranging from a minor second to a perfect fourth, within four semitones, while speakers M2 and M3 use MIs from a minor second to a perfect fifth, within six semitones. Close to speakers M2 and M3 is speaker M1, who uses MIs from a major second to a perfect fifth, in five semitones. The MIs of female speakers all range within four semitones with different starting and ending MIs: speakers F1 and F2 range from major second to tritone, while speaker F3 ranges from minor second to perfect fourth.

4.2.2. *Spatial relationship among linguistic tones*

The spatial relationship among different tones in a tone inventory is reflected through comparison between the MIs of

different tone pairs in a tone inventory. By comparing the MIs of different tone pairs, whether M is closer to H or L in Cantonese English can be figured out. In Figure 4, H:L has the highest MI. Since there are only three tones, H and L ought to be at opposite ends with M somewhere in between. Interestingly, instead of being in the middle of H and L, M is closer to H than L. This relationship among the tones in Cantonese English is so neat that it holds across the MI profiles of all six speakers. This unanimity displayed by the subjects suggests that the M is not randomly assigned to anywhere between H and L in Cantonese English, and hence has a stable position in the tonal space of Cantonese English.

4.2.3. Flexibility of linguistic tones

The flexibility of linguistic tones is reflected by the range of MIs, i.e. the extent to which the MIs of different speakers cluster for different tone pairs. The more unanimous MIs (minor or major second) for H:M confirm that the interval between these tones corresponds to minimally one semitone while the more flexible MIs (perfect fourth, tritone or perfect fifth) for H:L indicates that a wider range of MIs are allowed.

4.3. Comparing the new tones with canonical tones

Given the above tonal analysis of Cantonese English in terms of MIs, the next question is how the tone system of Cantonese English is similar to or different from that of its substrate language Cantonese. If the MIs of tones in Cantonese English can be related to Cantonese, not only does it show that MIs link speech tones and musical tones together, but it also serves as a link between the phonological tones in canonical tonal languages like Cantonese, and languages like Cantonese English where the tonal elements are probably substrate influence from one's tonal native language. With a similar methodology and the same subjects, this section compares the MIs of the new tones in Cantonese English and the canonical tones Cantonese.

The six phonemic tones in Cantonese are T1 [55], T2 [25], T3 [33], T4 [21], T5 [23] and T6 [22], according to [29]. Among the seven possible combinations of tone pairs in Cantonese and the three possible tone pairs in Cantonese English, the following tone pairs are selected for comparison: T1:T6, T1:T4 and T6:T4 in Cantonese, and H:M, H:L and M:L in Cantonese English respectively.

4.3.1. Distance between linguistic tones and tone space

The MIs for the tone pair whose tones are most distant from each other in Cantonese English (H:L) and Cantonese (T1:T4) are very close. They are the same for speakers M2, M3 and F1in both languages. The MI of speaker M1 is greater in Cantonese English by two semitones, while for speakers F2 and F3, their MI is smaller in Cantonese English by one semitone. H:L in Cantonese English and T1:T4 in Cantonese have the largest MI across speakers, ranging from perfect fourth to fifth. This suggests that the total pitch range used is much the same for Cantonese English and Cantonese.

For the other two tone pairs in both languages under comparison, while the smallest MI in Cantonese English and Cantonese are H:M and T6:T4 respectively, H:M is also the one with the smallest MI in the tone system of Cantonese English but T6:T4 is not in that of Cantonese. This could be due to the fact that the tonal space of Cantonese is more crowded than that of Cantonese English so that there can be even more fine-grained categorisation of tones.

4.3.2. Spatial relationship among linguistic tones

M is closer to H than L in Cantonese English, but T6 is closer to T4 rather than T1 in Cantonese across subjects. This may be due to having no perfect equivalent for Cantonese English M in Cantonese. The M in Cantonese English is slightly higher than T6 but not as high as T3 in Cantonese. Also, the unanimity displayed for the assignment of M suggests that the M in Cantonese English, like T6 in Cantonese, has a stable position in the tonal space, and hence in the tone inventory of Cantonese English.

4.3.3. Flexibility of linguistic tones

Unlike Cantonese, where the tone pair whose tones are most distant from each other, i.e. T1:T4, is also the most flexible tone pair in terms of MIs (perfect fourth, tritone and perfect fifth), the tone pair whose tones are second distant from each other, i.e. M:L is the most flexible tone pair (minor and major third, perfect fourth and tritone) in Cantonese English. The MIs of H:L are also quite flexible (perfect fourth, tritone and perfect fifth) but not as M:L, suggesting that a choice between transiting from H to L or from M to L should be available in tone-melody mapping.

The tone pair whose pitches are closest to each other in Cantonese English, i.e. H:M, is also the least flexible tone pair in terms of MIs (minor and major second), similar to T6:T4 (major third and perfect fourth) among the three tone pairs under comparison in Cantonese. That said, if all seven tone pairs are taken into account, the least flexible tone pair is T3:T6 (minor and major second), which is also the tone pair whose pitches are closest to each other.

5. Conclusions and implications

This paper has examined the correspondence between the new tones and MIs, and how the musical analogues relate to those for Cantonese. It has been demonstrated that the pitch levels of tones in Cantonese English correspond to the MIs in terms of the distance between/among tones and the flexibility of tonal intervals, given the converging ranges of MIs for different speakers, and similar MIs of different tone pairs for different speakers. It has also been shown that the MIs of tones in Cantonese English are related to those for Cantonese to some extent, with corresponding tone pairs in Cantonese English and Cantonese.

By demonstrating that such a musical treatment of linguistic tone is viable, it has been shown that MIs serve as a means to understand the tonal system of non-tonal languages whose speakers' native language is tonal, with characteristics of tones invisible in other existing approaches but visible through the MI glass. It has also extended the link between the use of pitch in speech and music. It seems promising to apply the proposed method to more new varieties of languages with a larger sample. Considering the convergence of tonal characteristics in Cantonese English, also in comparison with Cantonese, this paper supports the view that Cantonese English is an emerging tone language.

6. Acknowledgments

I thank Stephen Matthews and Diana Archangeli for their thought-provoking discussions. I also thank Lianhee Wee, Cathryn Donohue, 'lab rats' Winnie Cheung and Queenie Chan, the three anonymous reviewers and the six subjects.

7. References

[1] Yiu, S. S. Y., "Cantonese Tones and Musical Intervals". In W.S. Lee (Ed.) Proceedings of the International Conference on Phonetics of the Languages in China 2013 (ICPLC 2013), Hong Kong: the Organizers of ICPLC 2013 at the Department of Chinese, Translation and Linguistics, City University of Hong Kong, 155-158, 2013.

[2] Chao, Y.-R., "现代吴语的研究". 科學出版社. 3,4,74, 1956.

[3] Chao, Y.-R., "A System of Tone Letters". La Maître phonétique 45:24-27, 1930.

[4] Mohanan, K. P., "Describing the Phonology of Non-Native Varieties of a Language". World Englishes, 11, 111-128, 1992.

[5] Hung, T., "Towards a phonology of Hong Kong English". World Englishes, 19(3), 337-356, 2000.

[6] Luke, K. K., "Phonological re-interpretation: The assignment of Cantonese tones to English words". Paper presented at the 9th International Conference of Chinese Linguistics. National University of Singapore, 2000.

[7] Peng, L. and Setter, J., "The Emergence of Systematicity in English Pronunciations of Two Cantonese-speaking Adults in Hong Kong". English World-wide, 21(1), 81-108, 2000.

[8] Hung, T., "Word Stress in Hong Kong English: a preliminary study". HKBU Papers in Applied Language Studies, 9, 29-40, 2005.

[9] Sewell, A. and Chan J., "Patterns of variation in the consonantal phonology of Hong Kong English". English World-Wide, 31(2), 138-161, 2010.

[10] Wee, L.H., "Phonological patterns in the Englishes of Singapore and Hong Kong". World Englishes 27(3/4):480-501, 2008.

[11] Cheung, H.Y. W., "Span of high tones in Hong Kong English". In Proceedings of BLS 35, 72-82. Berkeley, CA, 2009.

[12] Yiu, S. S. Y., "Intonation of English Spoken in Hong Kong". BA Thesis, Hong Kong Baptist University, 2010.

[13] Gussenhoven, C., "Tone and Intonation in Cantonese English". TAL 3, Nanjing, May 26-29, 2012.

[14] Chan, M. K. M., "Tone and Melody Interaction in Cantonese and Mandarin songs". UCLA Working Papers in Phonetics vol. 68:132-169, 1987.

[15] Agawu, V. K., "Tone and Tune: The Evidence for Northern Ewe Music". Africa: Journal of the International African Institute 58, no. 2, 127-146. 1988.

[16] Wong, P. C. M. and Diehl, R. L., "How Can the Lyrics of a Song in a Tone Language Be Understood?" Psychology of Music 30, no. 2: 202-209, 2002.

[17] Ho, W. S. V., "The Tone-Melody Interface of Popular Songs Written in Tone Languages". In Proceedings of 9th International Conference on Music Perception and Cognition (ICMPC9), University of Bologna, Italy, August 22-26, 1414-1422, 2006.

[18] Wee, L. H., "Unraveling the Relation between Mandarin Tones and Musical Melody". Journal of Chinese Linguistics vol. 35, 1, 128-144, 2007.

[19] Wee, L. H., "Inquiry into the Musical Nature of Linguistic Tone". In Hsiao, Yuchau, Hui-chuan Hsu, L.H. Wee and Dah-An Ho (eds) Interfaces in Chinese Phonology. Taiwan: Institute of Linguistics, Academia Sinica, 139-160, 2008.

[20] Ho, W. S. V., "Fine-Tuning Tone-Melody Constraints through the Investigation of Mismatches in Cantonese Pop Music". Hong Kong: City University of Hong Kong, 2009.

[21] Sollis, M., "Tune-Tone Relationships in Sung Duna Pikono". Australian Journal of Linguistics 30, no. 1: 67-80, 2010.

[22] Chow, M. Y., "Singing the Right Tones of the Words: the Principles and Poetics of Tone-melody mapping in Cantopop". MPhil dissertation: The University of Hong Kong, 2012.

[23] Fok Chan, Y.-Y., "A Perceptual Study of Tones in Cantonese". Hong Kong: University of Hong Kong, 1974.

[24] Bauer, R. S., "Hong Kong Cantonese Tone Contours". In Studies in Cantonese Linguistics, edited by S. Matthews, 1-33. Hong Kong: Linguistic Society of Hong Kong, 1998.

[25] Nolan, F. "Intonational equivalence: an experimental evaluation of pitch scales". In Proceedings of the 15th international congress of phonetic sciences, 771-774, 2003.

[26] Fujisaki, H. "Prosody, information, and modeling: With Emphasis on Tonal Features of Speech". In Proceedings of Workshop on Spoken Language Processing,Mumbai,5–14, 2003.

[27] Boersma, P. and Weenink D., Praat: doing phonetics by computer [Computer program]. Version 5.3.57, retrieved 2012 from http://www.praat.org/, 2013.

[28] Xu, Y., Praat script ProsodyPro (version 4.3), 2012.

[29] Matthews, S. and Yip V., "Cantonese: A Comprehensive Grammar". London: Routledge, 1994, 2011.

Rhythmic Patterns in Native and Nonnative Mandarin Speech

Wentao Gu[1, 2] *and Keikichi Hirose*[2]

[1] Research Center for Language Information Technologies, Nanjing Normal University, China
[2] Graduate School of Information Science and Technology, The University of Tokyo, Japan
{wtgu, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Rhythm plays an important role in the naturalness of speech. This study compared rhythmic patterns of Mandarin speech between native speakers and two groups of L2 speakers whose first languages were Cantonese and English, respectively. The study started from isolated words, but focused on continuous speech, for which eleven durational metrics were used as objective rhythm indicators. The results on continuous speech showed that nonnative Mandarin gave a quite similar rhythmic mode as native one in terms of rate-normalized/independent metrics, but shifted towards the stress-timed class in terms of raw metrics, regardless of the rhythmic class of the L1. This seems to conflict with the L1 transfer effect and the results for isolated words, but it coincides with auditory impression and can be explained by speech rate difference and the lengthening effects associated with the change in prosodic structure.

Index Terms: rhythm, durational metrics, nonnative speech, Mandarin, Cantonese, English

1. Introduction

As an important prosodic property, rhythm plays a key role in the naturalness of speech. There used to be a tradition of classifying spoken languages in the world into three rhythmic classes, i.e., 'stress-timed', 'syllable-timed', and 'mora-timed' [1-3]. This classification was based on the notion of isochrony, which assumes the existence of units of nearly equal duration in speech: syllables for syllable-timed languages such as French and Italian, inter-stress intervals for stress-timed languages such as English and German, and morae for mora-timed languages such as Japanese.

However, experimental studies have failed to find any acoustic evidence to support the existence of isochronous units [4-6]. Thus, the so-called isochrony is only an impressionistic property which correlates with a number of phonological aspects such as syllable structure, vowel reduction, and stress [4]. Instead of searching for isochronous units, many recent studies came to find out acoustic metrics that could correspond roughly to the auditory impression of rhythmic distinctions, by inspecting the durational variability of consonantal and vocalic intervals, such as ΔC (standard deviation of consonantal duration), %V (percentage of vocalic duration) [5], and PVI (Pairwise Variability Index) of vocalic and consonantal durations [6, 7]. Similar metrics were also applied to syllable duration [8, 9]. These studies showed that such durational metrics could categorize spoken languages into different rhythmic classes. Also, they suggested that the difference in rhythm was not categorical; instead, various languages could be on a continuum between extreme rhythmic patterns [4, 6].

It has been widely recognized that English is a typical stress-timed language [2], in which there is perceptually a roughly constant amount of time between successive stressed syllables, and relatively low %V, high ΔC, and high PVIs tend to be measured acoustically [5, 6]. To accommodate the stress-timed rhythm, there is a tendency for unstressed syllables to be shortened. In contrast, Cantonese is deemed a typical syllable-timed language [8], in which successive syllables perceptually have roughly constant duration, and relatively high %V, low ΔC, and low PVIs tend to be measured acoustically. Mandarin (in mainland China) also shows a syllable-timed pattern, at least in read speech, but it is less typical than Cantonese as evidenced by various acoustic measures [8].

Besides rhythmic studies for native speech, there are also a few studies on rhythmic patterns for nonnative speech, e.g., English L2 by Mandarin and Cantonese L1 speakers [9], and the effects of L1 on L2 for Dutch, English, and Spain speakers [10]. This is important as objective rhythmic metrics may be applied in computer-assisted language learning systems.

The findings of these studies, however, are not consistent. In some cases nonnative speech is rhythmically between L1 and L2, coinciding with the general hypothesis of L1 transfer effect, while in other cases nonnative speech is rhythmically almost identical to L1 or L2, or even shows a pattern overshooting L2 – hence opposite to the L1 transfer effect. It is not easy to interpret such inconsistencies, though speech rate variation and selective lengthening have been used for explanation [9, 10]. This may lead us to question the validity of these rhythmic metrics for nonnative speech.

Since there have been few rhythmic studies on nonnative Mandarin speech, we attempt to fill the gap. We selected two groups of L2 Mandarin speakers who were native in American English and Hong Kong Cantonese, respectively. The reason for selecting these two groups is that English and Cantonese represent two opposite rhythmic classes as we described above.

2. Speech Data

2.1. Comparison in Phonology

The three languages concerned here, i.e., Standard Mandarin, Hong Kong Cantonese, and American English, contrast sharply in the phonological structure. Mandarin and Cantonese are tone languages of monosyllabic nature in the morpheme sense, while English is a stress language of polysyllabic nature.

Both Mandarin and Cantonese have a very simple syllable structure (C)V(C), and each syllable has a tone. In addition to monosyllabic nature in the morpheme sense, Cantonese has a higher frequency of monosyllabic words than Mandarin, thus further enhancing its monosyllabic nature. Cantonese does not have a contrast in lexical stress while Mandarin does – there is no lexical stress in the phonological sense but there is a neutral tone functioning as an unstressed syllable which usually has a shortened duration.

English, on the contrary, is a polysyllabic language with a more complex syllable structure: with the use of consonant clusters, English syllables can be (C)(C)(C)V(C)(C)(C)(C). Also, English is a stress language, for which the syllables in a polysyllabic word differ in the degree of lexical stress.

2.2. Materials

Read speech was used in the present study for the purpose of a controlled comparison among three groups. Two sets of Mandarin speech materials were designed. Speech Material I consisted of disyllabic and trisyllabic words which were used for analysis of timing in isolated words, while Speech Material II consisted of short stories which were used for investigating rhythmic patterns in continuous Mandarin speech.

Speech Material I included disyllabic and trisyllabic words with neutral tone (T0). The corpus of disyllabic words included four tonal combinations, with four lexical tones in the former syllable and T0 in the latter. With 10 words designed for each disyllabic tonal combination, there were altogether 40 disyllabic words. For trisyllabic words, T0 could occur in the mid and/or the latter syllable, and thus the corpus of trisyllabic words included 16 + 16 + 4 = 36 tonal combinations with T0. With two words designed for each trisyllabic tonal combination, there were altogether 72 trisyllabic words.

Speech Material II included three short stories, each with a length of about 200 syllables. One story was 'North Wind and Sun' as widely used in phonetic study. All three stories were easily understood by L2 Mandarin learners at the intermediate or advanced level, as confirmed by the nonnative subjects.

2.3. Subjects and data collection

Three groups of subjects were recruited: MM, CM, and EM. The MM group, including 3 males and 3 females with an average age of 20, consisted of native speakers of Mandarin, who were undergraduate students majoring in broadcasting and hosting arts, all professional in Mandarin pronunciation. The CM group, including 1 male and 5 females with an average age of 22, consisted of L2 Mandarin learners who were born in Hong Kong and were native in HK Cantonese. The EM group, including 3 males and 3 females with an average age of 22.5, consisted of L2 Mandarin learners who were born in USA and were native in American English. The subjects in both CM and EM groups were L2 Mandarin learners at the intermediate or advanced level. They had studied Mandarin for at least two years, and had already passed Level 5 of HSK, the Chinese proficiency test.

Speech recording was conducted in a sound-proof room after the subjects had got familiar with the reading materials and made sure that they had known the exact pronunciations. The recording was done at each subject's normal speech rate and was monitored by the experimenter. Once there was a mispronunciation or disfluency, the subject would be asked to repeat recording the word or the utterance until success.

2.4. Data analysis

By visual inspection of the waveform and the spectrogram, speech materials were segmented and labelled manually. Each word in Speech Material I was segmented into syllables, while Speech Material II was segmented into consonants, vowels, and pauses. Any pauses in the speech, either silent or not, were excluded from our analysis. On the basis of acoustic instead of phonological criteria, glides were classified as vowels, because the formant transition between a glide and a vowel nucleus in Mandarin is usually smooth, without an abrupt change.

Unlike for Speech Material I, labelling of Speech Material II was conducted not only for segments and tones but also for break indices, following a ToBI-like approach [11]. Three layers of break indices were adopted: prosodic word boundary (B1), minor prosodic phrase boundary (B2), and major prosodic phrase boundary (B3). A prosodic word is a basic unit that is tightly integrated in prosody. B2 differs from B1 in that B2 is accompanied by one of the following perceivable features: a short pause, a pitch resetting, or a final lengthening. B3 tends to be accompanied by a much longer pause than B2.

The labelling of break indices was more subjective than that of segments and tones. To ensure the reliability, we asked four labelers to do the labelling independently after an iterative training and discussion (the cross-labeler consistency reached 96% at the end of this stage). After their labelling, the consistency across the four labelers turned out to be 86%. We then finalized the labelling by double checking those apparently inconsistent labels.

For Speech Material II, the durations of vocalic and consonantal (i.e., intervocalic) intervals were measured. A vocalic interval is the section between the onset and the offset of a series of connected vowels/glides, while a consonantal interval is the section between the onset and the offset of a series of connected consonants. For Speech Material I, we simply measured the durations of individual syllables.

The following seven rhythmic metrics based on vocalic and consonantal interval durations [5, 7, 12] were calculated for each subject, and then were averaged across all subjects.

ΔC: the standard deviation of consonantal durations
ΔV: the standard deviation of vocalic durations
%V: the proportion of vocalic durations in the speech
VarcoC: (ΔC / mean consonantal duration) × 100
VarcoV: (ΔV / mean vocalic duration) × 100

$$rPVI\_C = \left(\sum_{k=1}^{m-1} \left| d_{Con_k} - d_{Con_{k+1}} \right| / (m-1) \right)$$

$$nPVI\_V = 100 \times \left(\sum_{k=1}^{m-1} \left| \left(d_{Vow_k} - d_{Vow_{k+1}} \right) / \left(\left(d_{Vow_k} + d_{Vow_{k+1}} \right) / 2 \right) \right| \right) / (m-1),$$

Here, VarcoC and VarcoV are rate-normalized metrics, which were introduced because ΔC and ΔV were found to be negatively correlated with speech rate [10]. PVI indicates the absolute durational difference between each pair of successive units [6]. Two PVI values, i.e., raw and rate-normalized, were calculated for consonantal and vocalic intervals, respectively, as vocalic duration was more sensitive to speech rate while consonantal duration carried more language variability [7].

Following [8, 9], four syllabic metrics were also calculated on the basis of syllable duration, including ΔS, VarcoS, rPVI_S, and nPVI_S, which were defined in the same way as their counterparts for consonantal and vocalic durations.

3. Results

3.1. Timing in isolated words

We started our study from isolated words in Speech Material I. The average percentages of duration for T0 syllables in a word are shown in Table 1, where X indicates any of the four lexical tones. In the case of trisyllabic words with two T0 syllables, the percentages of all three syllables are presented. In all cases MM gives the minimal durational ratio of T0 (in the case of X+T0+T0 the mid syllable should be the shortest, and hence the ratio of the mid T0 is compared), indicating that nonnative speakers did not shorten T0 adequately and hence decreased the durational variation in a word. In other words, nonnative speakers did not differentiate unstressed syllables from others

appropriately. Thus, it is expected that the rhythmic patterns of their continuous speech might be shifted further towards the syllable-timed extreme, especially for Cantonese L1 speakers.

Table 1. *Percentages of duration of T0 syllable in a word.*

Group	X+T0	X+T0+X	X+X+T0	X+T0+T0		
				X	T0	T0
CM	46.0	26.1	33.6	32.7	33.8	33.5
EM	47.4	25.0	33.3	34.7	30.2	35.1
MM	38.6	23.3	30.0	37.6	28.0	34.4

3.2. Rhythmic measurements for continuous speech

We then investigated rhythmic patterns for continuous speech in Speech Material II, using all eleven metrics described above.

Table 2 gives the statistical results of comparing the means of various rhythmic metrics among three groups. The left shows the *p*-values for one-way ANOVA, while the right shows the results of Bonferroni post-hoc tests only for those metrics showing significant main effects. For ΔC, ΔV, rPVI_C, nPVI_V, and rPVI_S, there are main effects of group, and the differences are significant only between native and nonnative groups – there is no significant difference between CM and EM. Besides, there is a marginally significant effect of group for ΔS, but no significant effects for the other metrics.

Except for nPVI_V, all those metrics that differentiate native and nonnative speech are non-rate-normalized; most of the rate-normalized metrics, however, have not provided any distinction between native and nonnative speech. The average speech rates turned out to be 4.14, 4.16, and 5.05 syllables per second for CM, EM, and MM, respectively. There was a main effect of group ($p < 0.001$). Bonferroni post-hoc tests showed that there was no significant difference between CM and EM; there was, however, a significant difference between MM and CM at $p < 0.01$, as well as between MM and EM at $p < 0.001$. This not only verifies that native speech is generally faster than nonnative speech, but also shows that the two nonnative groups have almost equal language proficiency in Mandarin.

It has been shown that ΔC, ΔV, and rPVI_C are negatively correlated with speech rate [10]. Therefore, the observed differences in ΔC, ΔV, and rPVI_C between native and non-native speakers might be the results of speech rate differences. While using VarcoV instead of ΔV helps capture rhythmic patterns better, rate-normalization of consonantal metrics such as ΔC and rPVI_C removes phonotactic differences between languages and hence cannot discriminate rhythmic classes, as shown in [6, 10]. Hence, we adopt raw metrics for consonants and rate-normalized/independent metrics for vocalic intervals.

Table 2. *Significance levels for the differences in rhythmic scores between the three groups.*

Metric	ANOVA	Bonferroni post-hoc test		
		CM–EM	CM–MM	EM–MM
ΔC	0.010**	0.614	0.013*	0.005**
ΔV	0.041*	0.536	0.017*	0.057†
%V	0.475	–	–	–
VarcoC	0.489	–	–	–
VarcoV	0.564	–	–	–
rPVI_C	0.013*	1.000	0.062†	0.016*
nPVI_V	0.024*	0.847	0.208	0.023*
ΔS	0.058†	–	–	–
VarcoS	0.499	–	–	–
rPVI_S	0.037*	1.000	0.142	0.046*
nPVI_S	0.275	–	–	–

* indicates $0.01 < p < 0.05$, while ** indicates $p < 0.01$.
† indicates marginally significant.

To compare the results with previous findings, in Figure 1 we plot the average values of ΔC and %V for different languages, including Mandarin [13], Cantonese [8], and seven other languages [5]. It can be seen that native Mandarin and Cantonese should both be classified as syllable-timed, though the distance in %V from these two to the other four syllable-timed languages is even farther than between the four syllable-timed languages and the three stress-timed languages. For non-native Mandarin, CM and EM are quite close to each other. They share almost the same %V with MM, but both have a much larger ΔC than MM (even larger than the stress-timed languages), showing a tendency of shifting towards stressed-timed languages in the dimension of ΔC, regardless of the L1. On the whole, all the languages in Fig. 1 can be clustered into four sets in the ΔC vs. %V space; native Mandarin/Cantonese, as well as nonnative Mandarin, looks distinctly deviated from other languages, probably due to their monosyllabic nature.

In Figure 2, we plot the average values of nPVI_V and rPVI_C for different languages, including Cantonese [8] and five other languages [6]. It can be seen that native Mandarin and Cantonese should both be classified as syllable-timed (the position of Spanish in [6] is somewhat suspicious because it differs from the results in [10, 14] substantially). For non-native Mandarin, CM and EM are close to each other. They differ from MM mainly in rPVI_C, showing a tendency of shifting towards stressed-timed languages in the dimension of rPVI_C, regardless of the L1.

Among four syllabic metrics, Mok [8] found that rPVI_S gave the best separation between stress-timed and syllable-timed languages, ΔS gave the second best separation, while nPVI_S and VarcoS were not good separators. As shown in

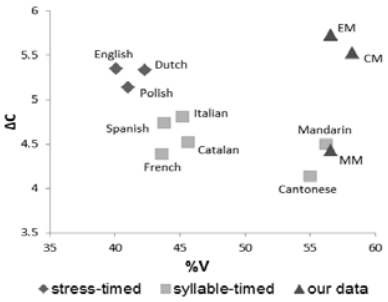


Figure 1: *ΔC and %V for different languages.*

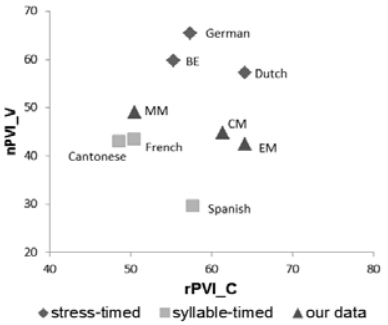


Figure 2: *nPVI_V and rPVI_C for different languages.*

Table 2, the syllabic metric that gives the best distinction among three groups is rPVI_S, followed by ΔS. This is in line with the result for separating stress-timed and syllable-timed languages in [8]. To compare with previous findings, Table 3 lists the average values of rPVI_S and ΔS, for three groups of subjects here and for six languages examined in [8] – English and German are stress-timed while others are syllable-timed.

We have noticed that MM in the present study and Mandarin in [8] give substantially different values. This is quite possibly due to speech rate difference between the materials used in the two studies – hence caution should be made in comparing absolute rhythmic scores between studies, as suggested in [10]. In spite of this, CM and EM showed a tendency of shifting from MM towards the stressed-timed class in terms of both rPVI_S and ΔS. This also coincides with the results for ΔC and ΔV (the latter is not illustrated here). It should be noted that rPVI_S, ΔS, ΔC, and ΔV showing the same tendency are all non-rate-normalized metrics.

Table 3. *Syllabic metrics for different languages.*

Language	rPVI_S	ΔS
English	115.50	88.74
German	99.62	80.78
Mandarin	86.08	75.80
Italian	82.68	67.61
French	75.89	55.30
Cantonese	63.62	57.48
EM	75.70	63.41
CM	72.44	60.32
MM	60.20	51.91

4. Discussion

Comparison of rhythmic characteristics for CM, EM, and MM on the basis of durational metrics has shown that nonnative Mandarin gives almost the same rhythmic mode (i.e., syllable-timed) as native Mandarin in terms of rate-independent (%V) or rate-normalized metrics, but it is shifted towards the stress-timed class in terms of raw durational metrics, regardless of the rhythmic class of the L1. Surprisingly, this does not accord with our expectation from the observation on isolated words, and also conflicts with the general hypothesis (i.e., L1 transfer effect) that the rhythmic pattern of nonnative speech should be intermediate between L1 and L2. However, this does coincide with our auditory impression of nonnative Mandarin speech – basically syllable-timed but meanwhile accompanied by mistaken stress assignment and relatively frequent breaks.

The first conflict can be explained by the fact that stress in continuous speech is not a simple copy of lexical stress pattern but is affected by many other factors. This has been confirmed by our auditory impression that some nonnative subjects assigned stress mistakenly in their utterances, enlarging the contrast of stressed vs. unstressed syllables. It is more difficult to interpret the second conflict, for which the differences in speech rate (viz., nonnative speech is generally slower due to the lack of fluency) can be the best explanation, especially for the larger raw metrics for nonnative speech. In addition, a complex and selective lengthening effect may also account for the observed results, as already mentioned in [9, 10].

This can be further analyzed from the view of prosodic structure. Table 4 shows the numbers of prosodic boundaries at different layers for all subjects. While there is little difference between CM and EM, it is obvious that nonnative speakers produced more prosodic boundaries than natives, especially for B1. The total numbers of higher-layer prosodic boundaries (B2 and B3) are similar, but nonnative subjects have fewer B2 and more B3 than natives.

The results in Table 4 are in line with our auditory impression that nonnative subjects were less fluent in speech and tended to divide an utterance into more chunks, sometimes with longer pauses in between, and even produced some words syllable by syllable. Table 5 lists the distribution of prosodic words with different numbers of syllables for all subjects, showing that CM and EM produced monosyllabic and disyllabic prosodic words more frequently than MM. Besides the lack of fluency, the stronger monosyllabic nature of Cantonese (than Mandarin) may also contribute to the large number of monosyllabic prosodic words for CM.

Because prosodic boundaries are usually accompanied by segmental lengthening such as word-initial and phrase-final lengthening, insertion of more prosodic boundaries leads to more lengthening, which may contribute to a higher variation of duration, causing the larger raw durational metrics.

Table 4. *Numbers of prosodic boundaries.*

Prosodic boundary	CM	EM	MM
B1	1047	1001	801
B2	144	165	247
B3	271	295	196
Total	1462	1461	1244

Table 5. *Numbers of prosodic words in various lengths.*

Group	1-syl	2-syl	3-syl	4-syl	5-syl	≥ 6-syl
CM	226	820	310	84	13	9
EM	184	849	338	74	12	4
MM	106	585	316	166	44	27

5. Conclusions

We have compared the rhythmic patterns of Mandarin speech between native speakers and two groups of L2 speakers who were native in Cantonese and English, respectively. Study on isolated words showed that nonnative speakers did not reduce unstressed syllables adequately. For continuous speech, eleven durational metrics were used to analyze rhythmic patterns. It was shown that nonnative Mandarin gave a quite similar rhythmic mode as native Mandarin in terms of rate-normalized or rate-independent metrics, but shifted towards stress-timed languages in terms of raw metrics, regardless of the rhythmic class of the L1. The result can be explained by speech rate difference and lengthening effects associated with the change in prosodic structure. It coincides with our auditory impression of L2 speech, for which the perceived rhythm may not be classified in a traditional binary/ternary way, but is usually a mixture. In this case, we may need to find better metrics, and the relationship between rhythmic metrics and other measures related to fluency and naturalness needs to be further studied.

6. Acknowledgements

This work is supported jointly by the National Social Science Fund of China (10CYY009 and 13BYY009), the Major Programs for the National Social Science Fund of China (13&ZD189), and the key project funded by the Jiangsu Higher Institutions' Key Research Base for Philosophy and Social Sciences (2010JDXM024).

7. References

[1] Pike, K.L., The Intonation of American English, Ann Arbor: University of Michigan Press, 1945.

[2] Abercrombie, D., Elements of General Phonetics, Edinburgh: Edinburgh University Press, 1967.

[3] Ladefoged, P., A Course in Phonetics, New York: Harcourt Brace Javanovich, 1975.

[4] Dauer, R.M., "Stress-timing and syllable-timing reanalyzed", Journal of Phonetics, 11: 51-62, 1983.

[5] Ramus, F., Nespor, M., and Mehler, J., "Correlates of linguistic rhythm", Cognition, 73: 265-292. 1999.

[6] Grabe, E. and Low, E.L., "Durational variability in speech and the rhythm class hypothesis", in N. Warner and C. Gussenhoven [Eds], Papers in Laboratory Phonology, 7: 515-546, Berlin: Mouton de Gruyter, 2002.

[7] Low, E.L., Grabe, E., and Nolan, F., "Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English", Language and Speech, 43: 377-401, 2000.

[8] Mok, P., "On the syllable-timing of Cantonese and Beijing Mandarin", Chinese Journal of Phonetics, 2: 148-154, 2009.

[9] Mok, P. and Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", Proc. Speech Prosody, pp. 423-426, Campinas, Brazil, 2008.

[10] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics, 35: 501-522, 2007.

[11] Li, A., "Chinese prosody and prosodic labeling of spontaneous speech", Proc. Speech Prosody, 39-46, Aix-en-Provence, France, 2002.

[12] Dellwo, V., "Rhythm and speech rate: A variation coefficient for ΔC", in P. Karnowski, & I. Szigeti [Eds], Language and language Processing, 231-241, Frankfurt: Peter Lang, 2006.

[13] Lin, H. and Wang, Q., "Mandarin rhythm: An acoustic study", Journal of Chinese Language and Computing, 17(3): 127-140, 2007.

[14] Ramus, F., "Acoustic correlates of linguistic rhythm: Perspectives", Proc. Speech Prosody, 115-120, Aix-en-Provence, France, 2002.

12 Thursday 1

Laughing, Breathing, Clicking - The Prosody of Nonverbal Vocalisations

Jürgen Trouvain

Phonetics, Saarland University, Saarbrücken, Germany
trouvain@coli.uni-saarland.de

Abstract

When analysing human spoken communication the focus on the linguistic side lies on speech with its verbal message, whereas the focus on the non-linguistic side usually is on the visually transported information such as gestures and facial expression. However, speech, especially in talk-in-interaction, also features numerous nonverbal vocalisations including various forms of laughter and inhalation noises as their most frequent forms. Although nonverbal vocalisations are usually short in duration they may provide rich information on linguistic, paralinguistic and extralinguistic levels including prosodic phrasing, cognitive load, affective state or speaker identity. The paper provides an overview of the phonetic and prosodic structure and the timing of laughter and audible breathing. Special attention is given on conversational speech, where we can frequently find situations in which interlocutors overlap temporally and on apical click sounds that often occur with inhalation before upcoming articulation but also during word-finding difficulties.

Index Terms: laughing, speech respiration, clicks, paralinguistics, pauses

1. Introduction

Speech communication is concerned with the analysis and processing of *verbal* communication. In contrast, *nonverbal* communication is often associated with *visual* information like gestures from hands, arms, eyes and other parts of the body, particularly the face. However, there is also *vocal* nonverbal communication. Crystal [6] divides the paralinguistic features into *voice qualifiers* (such as whispery, breathy or creaky voice) and *voice qualifications* (like laugh, giggle, sob or cry). The latter group belongs with physiological reflexes (like sneezing, coughing or snoring), to non-word vocalisations that are termed nonverbal vocalisations (NVVs) [47] or non-lexical sounds [51].

Although it is a matter of dispute what counts as verbal and what as nonverbal, for many vocalisations there is no doubt about their non-linguistic status, such as for *vegetative sounds* or physiological reflexes. Snoring, moaning (e.g. in sports), swallowing sounds, chewing noises, hiccup, coughing, sneezing, clearing the throat, yawning or panting (after physical exercise) are not primarily communicative and not all are under voluntary control. Typically, vegetative sounds are not learned. However, there are vegetative sounds that require some level of learning such as lip smacking or blowing one's nose. Some vegetative sounds can be used deliberately like clearing the throat ("ehem") to indicate e.g. "I'm here now". Thus, deliberate vegetative sounds require pragmatic knowledge and the control of the vocal apparatus.

Affect sounds include vocalisations such as laughing, weeping, cheering, crying aloud or screaming and many other types. Conventionalised forms of these affect sounds include the deliberate use of moaning and yawning as well as imitations of coughing and snoring. Often, these vocalisations are called affect bursts, e.g. [35].

Sometimes all sorts of NVVs occur under the umbrella of *interjection*s or *sound objects* [33] as words or utterances with either an emotional and/or an affective connection such as "ouch" or "wow" or as imitative expressions like "miaow" or "knock-knock". There are various grades of lexicalisation among the interjections: "Damned!" or "Shit!" are clearly verbal vocalisations whereas "woosh" or "bing" seem to be less conventionalised. Some interjections are affective words with an ungrammatical phonology such as "pst" or "shh" (no vowels) and "ts-ts-ts-ts" (clicks).

There are further potential candidate utterances for NVVs on the basis of their over-simple or ungrammatical phonotactics. *Hesitation particles*, also known as fillers or filled pauses, such as "uh" or "uhm", can be phonetically regarded as targetless vowels plus a potential neutral nasal consonant. *Feedback utterances* or response tokens include humming signals like "hm" or "yeah" and "uhu" which require hardly any vocal tract control. Usually they are used as backchannel signals but potentially also for asserting and other kinds of attitudinal expression in conversations.

A universal phonetic behaviour is the use of *melodies* with one's own vocal apparatus. Melodies without text can be hummed, sung or whistled but probably not in everyday conversation. In this context the melodies and utterances of babies and toddlers in their pre-linguistic phase should be mentioned where prosody is presumably used without any articulatory targets in the vocal tract.

In conversations, as the most common form of speech communication, NVVs seem to occur more often than in read aloud speech and other forms of controlled speaking situations. An analysis of NVVs in six annotated corpora of conversational speech [47] revealed that breathing noises and laughter were by far the most frequent NVVs in the inspected data. Interestingly, laughs were always present as an annotation category in the different corpora whereas breathing (or similar concepts such as in- or exhalation) was not. Hesitation particles and feedback utterances were usually considered as words and were therefore not counted as NVVs.

In the following sections an overview is given with respect to laughing, breathing and clicking - three phenomena that are linked to the prosodic concept of *pause*. Usually pauses are classified as filled and unfilled pauses [15] whereas the latter are often regarded as silent pauses. Admittedly this 'silence' often contains audible phonetic activity.

2. Laughing

Stereotypically, laughter is associated with a vocal expression of joy that is often spelled "haha". However, investigations of the acoustics of laughter show a huge range of variability for several parameters of "haha"-like laughter. For instance the 'vowel' as the vocal tract reflection in laughing is highly variable between laughs but also in the same laughs [1]. This is also valid for the number of reduplicated laughter 'syllables'

and the duration of laughs. In longer laughs, inspiratory breathing can occur at locations of the 'consonant' and there can be an onset and an offset before and after the staccato-like structure of the "haha"-syllables. We note that such a stereotypical laugh shows a great degree of variability and also complexity [42], see Fig. 1[1].

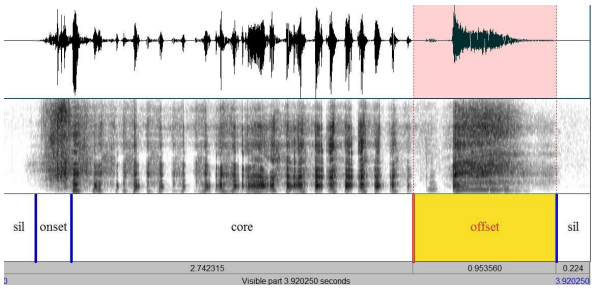


Figure 1: *A complex laugh (as a feedback utterance) with a strong inhalation as an offset (spk R06, duration: 3.9 sec).*

The stereotypical laughs belong to *song-like* laughs [1]. However, not all laughs are alike, as Bachorowksi et al. [1] show in several studies. One important distinction is whether a laugh is voiced or unvoiced [16]. Lab experiments show that females and males use voicing in laughs differently (females prefer voiced laughs) [1, 16]. Unvoiced laughs also have a tendency for conspiracy and less trust [5]. Forms of unvoiced laughter include snort-like and grunt-like variants and sometimes just a short forced exhalation. It becomes clear that laughter is *not only one* form with some variations of this main form but rather *a bundle of forms* with variations for each of these forms.

A special form is the so-called speech-laugh [29, 41] where laughter is produced simultaneously with articulated speech. Note that speech-laughs are different from smiled speech, and that speech can be affected by both, laughing and smiling. In Crystal's terminology [6] speech-laughs and smiled speech belong to voice qualifiers whereas laughs, coughs, sneezes and other NVVs belong to voice qualifications. Speech-laughs can represent a considerable number of all laughs in spontaneous data sets [29, 41, 49] and they are mainly used as self-comments.

Laughter is mainly associated as a signal of emotion displaying happiness, joy, amusement and other forms of well-being. Apart from these positive characteristics there are also negative emotions like maliciousness or simply nervousness. Further important functions of laughter are social bonding or creating affiliation [19]. Presumably, one's decision to join a laughing event, or not to join it, can serve various social functions in addition to transporting affective information.

Investigations of laughs in dialogues reveal that in a considerable number of laughs both speakers overlap with each other (Fig. 2 - as a contrast see a laugh overlapping with speech in Fig. 3). This shared laughter often happens at locations where the speaker with the turn invites the partner to join in the common laugh in order to take the turn [19, 28, 48].

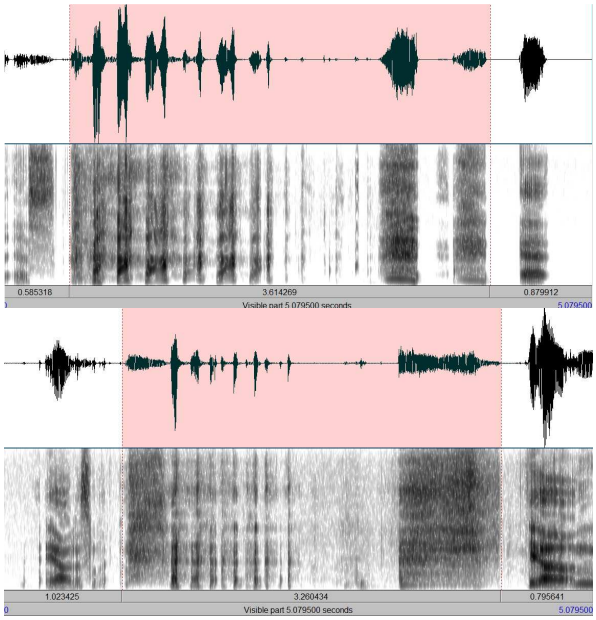


Figure 2: *Overlapping laughter (coloured) with speaker at top (L06) starting to laugh after his speech, speaker at bottom (R06) joining in and taking the turn (duration: 5.0 sec).*

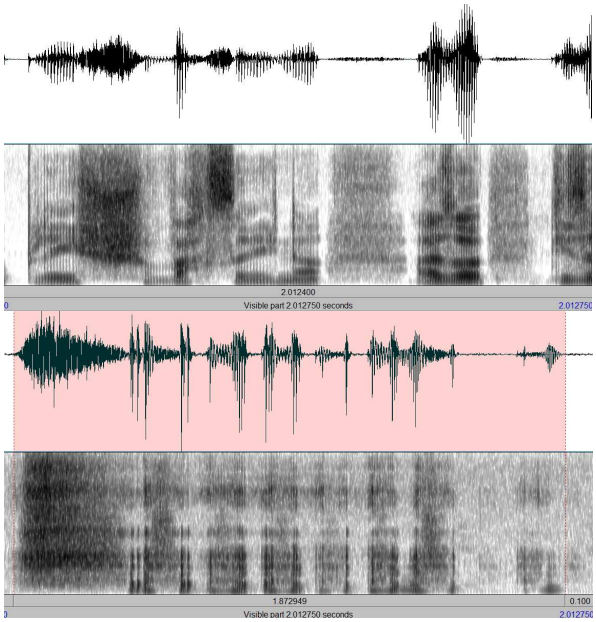


Figure 3: *A laugh at the bottom (R06), starting with an inhalation onset, overlapping with speech of the speaker at the top (L06) (duration: 2.0 sec).*

The prosodic organisation of laughter can be considered at three different levels:

- *The laugh itself.* A voiced laugh follows certain variations of rhythmical, pitch and intensity patterns. Song-

[1] All figures show waveforms and spectrograms (0-8 kHz) of excerpts from the "Lindenstrasse" corpus [18] (six German dialogues, separate channel for each speaker). Acronyms like L06 refer to the left-channel speaker from dialogue 06. L/R02 are female, L/R 06 are male.

like laughter is sometimes characterised as staccato-like. However, a strict application of staccato-like replications of "laugh-syllables" in manipulated laughter leads to the percept of unnaturalness [21].

- *Laugh integrated in the speech*. Laughs are not independent of the preceding articulation phases, e.g. the intensity of a laugh is adapted to the intensity of the preceding speech (cf. [46]). In addition, speech-laughs as vocal productions of speech and laughing by the same speaker at the same time are paralinguistic voice qualifiers and can be seen as tone of voice.
- *Laughing as a construct of interactional behaviour*. A considerable number of laughs in conversations are produced as speaker-overlapping vocalizations. These overlapping laughs (laugh of one speaker overlaps with laugh of the other) show significantly higher values in terms of fundamental frequency, intensity, duration and voicing [48].

3. Breathing

Respiration in speech usually leads to audible noises of breathing which can strongly vary between individuals in terms of duration and intensity of the frication noise [23]. Inhalation noises can be distinguished from exhalation noises and both types can occur as oral or nasal or combined oral-nasal sounds [20].

In read but also in spontaneous speech, audible inhalation noises are usually found in pauses at major prosodic breaks, while pauses that include breathing noises are generally longer than those without breathing [17]. There seems to be a correlation between breath pauses and higher-ranked constituents of the prosodic hierarchy. However, in many prosodic annotation schemes such as ToBI [2], breathing information is not used for determining the boundary strength but should be treated under 'miscellaneous'. Obviously respiration plays a role in controlling and planning linguistic units of various size. The planning of longer phrases is usually indexed by a deeper inhalation with a subsequent, more intensive and/or longer inhalation noise compared to the planning of shorter phrases [11, 34, 53].

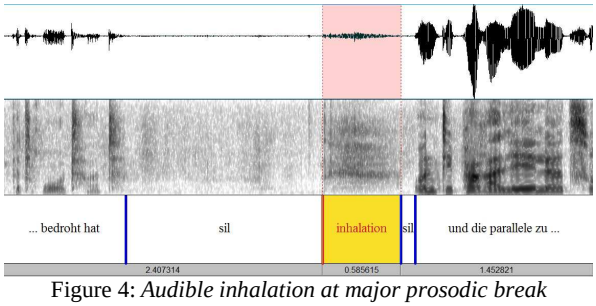


Figure 4: *Audible inhalation at major prosodic break after silence (filled by interlocutor with feedback expression) and before next phrase (speaker R02, duration: 4.5 sec).*

Despite the correlation between inhalation and prosodic planning, the respiratory kinematics of inhalation does not necessarily lead to *audible* breathing noise, as is the case in quiet breathing but also sometimes in speech [45].

Audible breathing can have an impact on how listeners perceive speech tempo. For instance, horse race commentaries are usually described as getting faster the closer the horses are coming to the finish. However, acoustic analyses of those commentaries [43] reveal that the articulation rate remains the same over the race and that the number of pauses increases towards the end instead of decreasing as expected. The main characteristic of these pauses in the final part is that they are shorter and filled with strong inhalation noise - together with an immense increase of the mean pitch.

Paralinguistically, inhalation noises are used for the display of affect such as startle and surprise [35]. Breathing noises are also components of cultural patterns, e.g. as markers of politeness in Korean [54].

Individual patterns of audible breathing and its acoustic correlates are of great interest for forensic research. Duration, intensity and spectral distribution of the frication noise indicate typical differences between individuals [23, 24]. It remains an open question to what extent audible breathing patterns can be reliably associated with the corresponding speaking voice and whether it is possible to impersonators to imitate the breathing of another person.

Attempts to integrate inhalation sounds in speech synthesis are rare. Studies either tested the recall rate and the preference of synthesised sentences with and without preceding breath sounds [52, 44], or inhalation noises were integrated in expressive synthesis as affective sounds [39]. A further example is the modelling of inhalation pauses for speech synthesis beyond single-sentence prosody [3].

As already mentioned in the previous section, audible inhalation can represent a substantial part of laughter, either as an inter-vocalic part dividing two bouts of a voiced song-like laugh or as an offset in a complex laugh (see Fig. 1). Inhalation also plays a role in producing click sounds (in non-click languages) which will be presented in the next section.

4. Clicking

Usually clicks are associated with phonemes occurring in languages in the Southern part of Africa [25, 26]. In contrast to vowels and pulmonic consonants clicks are plosives produced with an ingressive velaric airstream mechanism [22] with two closures, one in the front (i.e. with the lips or the tip of the tongue) and one in the back at the velum, forming a small pocket of air. While both closures are maintained, the tongue moves down. The result is an enlargement of the air pocket and thus a decrease of the pressure of the air therein. Then the front closure is released, followed by the release of the back closure resulting in the click sound.

In 'non-click languages' apical clicks are used as paralinguistic signals, e.g. to express disapproval but also to indicate "yes" or "no" [12]. They can also be used for imitation (e.g. horses) or for addressing animals. Another type are so-called *weak clicks* which are non-intended sounds that come as a coarticulatory by-product [10, 27, 38] when consonants with a closure at the alveolar ridge are followed by a consonant with a velar closing gesture. The release of the alveolar closure followed by the release of the velar closure can lead to a click with a low intensity. Moreover, clicks are used in musical styles with vocal percussion such as beatboxing [32].

In contrast to all the aforementioned uses and types of clicks, recent studies of conversational data in non-click languages such as English and German show that clicks are frequently used for marking new sequences [14, 31, 40, 55, 56], sometimes also to express a stance [31] but also before

feedback utterances and when searching for the right words [40], see Figs. 5 and 6.

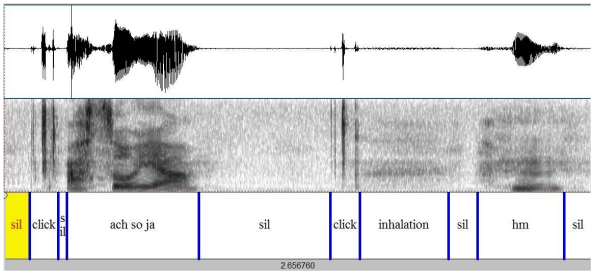


Figure 5: *Two clicks: first click between silence and the feedback utterance "ach so ja" ("I see, yes"), second click between silence and inhalation noise, followed by silence and the feedback token "hm" (speaker R02, duration: 2.6 sec).*

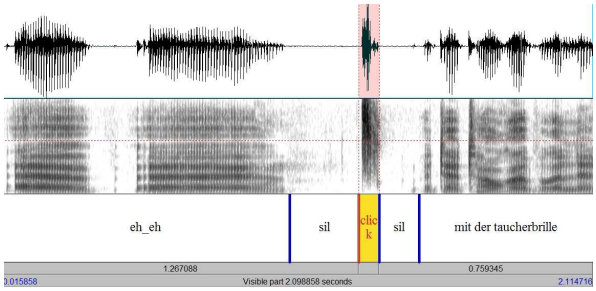


Figure 6: *A click during a word search with two hesitation particles followed by silence, the click, silence and fluent speech (speaker R06, duration: 2.1 sec).*

Interestingly, most speakers of the inspected data sets seem to use clicks, although individual speakers show different frequencies of occurrences. Despite the observed individuality there are limitations of the individual clicking behaviour to be reliably used in forensic phonetics [14].

Producing an apical click is normally unproblematic for a speaker of a "non-click language". The articulation of quasi-lexical items such as "ts-ts" to express disapproval requires a planning of the apical and a tongue body gesture together with inhalation. In contrast to such a conscious choice, the clicks in spontaneous discourse probably occur as by-products of an increased inhalation. Articulatory measurements of speech in preparation show that the tip of the tongue together with the increased inhalation may cause click sounds [37]. Although a velaric airstream cannot be excluded, it is more likely that an inhalation gesture with a sudden and strong vertical downwards movement of the larynx combined with an increased glottal opening provides the necessary negative pressure [9].

It remains an interesting topic for future research whether there is a universal tendency to use clicks as indices for new sequences and other pragmatic functions. The variability of the paralinguistic clicks across languages is still to be explored, particularly their phonetic substances. More knowledge about the phonetics of clicks in languages other than English and German is needed before we can reach a better understanding of what is a vegetative by-product and what is part of a linguistic system.

5. Discussion and conclusion

In vocal communication, particularly in interaction with interlocutors, numerous nonverbal vocalisations can be found. Despite various attempts to describe NVVs [6, 7, 30, 47, 51], a generally accepted framework including a theoretical foundation is missing. For instance, it is a matter of debate whether and which NVVs should be considered as 'conversational grunts' [7, 51].

From the perspective of vocal production, NVVs seem to entail no (or a low) active control of vocal tract configurations. NVVs are mainly characterised by their activities at the subglottal and glottal level and by their temporal control.

Looking at the prosody of NVVs we can see that some NVVs like a voiced laugh (see Fig. 1) show a complex prosodic make-up that can be regarded independently of its neighbouring context. Often NVVs like inhalation noises reveal how they are embedded in their context. Duration and intensity of an inhalation noise provide information on various levels: on the prosodic-syntactic level about the length of the upcoming phrase [11, 53] and the strength of the break [17]; on the prosodic-pragmatic level together with clicks about new sequences [55]; on the paralinguistic level about the degree of arousal [35, 43]; and on the extralinguistic level about patterns typical for certain individuals [13, 23, 24]. In talk-in-interaction, speakers change the prosodic shape of NVVs in accordance with the communication partner as it is the case for speaker-overlapping laughter [48].

Many NVVs occur in speaking situations that are listener-oriented and embedded in a communicative context. They are rarely observed (and not always welcome) in speaker- and text-oriented speech. Speech synthesis, mainly performed as text-to-speech conversion of single utterances, should not ignore NVVs when the aim is to generate more natural and more expressive speech [4]. Synthesis of laughter in isolation [50] is a promising starting point. However, conversational speech synthesis requires a deeper knowledge of the prosody of NVVs. On the recognition side in speech technology, initiatives like the paralinguistic challenges [36] show that NVVs are seen as important social signals.

Most spontaneous discourse (beyond single utterances) contains NVVs. This paper attempted to present the complexity of NVVs by touching upon laughing, breathing and clicking. For a better understanding of the prosody of vocal communication it is time to move the NVVs from the 'miscellaneous' tier to a more systematic description.

6. Acknowledgements

The author would like to thank Bernd Möbius for valuable feedback on an earlier draft of this paper.

7. References

[1] Bachorowski, J.-A., Smoski, M.J. & Owren, M.J. "The acoustic features of human laughter", JASA 111: 1582-1597, 2001.

[2] Beckman, M.E. & Ayers Elam, G. "Guidelines for ToBI Labelling", version 3.0, March 1997, Ohio State University.

[3] Braunschweiler, N. & Chen, L. "Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS", Proc. 8th ISCA Speech Synthesis Workshop, Barcelona, 2013.

[4] Campbell, N. "Conversational speech synthesis and the need for some laughter", IEEE Transactions on Audio, Speech, and Lang Proc 14: 1171 - 1178, 2006.

[5] Cirillo, J. & Todt, D. "Perception and judgement of whispered vocalisations", Behaviour 142: 98-125, 2005.

[6] Crystal, D. "Prosodic Systems and Intonation in English", Cambridge University Press, 1969.

[7] Dingemanse, M., Torreira, F. & Enfield, N. J. "Is 'Huh?' a universal word? Conversational infrastructure and the convergent evolution of linguistic items. PLoS One 8(11), 2013.

[8] Dryer, M.S. & Haspelmath, M. (eds.) The World Atlas of Language Structures Online. Leipzig: MPI for Evolutionary Anthropology. (Accessed on 2014-01-14 at http://wals.info/.)

[9] Fuchs, S. & Rodgers, B. "Negative intraoral pressure in German: Evidence from an exploratory study", JIPA 43: 321-327, 2013.

[10] Fuchs, S., Koenig. L.L. & Winkler, R. "Weak clicks in German?" Proc. 16th ICPhS, Saarbrücken, 449-453, 2007.

[11] Fuchs, S., Petrone, C., Krivokapić, J. & Hoole, Ph. "Acoustic and respiratory evidence for utterance planning in German", J Phon 41: 9-47, 2013.

[12] Gil, D. "Para-linguistic usages of clicks", In: [8] Dryer & Haspelmath (eds.) WALS online. Chapter 142, 2013.

[13] Gold, E. & French, P. "An international investigation of forensic speaker comparison practices", Proc. 17th ICPhS, Hong Kong, 751-755, 2011.

[14] Gold, E., French, P. & Harrison, P. "Clicking behaviour as a possible speaker discriminant in English", JIPA 43: 339-349, 2013.

[15] Goldman Eisler, F. "Psycholinguistics: Experiments in spontaneous speech", New York: Academic Press, 1968.

[16] Grammer, K. & Eibl-Eibsfeldt, I., "The ritualisation of laughter. In: Koch, W.A. (ed) Die Natürlichkeit der Sprache und der Kultur. Bochum: Brockmeyer: 192-214, 1990.

[17] Grosjean, F. & Collins, M. "Breathing, pausing and reading", Phonetica 36: 98-114, 1979.

[18] IPDS "Video Task Scenario: 'Lindenstrasse' – The Kiel Corpus of Spontaneous Speech" (Volume 4, DVD) Institut für Phonetik und Digitale Sprachsignalverarbeitung, University of Kiel.

[19] Jefferson, G., Sacks, H. & Schegloff, E., "Notes on laughter in the pursuit of intimacy", In: Button & Lee (eds.), Talk and Social Organisation, Clevedon: Multilingual Matters: 152–205, 1987.

[20] Kienast, M. & Glitza, F. "Respiratory sounds as an idiosyncratic feature in speaker recognition", Proc. 15th ICPhS, Barcelona: 1607-1610, 2003.

[21] Kipper, S. & Todt, D. "The role of rhythm and pitch in the evaluation of human laughter", J Nonverbal Beh 27: 255-272, 2003.

[22] Ladefoged, P. & Maddieson, I. "The Sounds of the World's Languages", Oxford: Blackwell, 1996.

[23] Lauf, R.. "Aspekte der Sprechatmung: Zur Verteilung, Dauer und Struktur von Atemgeräuschen in abgelesenen Texten", In: Braun, A. (ed.) Beiträge zu Linguistik und Phonetik. Stuttgart: Franz Steiner Verlag: 406-420, 2001.

[24] Link, L. "Individualtypische Aspekte des Atemgeräusches. Eine experimentalphonetische Untersuchung", M.A. thesis. Marburg University, 2012.

[25] Maddieson, I. "Presence of uncommon consonants", In: [8] Dryer & Haspelmath (eds.) WALS online, Chapter 19.

[26] Maddieson, I., "Patterns of sounds", Cambridge: CUP, 1984.

[27] Marchal, A. "Des clics en français?" Phonetica 44: 30–37, 1987.

[28] McFarland, D.H., "Respiratory markers of conversational interaction " J Sp Lang Hear Res 44:128-43, 2001.

[29] Nwokah, E.E., Hsu, H.-C., Davies, P. & Fogel, A. "The integration of laughter and speech in vocal communication: a dynamic systems perspective. " J Sp Lang Hear Res 42: 880-894, 1999.

[30] O'Connell, D.C. & Kowal, S. "Communicating with One Another", New York: Springer, 2008.

[31] Ogden, R. "Forms and functions of clicks in English conversation", JIPA 43: 299-320, 2013.

[32] Proctor, M. Bresch, E., Byrd, D., Nayak, K. & Narayanan, S. "Paralinguistic mechanisms of production in human 'beatboxing': A real-time MRI study", JASA 133: 1043-1054, 2013.

[33] Reber, E. "Affectivity in Interaction: Sound objects in English", Amsterdam/Philadelphia: John Benjamins, 2012.

[34] Rochet-Capellan, A. & Fuchs, S. "The interplay of linguistic structure and breathing in German spontaneous speech", Proc. Interspeech, Lyon: 2014-1018, 2013.

[35] Schröder, M. "Experimental study of affect bursts", Speech Communication 40: 99-116, 2003.

[36] Schuller, B. et al. "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism", Proc. Interspeech, Lyon, 2013.

[37] Scobbie, J.M., Schaeffler, S. & Mennen, I. "Audible aspects of speech preparation", Proc. 17th Int'l Congress of the Phonetic Sciences, Hong Kong, 1782-1785, 2011.

[38] Simpson, A.P. "Acoustic and auditory correlates of non-pulmonic sound production in German", JIPA 37: 173-182, 2007.

[39] Sundaram S. & Narayanan S. "An empirical text transformation method for spontaneous speech synthesizers", Proc. Interspeech, Geneva: 1221-1224, 2003.

[40] Trouvain, J. "Clicks in German", submitted.

[41] Trouvain, J. "Phonetic aspects of 'speech-laughs'", Proc. Confer. on Orality & Gestuality, Aix-en-Provence: 634-639, 2001.

[42] Trouvain, J. "Segmenting phonetic units in laughter", Proc. 15th ICPhS., Barcelona: 2793-2796, 2003.

[43] Trouvain, J. & Barry, W.J., "The prosody of excitement in horse race commentaries", Proc. ISCA-Workshop on Speech and Emotion, Newcastle (N. Ireland), 86-91, 2000.

[44] Trouvain, J. & Möbius, B. "Einatmungsgeräusche vor synthetisch erzeugten Sätzen -- Eine Pilotstudie", Proc. 24. Konfer. Elektron. Sprachsignalverarbeitung, Bielefeld: 50-55, 2013.

[45] Trouvain, J. & Möbius, B. "Individuelle Ausprägung von Atmungspausen in der Mutter- und in der Fremdsprache als Anzeichen kognitiver Belastung", Proc. 25. Konfer. Elektron. Sprachsignalverarbeitung, Dresden: 177-184, 2014.

[46] Trouvain, J. & Schröder, M "How (not) to add laughter to synthetic speech", Proc. Workshop on Affective Dialogue Systems, Kloster Irsee: 229-232, 2004.

[47] Trouvain, J. & Truong, K. "Comparing non-verbal vocalisations in conversational speech corpora", Proc. 4th Int'l Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul: 36-39, 2012.

[48] Truong, K. & Trouvain, J. "On the acoustics of overlapping laughter in conversational speech", Interspeech, Portland, 2012.

[49] Urbain, J. et al. "The AVLaughterCycle Database", Proc. LREC, Malta: 2996-3001, 2010.

[50] Urbain, J. "Acoustic Laughter Processing", PhD thesis, University of Mons, 2014.

[51] Ward, N. "Non-lexical conversational sounds in American English", Pragmatics and Cognition 14: 113-184, 2006.

[52] Whalen, D.H., Hoequist, Ch.E. & Sheffert, S. "The effects of breath sounds on the perception of synthetic speech", JASA 97: 3147-3153, 1995.

[53] Winkworth, A.L., Davis, P.J., Adams, R.A. & Ellis, E. "Breathing patterns during spontaneous speech", J Sp Lang Hear Res 38: 124-144, 1995.

[54] Winter, B. & Grawunder, S. "The phonetic profile of Korean formal and informal speech registers", J Phon 40: 808-815, 2012.

[55] Wright, M. "Clicks as markers of new sequences in English conversation", Proc. 16th ICPhS, Saarbrücken: 1069–1072, 2007.

[56] Wright, M. "On clicks in English talk-in-interaction", JIPA 41: 207–229, 2011.

Tuning in to whispered boundary tones

Willemijn Heeren [1], *Sarah Bibyk* [2], *Christine Gunlogson* [3], *Michael K. Tanenhaus* [2]

[1] Leiden University Centre for Linguistics, Leiden University, The Netherlands
[2] Department of Brain and Cognitive Sciences, University of Rochester, NY, USA
[3] Department of Linguistics, University of Rochester, NY, USA

w.f.l.heeren@hum.leidenuniv.nl, {sbibyk,mtan}@bcs.rochester.edu, gunlog@ling.rochester.edu

Abstract

Very little is known about how listeners incorporate "intonational" information in whispered speech during online language processing. We present data showing that listeners can incorporate information about boundary tones in whispered speech rapidly, but this process is complicated by additional structural biases as well as by the fact that speakers do not produce cues to boundary tones consistently in whisper. Listeners, however, are able to adapt to these differences in order to correctly identify different boundary tones in whisper.
Index Terms: boundary tones, online processing, whispered speech

1. Introduction

In earlier work we investigated the online processing of high (H%) versus low (L%) boundary tones in normal speech, and found that they are processed as quickly as pitch accents [1-5], and with very few interpretation errors. The main acoustic cue to boundary tones, and intonation in general, is thought to be the speaker's fundamental frequency (f0). But when a speaker whispers, f0 is not being produced. This paper begins to address how listeners process boundary tones produced in whispered speech when f0 is absent.

Offline studies have shown that listeners can identify and discriminate boundary tones in whispered speech [6,7]. Performance is worse than in normal speech, but well above chance level, indicating that there are prosodic cues to boundary tones available in whispered speech as well. However, it is unclear how listeners process whispered boundary tones online and which cues they use in doing so. Using a targeted language game, an acoustic analysis of multiple speakers' boundary tones, and a crowd-sourcing perception experiment, the online processing of and adaptation to whispered boundary tones were investigated.

2. Online processing of whispered boundary tones

In earlier research, we developed a "targeted language game" using the visual world eye-tracking paradigm [8] that is sensitive to the time-course of processing boundary tones: [9]. Here we used that paradigm to study the online processing of whispered speech. The participant played a card game against the computer by means of a verbal interaction, and the game was designed in such a way so that on critical trials only the boundary tone indicated whether the computer's move was a statement (signaled by L-L%) or a yes-no question (H-H%). Syntactic cues were removed by having the computer use elliptical sentences of the form '*Got a <card category>*', which could be elliptical versions of either '*I have got a <card category>*.' or '*Have you got a <card category>?*'. The game elicited different actions (thus different fixation patterns) from the participant in response to questions vs. statements, allowing us to assess listeners' online categorization of boundary tones.

The goal of the game was for the opponents to discard cards from their (virtual) hands by matching them to a match card, a face-up card in the middle of the screen (Fig. 1a). Each player also had a stack of block cards, the top one of which could be used to block the other's matches (Fig. 1b). Upon perceiving a question from the computer, the player would look at the playing cards, and upon perceiving a statement, the player would look at the block card. There were four card categories, each represented by a black and white line drawing shown on the match, playing or block card (shoe [ʃuː], wheel [wiːl], candy [kændi], window [wɪndoʊ]). The center of the match card was placed equidistant to the center of the mean size of the playing card set, and the center of the block card.

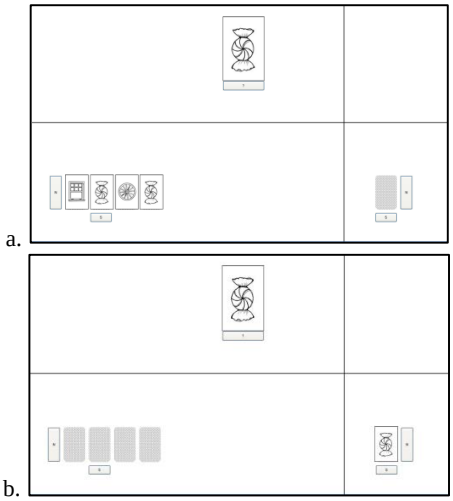


Figure 1: *Screenshots of the game with the match card (top center), the player's playing cards (bottom left) and block card (bottom right); the player (a) looks for possible matches when asked; and (b) looks to block the computer's stated move.*

2.1. Materials

The computer's whispered moves were pre-recorded: eight were target sentences (4 statements/ 4 questions) and nine were fillers (3 statements/ 6 questions). Fillers were used to introduce syntactic variation into the computer's speech (e.g., '*Do you have a candy?*'). Also recorded as computer 'utterances' were moves to keep track of turns during the game (e.g., '*It's your/my turn.*'), to respond to the player's

questions for a match (e.g., '*Yes.*' or '*No.*') and to block the player's match (e.g., '*I am blocking you.*'). The utterances representing the computer's moves were recorded in a sound-treated booth using an Audio-technica ATM75 head-worn microphone and a Marantz PMD 670 solid state recorder (mono, 32 kHz, 16 bits). The speaker was a 23-year-old female native speaker of American English.

Table 1 presents acoustic measurements taken over the target words' final syllables, that is the boundary tone landing sites. It shows mean intensity, duration, and the first through third formants, measured over the mid 50 ms of a vowel using the Burg method implemented in PRAAT [10]. The measurements show a comparable duration for statements and questions, a higher intensity in questions than statements, and in many cases higher formant values for questions.

Table 1. *Acoustic content of the final vowels, per target word, spoken as statement (S) or question (Q).*

	Speech act	Int. (dB)	Dur. (ms)	F1 (Hz)	F2 (Hz)	F3 (Hz)
candy	S	54.0	171	452	2720	3089
	Q	60.8	169	537	2925	3352
shoe	S	60.5	323	570	1679	2748
	Q	62.8	318	611	1832	2900
wheel	S	66.2	390	642	2551	2901
	Q	67.8	399	771	2653	3014
window	S	65.1	254	790	1766	2744
	Q	70.7	236	858	1683	2844

2.2. Participants and procedure

Fifteen American English participants were recruited at the University of Rochester, NY, USA (informed consent obtained). They were given both written and oral instructions. During testing the verbal interaction was recorded using a Realistic 33-984A Highball dynamic unidirectional table microphone placed between the computer speakers and the player, so that it would register both interlocutors. Eye movements were recorded using a head-worn ASL eye-tracker at 30 Hz, and a Sony DSR-30 digital video recorder, with Sony 184 DVCAM digital videotapes. Before the test started, the participant played a practice game that contained all possible game situations. Calibration was checked throughout the test game and always occurred before a new turn for the participant. The entire session lasted 30 to 45 minutes.

The order of events (question vs. statement) during the game was fully determined, but the order for the card items (candy, shoe, etc.) was rotated and balanced across four lists. For target utterances, the wave file started 1400 ms after the match card had changed. For filler items, the preceding silence was variable, but at least 1200 ms, thus introducing variation to increase the naturalness of the computer's utterances, as a real player also would not respond at regular intervals. The computer utterances were played to the participants at a comfortable listening level over computer speakers.

2.3. Results and discussion

The 33 ms video frames were coded manually from the onset of a target utterance until the participant's verbal response, i.e. a variant of '*I can('t) match/block*'. Five locations were coded: (1) the playing cards, (2) the match card, (3) the block card, (4) other on-screen locations, and (5) track loss. Saccade-initial frames were counted as fixations to the landing site.

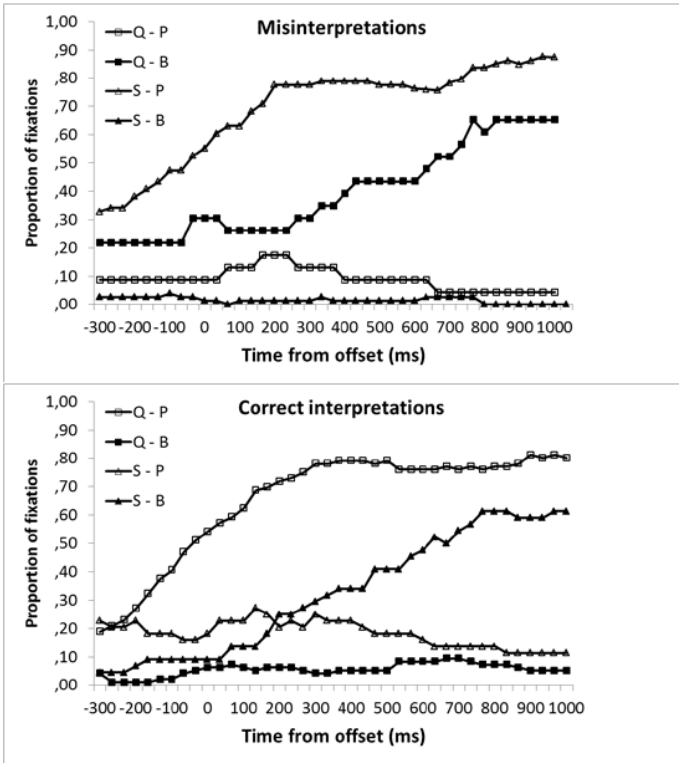


Figure 2: *Proportions of fixations to the playing cards (P) and block card (B), in the case of intended questions (Q) and statements (S). Misinterpretations and correct interpretations are plotted separately.*

Out of 240 trials, 140 were correctly interpreted (58%), 99 were misinterpreted (41%), and one trial was misunderstood. Misinterpretation means that intended questions were interpreted as statements, and the other way around. When looking at the division of correctly versus incorrectly interpreted trials per speech act, 76 out of 120 statements were misinterpreted, but only 23 out of 120 questions. On average, questions, but not statements, were correctly recognized above chance level.

Because of the large number of misinterpretations, eye movements to both correctly interpreted and misinterpreted trials were analyzed separately (see Fig. 2). For questions and statements the ratio of targets divided by the sum of targets and distractors, T/(T+D), was computed and compared in a two-tailed paired samples t-test, see [11]. This was done for two intervals, one before target offset (-300 ms to 0 ms) and one after target offset (0 ms to 300 ms). These intervals were chosen because they represent the regions before and after where the earliest effects of boundary tones are expected.

For correctly interpreted trials, no significant difference between the speech acts was found during the first analysis interval, $t(7)=0.8$, n.s, whereas during the second interval, a significant difference was found, $t(12)=2.7$, $p=0.020$. This pattern is in principle consistent with sensitivity to prosodic cues. When looking at the questions, however, there is an early preference for the playing cards that does not seem to be time-locked to the boundary tone (< 0 ms). In statements, the pattern of fixations is consistent with interpretation of the boundary tone, showing an increase in looks to the block card once the boundary tone becomes available (> 0 ms).

For misinterpreted trials, both analysis intervals showed a significant difference between the speech acts, $t(6)=-5.2$, $p=0.002$ and $t(8)=-5.1$, $p=0.001$, respectively. Eye-movements therefore do not seem to be time-locked to the prosodic events. In the case of misinterpreted questions (as statements), there is an initial bias to look at the blocking card early in the utterance, followed by a later increase in looks to the blocking card which is too delayed to reflect a time-locked response to the boundary tone. In the case of misinterpreted statements (as questions) there is an initial bias to look to the playing cards which begins to increase well before the final syllable.

The results show only a weak reliance on prosodic cues in whisper. Firstly, many trials were misinterpreted which indicates that prosodic cues to speech act were not correctly used or not used at all. Secondly, in correctly interpreted questions the proportion of fixations to the target increased well before boundary tone information became available. Only in correctly interpreted statements, does the time course of fixation proportions suggest that prosodic cues were used. Because of the predetermined order for moves in the game, the majority of statements occurred during the second half of the game. We used a mixed-effects logistic regression to explore the effect of trial number and speech act type on participants' answers. The model with the fixed and random effects structure most justified by the data (as accessed by model comparison) did not contain a significant intercept ($B=0.19$, $z=0.80$, $p=0.42$) nor a main effect of trial number ($B=0.057$, $z=0.60$, $p=0.55$), but it did show a significant main effect of speech act type ($B=-1.21$, $z=-6.54$, $p<0.001$); questions were more likely to be answered correctly than statements. The interaction between type and trial number was also significant ($B=0.22$, $z=5.25$, $p<0.001$). Separate logistic regressions by speech act type revealed that participants improved in performance across trials only on statements, $B=0.28$, $z=3.89$, $p<0.001$. On questions listeners actually became worse across trials ($B=-0.21$, $z=-2.99$, $p=0.003$), but remember that overall participants performed better on questions than on statements.

Earlier research [e.g. 6,7] and Table 1 suggest that acoustic cues to speech act type were present. There are several explanations as to why this information was not fully used. Listeners may either be relatively insensitive to the prosodic information that the speaker attempted to convey, perhaps because whisper as a speech mode is not used very often and the cue that normally carries boundary tones, f0, is absent. Listeners might also find it difficult to extract prosodic information because speakers may not provide consistent cues to intonation in whisper. These possibilities were addressed further in section 3.0.

3. Variation in boundary tone realization

Gotta-utterances from three additional speakers were recorded and speaker strategies for signalling prosody were compared to explore if the listeners' difficulty with prosodic cues in whisper may be explained by varying speaker strategies.

3.1. Method

Two male and one female native speaker of American English were recorded [Shure SM57 microphone, mono 44.1 kHz, 16 bit] using a script that took speakers through a game scenario intended to elicit each target utterance twice. Participants were told the outline of the game and then asked to imagine themselves as best as possible in the game scenario, saying the sentences how they would say them if they were really playing. They were not asked to make the questions and statements as acoustically distinct as possible; they were just told to be as clear as possible given that they would be whispering. They were also told to take their time and read the scenario descriptions carefully so that they would be in the right mindset for producing the utterances.

Recordings were annotated at the segment level, and for all final-syllable vowels we measured the relative syllable duration, mean intensity, and the first through third formants over the vowel's mid 50 ms using the Burg method implemented in PRAAT. Formant measurements were visually verified in the spectrogram (some F1s, mainly of [u], could not be determined). These acoustic characteristics were selected as they have been put forward as cues to whispered tones and/or intonation [12-15]. Comparisons between the speech acts were done per speaker, using paired samples Wilcoxon signed ranks tests.

3.2. Results and discussion

Relative vowel duration was longer in questions than statements for Speaker 1 only, $Z=-2.5$, $p=0.012$. The difference was marginal for Speaker 2 ($p=0.058$), and non-significant for Speaker 3 ($p=0.16$). Mean vowel intensity showed no significant differences between statements and questions for any of the speakers. Due to data sparsity, effects in F1 were hard to measure, but no significant differences were found, and no consistent trends were observed. For F2, Speaker 3 produced a significantly higher second formant (F2) in questions compared to statements ($Z=-2.1$, $p=0.036$). The mean difference was 82 Hz. Speaker 2 showed a tendency in the same direction with a 119 Hz mean difference ($p=0.077$). As for F3, speakers 1 and 3 had a higher value in questions than statements, with mean differences of 120 Hz and 96 Hz, respectively (both $Z=-2.5$, $p=0.012$). Speaker 2's data showed a comparable trend with a 92 Hz mean difference.

The results show that speakers vary in the acoustic dimensions that they use to signal different speech acts, along three dimensions that have been proposed as potentially contributing to the expression of intonation in whisper. Speakers provided both durational and spectral cues, but not in exactly the same way. Spectral differences, which are assumed to provide the most direct cue, were also present in the productions of Experiment 1's speaker. Duration may be taken as a secondary cue, through lengthening of the speech act that requires most production effort, the question. Two out of three speakers made a durational difference, but this had not been the case for the speaker of Experiment 1. Whereas that speaker seemed to vary vowel intensity with speech act, a comparable intensity difference was not found for the other three speakers.

4. Tuning in to whispered prosody

Taken together, results of Experiment 1 and the analysis of multiple speakers suggest that listeners' difficulty with the extraction of prosodic information from whispered utterances may arise because different speakers provide different cues to prosody in whisper. Therefore listeners require exposure to a particular speaker to learn the relevant cues that remain when f0 is absent. The fact that 'tuning in' took place is supported by the increased number of correctly interpreted trials in the second half of Experiment 1. To further investigate if participants are in fact tuning in to prosodic information or

were just making lucky guesses, listeners were exposed to different patterns of acoustic cues to test the hypothesis that they adapt to and use the cues in each case.

4.1. Method

A Web survey was conducted in which participants provided responses to whispered audio samples of statements and questions produced by one of the three speakers from Experiment 2. Participants were instructed that they would hear a whispering speaker play a simple card game, and that they would be asked to indicate whether they heard the player make a match (=statement response) or ask for a card (=question response), by clicking one of two buttons. Feedback on the correctness of their answers was provided, both during practice and testing. During practice, sentences from Experiment 1's fillers were used (i.e. different speaker, no elliptical structure). During testing, there were 16 trials: 2 repetitions of each target word (4) in each speech act (2).

Using the online crowd-sourcing service Amazon's Mechanical Turk, 258 Human Intelligence Tasks (HITs) were posted for (self-reported) American English participants (253 unique participants). Per speaker, four lists were used, each with a different order for the response options. In a pretest it was established that listeners' home equipment was set up to perceive whispered speech well. Ten participants were eliminated because sound files did not play properly during the pretest. Three additional participants were eliminated due to experimenter error, leaving 240 participants (80/speaker, 20/list). Participants were paid $2 for their efforts.

4.2. Analysis, results and discussion

The data were analysed with a hierarchical mixed-effects logistic regression with Trial Number, Speech Act Type, and Speaker as predictors using the maximal random effects structure as justified by the model, which included random intercepts by subject and by item, as well as random slopes by item over trial. The main effects of trial, type, and speaker were entered first, followed by the two-way interactions, followed by the three-way interaction.

We used model comparison to select the model most justified by the data. The final model included the intercept, and the main effects of trial, type, and speaker. This model accounted for a significant portion of the variance above and beyond the model with just the intercept $\chi^2(4)=22.62$, $p<0.001$. The intercept itself was significant ($B=0.22$, z=2.53, $p<0.05$), indicating that on average listeners performed above chance. The main effect of trial was significant ($B=0.056$, z=2.82, $p<0.01$); on average, as trial number increased, performance improved. The main effect of type was also significant ($B=-0.22$, z=-2.67, $p<0.01$); on average questions were more often answered correctly as compared to statements. The intercepts for Speakers 1 and 3 were significantly different from each other ($B=0.10$, z=3.27, $p<0.01$). The intercepts for Speakers 1 and 2 were not significantly different ($B=0.050$, z=1.57, $p=0.12$). The intercepts for Speakers 2 and 3 were marginally different ($B=0.052$, z=1.71, $p=0.088$). Separate logistic regressions for each speaker revealed a significant intercept for speaker 1 ($B=0.33$, z=2.67, $p=0.008$), a marginal intercept for speaker 2 ($B=0.199$, z=1.66, $p=0.097$), and a non-significant intercept for speaker 3 ($B=0.043$, z=0.31, $p=0.76$).

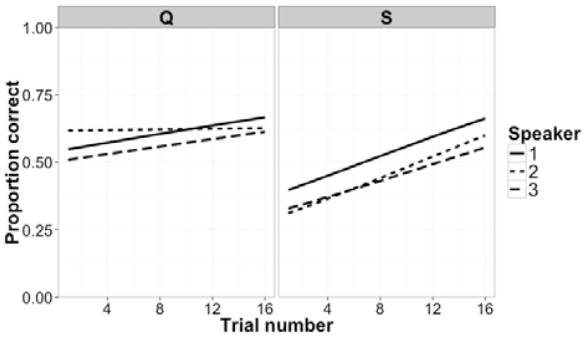


Figure 3: *Proportions of correct answers across trials plotted for the three speakers separated by questions (left panel) and statements (right panel).*

First, as Figure 3 shows, there was an overall trend across all speakers for listeners to improve performance as trial number increased. Thus, while it is true that different speakers signal prosody differently in whisper, listeners are able to adapt to those differences. This result contrasts with Experiment 1 where a main effect of trial was not found. Recall, however, that only one ordering for question and statement moves was used in Experiment 1, whereas four lists were used here. Second, there was a trend for listeners to perform better on questions than on statements (see Fig. 3), just as in Experiment 1. This difference seems to indicate a bias to interpret the "got a" construction as a question as opposed to a statement, see [16]. Third, we have some evidence that productions of different speakers have different effects on how listeners are able to interpret their intentions in whispered speech. The listeners who received Speaker 1's stimuli performed above chance overall (61% correct), in comparison to the listeners receiving Speaker 2, who were marginally above chance (58%), and Speaker 3 who were not statistically different from chance (54%). Possibly, listeners tuned in more quickly to Speaker 1 because that speaker significantly changed two acoustic dimensions, spectral and durational, between the speech acts rather than just one, spectral or durational.

5. Conclusion

We have replicated the effect that listeners can distinguish boundary tones in whispered speech. In addition we have provided provisional evidence that speakers can incorporate information about the boundary tones in real time processing of whispered sentences. This effect is complicated by the fact that listeners also appear to take into account cues from the lexical content of the utterance (a question bias for "got a"), and by the fact that cues to boundary tones are more difficult to distinguish in whispered speech. In addition, speakers have different strategies for how they signal boundary tones in whisper. Therefore, compared to the more systematic f0 cue in regular speech, listeners may need more exposure before they can utilize (speaker-dependent) systematic cues to prosody in whispered utterances. We note that there is emerging evidence that listeners adapt to the reliability of different prosodic cues for individual speakers even in normal speech [17]. Therefore, the mechanisms used by listeners to process prosody in whispered speech might partially be similar to those used when processing normal speech.

6. References

[1] Dahan D., Tanenhaus M. K. and Chambers C. G., "Accent and reference resolution in spoken-language comprehension", J. Mem. Lang., 47:292-314, 2002.

[2] Watson, D. G., Gunlogson, C. A. and Tanenhaus, M. K., "On-line methods for the investigation of prosody", in I. Mleinek [Ed.] Methods in Empirical Prosody Research. Berlin: Mouton de Gruyter, 2006.

[3] Weber, A., Braun, B. and Crocker, M. W., "Finding referents in time: Eye-tracking evidence for the role of contrastive accents", Lang. Speech, 49:367-392, 2006.

[4] Ito, K. and Speer, S. R., "Anticipatory effects of intonation: eye movements during instructed visual search", J. Mem. Lang., 58:542-573, 2008.

[5] Watson D. G., Tanenhaus M. K. and Gunlogson C. A., "Interpreting pitch accents in online comprehension: H* vs. L+H*", Cogn. Science, 32:1232-1244, 2008.

[6] Fónagy, J. (1969). "Accent et intonation dans la parole chuchotée," Phonetica, 20:177-192, 1969.

[7] Heeren, W. F. L. and Van Heuven, V. J., "Perception and production of boundary tones in whispered Dutch", in Proc. Interspeech 2009, Brighton2411-2414, 2009.

[8] Brown-Schmidt, S. and Tanenhaus, M. K., "Real-time investigation of referential domains in unscripted conversation: A targeted language game approach", Cogn. Science, 32: 643-684, 2008.

[9] Bibyk, S., Heeren, W., Gunlogson, C. and Tanenhaus, M. K., "Asking or telling? Real-time processing of boundary tones", LSA annual meeting, Boston, January 3-6 2013, 2013.

[10] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]," retrieved from http://www.praat.org/, 2013.

[11] Dahan, D. and Tanenhaus, M. K., "Looking at the rope when looking for the snake: conceptually mediated eye movements during spoken-word recognition", Psychon. B. Rev., 12:453-459, 2005.

[12] Higashikawa, M. and Minifie, F. D., "Acoustic–perceptual correlates of 'whisper pitch' in synthetically generated vowels," J. Speech, Lang. Hear. Res., 42:583-591, 1999.

[13] Denes, P., "A preliminary investigation of certain aspects of intonation," Lang. Speech, 2:106-122, 1959.

[14] Liu, S. and Samuel, A. G., "Perception of Mandarin lexical tones when F0 is neutralized," Lang. Speech, 47:109-138, 2004.

[15] Meyer–Eppler, W., "Realization of prosodic features in whispered speech," J. Acoust. Soc. Am., 19:104-106, 1957.

[16] Bibyk, S., Heeren, W., Gunlogson, C. and Tanenhaus, M. K., "Asking or Telling - Real-time processing of boundary tones", submitted.

[17] Kurumada, C., Brown, M. and Tanenhaus, M. K., "Pragmatic interpretation of contrastive prosody; It looks like speech adaptation", in N. Miyake, D. Peebles and R. P. Cooper [Eds.], Proc. 34th Annual Conference of the Cognitive Science Society, 647-652, 2012.

"A little more ironic" – Voice quality and segmental reduction differences between sarcastic and neutral utterances

Oliver Niebuhr

Department of General Linguistics, ISFAS, Kiel University, Germany
niebuhr@isfas.uni-kiel.de

Abstract

The presented production experiment analyzes the phonetic differences between neutral (i.e. sincere) and sarcastically ironic utterances in German. Results show in line with previous studies that sarcastic irony is expressed by longer utterance durations, lower and flatter F0 contours, and a lower intensity level. Moreover, extending previous findings, sarcastic irony is also characterized by a more variable (in tendency breathier) voice quality and a higher degree of segmental reduction, probably reflecting the speakers' dissociation from the wording of their utterances.

Index Terms: irony, sarcasm, intonation, emphasis, reduction

1. Introduction

Ironic statements convey the opposite of what the speaker has put into words. Even in this admittedly oversimplified common sense definition (e.g., [1,2]), irony is anything else but a marginal aspect of speech communication. For example, estimated on the basis of spontaneous dialogue corpora [8,23], one hour of speech contains about 10-15 instances of ironic expressions. They represent a type of emphasis at the utterance level. Such utterance-level types of emphasis can be distinguished from other word-level types of emphasis [3,4,5]. While lexical means like extreme exaggerations or question tags are optional irony signals, prosodic signals seem to be a constant companion of ironic utterances. Typically, and particularly in lexically unmarked ironic utterances, prosodic meanings are selected such that they clash with the verbal meanings. In this way, they can indicate irony and express at the same time what the speaker actually wants to say (in combination with the communicative and visual contexts).

Since there is an infinite number of different ironic utterances, and since the prosodic meanings have to be adjusted to clash with each of their verbal meanings, it is obvious that *the* prosody of irony cannot be determined. Therefore, phonetic studies trying to pinpoint specific prosodic exponents of irony have either failed or – implicitly or explicitly – investigated a particular subtype of irony. The most frequently investigated subtype of irony concerns the clash between positive verbal and negative prosodic meanings. It is often referred to as sarcasm, although sarcasm is actually just an attitudinal concept of a provocatively negative kind, which can in principle also be directly expressed in speech without being exploited for an ironic meaning clash [1].

The general aim of the present paper is to continue the line of phonetic research on this sarcastic subtype of irony. However, for the sake of simplicity and because irony is the determining concept, the more accurate term 'sarcastic irony' will be used interchangeably with 'irony' in the following. Detailed experimental-phonetic analyses of sarcastic irony have begun only about a decade ago, which makes sarcastic irony a relatively young field of research. The phonetic picture that has emerged so far from analyses of English, Italian, French, German, and Cantonese can be condensed as follows: Compared with neutral utterances, utterances with sarcastic irony are produced longer (i.e. at lower speaking rates) and with clear changes in the levels and ranges of F0 and intensity [6,7,8]. F0 patterns seem to be lowered and narrowed in Western Germanic languages like English and German [9, 10,11,12], but raised and/or widened in Cantonese and Roman languages like Italian or French [13,14,15]. Findings on intensity changes were a bit less consistent, but typically went in the direction of a lower, flatter intensity contour [13,16] (cf. [9,12] for exceptions).

The summarized findings call for an extension in two directions. (1) What about voice quality whose relevance as "the fourth prosodic dimension" as recently been stressed and promoted by [17]? (2) Since modern phonetic research also stresses the strong linkages of segments and prosodies (e.g. [18]), what about the segmental characteristics of sarcastic irony? The present study addresses these two questions on the basis of Standard Northern German.

As regards question (1), there are impressionistic indications from [19] that sarcastic irony is produced with a softer, i.e. breathier voice quality. With respect to question (2), one may expect a higher degree of speech reduction in ironic utterances for the following reason. It has recently been argued in [20,21] that the degree of speech reduction in an utterance is not just the result of an articulatory balance between the competing demands of economy and comprehensibility in a given communicative situation, cf. [27]. Rather, the articulatory effort invested by the speaker also signals his/her attitudes towards the dialogue partner or the content of the message. For example, utterances like "good morning", "I do not know" (cf. also [22]), "I am sorry", or "thank you" are less reduced when they not just function as routine statements, but sincerely serve to initiate a conversation, or express compassion, cooperativeness, remorse, or gratitude, cf. also [29]. Since irony requires that speakers distance themselves from the semantic con-tent of their utterances, it is reasonable to assume that they mark this distance by a higher degree of segmental reduction.

In summary, the hypotheses tested in this study are that ironic utterances in Standard Northern German are longer, breathier, more reduced, and have lower and flatter F0 and intensity contours than their neutral counterparts. Design and procedure of the corresponding production experiment will be outlined in the following section.

2. Method

2.1. Target Sentences

The study was based on 20 target sentences. They were designed to meet three criteria. First, they were plain and syntactically simple constructions of four to eight frequent German words. This allowed the sentences to be produced fluently and with a clear and consistent prosody. Second, in order to facilitate speech reduction processes, each sentence contained at least two so-called phonetically weak forms in the

sense of [24], i.e. function words like pronouns, auxiliary verbs, modal particles, articles, prepositions, or conjunctions, which are by default unstressed. Third, the sentences were formulated such that they can be directed towards an interlocutor. In addition, they made positive statements about this interlocutor or a third party and in this way lay the ground for realizing the sentences not only in a neutral fashion, but also with sarcastic irony by inverting their positive propositions. The 20 target sentences are provided in the Appendix.

2.2. Speakers and Recording Procedure

The target sentences were produced by 10 speakers of Standard Northern German, 5 males and 5 females. They were students at Kiel University and 23-29 years old.

The recordings were conducted at the speakers' homes in rooms which were silent and as anechoic as possible (i.e. typically in the speakers' living rooms). It was assumed (and supported by feedback after the recordings) that familiar surroundings would make it easier for the speakers to get into the kind of mood that is needed to express sarcastic irony. For the same reason, the recordings were conducted by a student of phonetics – DT – who was a good friend of all of the speakers.

Each speaker had a few minutes time prior to the recording to familiarize him/herself with the target sentences. Then, the speaker received the oral instruction to produce each target sentence in an informal everyday fashion while addressing his/her friend DT. The target sentences were presented separately on file cards in order to avoid any list effects on speech production. The speakers were free to repeat each sentence until they were satisfied with the result; and in fact most speakers made use of this possibility several times due to unsatisfactory expressions of irony or insufficient informality.

The group of 10 speakers was divided into two subgroups. One subgroup produced the target sentences first in a neutral fashion and then with sarcastic irony. The other, equally large subgroup started with the expression of sarcastic irony and then uttered the target sentences again in a neutral way. The sentence order was re-randomized after each round and for each new speaker. The concept of sarcastic irony was explained to the speakers on request by means of instances in comic strips, i.e. without using auditory examples.

The elicited ironic and neutral target sentences were recorded digitally with a HiFi speech recorder (Zoom H2n). Instead of using the built-in microphone, recordings were made with a head-mounted microphone so that the relative distance between mouth and microphone was constant during the recording. An entire recording session took about 15 minutes.

Pitch accents in the neutral and ironic sentences fell on the same words and were free from additional positive or negative emphatic intensification [3]. Spot checks just revealed a slightly greater number of high-tone (H* or L+H*) accents in the neutral set, reflecting the positive sentence semantics.

2.3. Filtering out Unclear Cases

Phonetic analyses were preceded by a small perception experiment. It served to cross-check whether the target sentences selected by the speakers to convey sincerity (neutrality) and sarcastic irony would in fact be clearly associated with these functions by independent listeners.

To that end, all target sentences were cut out of the recorded sound files and normalized in amplitude with Adobe Audition (www.adobe.com/Audition) to 90% of the possible dynamic range. Then, they were arranged in four differently randomized orders in which the sentences were separated from each other by a pause of five seconds and a beep. The so-assembled stimulus lists were finally handed over as a single sound file to four naïve German listeners together with a corresponding response sheet.

The naïve participants were informed that they would hear a set of isolated sentences from 10 speakers of Standard Northern German. They were furthermore asked to play the sound files in a silent room and judge spontaneously for each sentence whether it was meant sincerely or ironically. Conducting the perception experiment took about 45 minutes.

Based on the retrieved response sheets, target sentences were excluded from further analyses if they were not correctly identified as neutral or ironic in at least 75% of the cases. Although sentences of both types were affected by this filtering, it was primarily the ironic sentences that failed to meet the 75% threshold. Therefore, in order to create equally large samples for the phonetic analyses, additional sentences were randomly excluded from the neutral sample. In summary, the perceptual filtering procedure resulted in 142 target sentences, subdivided into two samples of 71 clearly neutral and 71 clearly ironic tokens.

2.4. Phonetic Analyses

The phonetic analyses of the remaining, clear target sentences included acoustic as well as auditory measurements. The acoustic measurements covered all four prosodic dimensions, i.e. F0, intensity, duration, and voice quality. F0 and intensity were represented by four parameters each. In summary, the following ten parameter values (a)-(j) were determined for each target sentence:

- (a) F0 minimum of the sentence, measured in semitones relative to 100/200 Hz for male/female speakers,
- (b) F0 maximum of the sentence, measured in semitones relative to 100/200 Hz for male/female speakers,
- (c) F0 range in semitones, i.e. (b) subtracted by (a),
- (d) average F0 level, calculated in semitones relative to 100/200 Hz for male/female speakers,
- (e) intensity minimum in dB within a non-silent sound section of the target sentence,
- (f) intensity maximum in dB,
- (g) intensity range in dB, i.e. (f) subtracted by (e),
- (h) average intensity level in dB, disregarding silent (sound) sections of the target sentence,
- (i) voice quality in terms of the amplitude difference in dB between the first and the second harmonic (H1-H2), measured in the midpoints of all vowels of a target sentence and then averaged across these vowels,
- (j) tempo, represented by the total sentence duration.

All measurements except for (i) were conducted with PRAAT [28], using the default settings. The F0 measurements (a)-(b) and (d) omitted octave errors and, as far as possible, also microprosodic perturbations. Intensity measurements were not normalized as the use of head-mounted microphones assured a constant speaker-microphone distance. The voice-quality measurements for (i) were made with WaveSurfer (www.speech.kth.se) on the basis of narrow-band FFT spectra.

The acoustic-prosodic analysis was complemented by a careful auditory analysis targeted at the degree of segmental reduction in a target sentence. This degree was measured by successively playing short sections of the target sentence and counting the number of assimilations, elisions, and lenitions of sound segments in these sections relative to their canonical full form in Standard German according to [25]. The reduction countings of the individual sections were then summed up for each target sentence. Assimilations could, for example, concern voice and place of articulation. Lenitions included vocalizations of consonants, centralizations of vowels, and changes of obstruents to approximants. The auditory analysis was performed with headphones by DT in a silent room and subsequently cross-checked by a trained phonetician (ON). DT and ON were Standard German listeners. Cases of disagreement were supplemented by visual inspections of spectrograms and/ or waveforms and discussed until a final decision was reached.

3. Results

The acoustic-prosodic measurements made within the selected 2x71 target sentences were analyzed with a multivariate ANOVA. It was based on the 2-level fixed factor Sentence Type (ironic vs. neutral); and since the sentences of the ironic and neutral sample were not equally distributed across the 10 speakers, possible differences between the two samples were analyzed in terms between-subject effects. The MANOVA resulted in a highly significant main effect of Sentence Type ($F[10,131]= 69.696$; $p<0.001$; $\eta_p^2= 0.842$). All dimensions of acoustic-prosodic parameters contributed to this main effect.

As regards the F0 parameters (a)-(d), the largest difference between the ironic and neutral sentences and thus the strongest effect in terms of partial eta-squared lies in the height of the F0 maximum ($F[1,140]= 193.416$; $p<0.001$; $\eta_p^2= 0.580$). As is illustrated in Figure 1, this maximum was produced considerably (i.e. 3 semitones) lower for the ironic than for the neutral sentences. The same, though less clearly, applied to the F0 minimum ($F[1,140]= 12.816$; $p<0.001$; $\eta_p^2= 0.084$), and in accord with the lowering of these two extreme F0 values, the overall F0 level was also about a major third lower for the ironic than for the neutral sentences ($F[1,140]= 116.366$; $p<0.001$; $\eta_p^2= 0.454$). In addition, the sentences expressing irony were also characterized by a significantly narrowed F0 range, representing the second strongest F0 effect ($F[1,140]= 163.532$; $p<0.001$; $\eta_p^2= 0.539$).

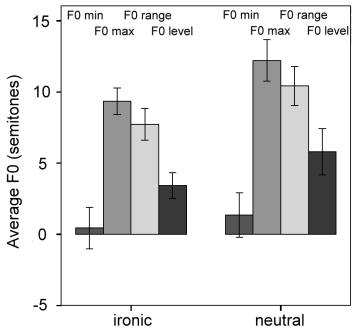


Figure 1: *Means and standard deviations of the four F0 parameters (a)-(d), measured in ironic (left) and neutral (right) target utterances; n=71 for each bar.*

The biggest intensity contributions to the significant main effect of Sentence Type come from the intensity minima and maxima. Figure 2 shows that both clearly decrease (between 4-9 dB) from the neutral to the ironic sentences (Int-min: $F[1,140]= 72.334$; $p<0.001$; $\eta_p^2= 0.341$; Int-max: $F[1,140]= 71.643$; $p<0.001$; $\eta_p^2= 0.339$). The intensity range does not differ significantly between the ironic and neutral sentences. However, parallel to the change in F0 level, the intensity level was also significantly lower for the ironic than for the neutral sentences ($F[1,140]= 37.594$; $p<0.001$; $\eta_p^2= 0.212$).

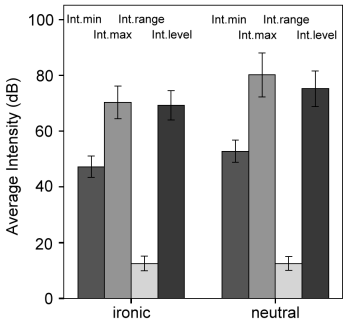


Figure 2: *Means and standard deviations of the four intensity parameters (e)-(h), measured in ironic (left) and neutral (right) utterances; n=71 for each bar.*

From the remaining two prosodic parameters (i)-(j), i.e. harmonic amplitude difference H1-H2 (voice quality) and sentence duration, only the latter differed significantly between the ironic and neutral sentences ($F[1,140]= 58.589$; $p<0.001$; $\eta_p^2= 0.288$), cf. Figure 3. Sentence durations were on average about 25% longer in the ironic than in the neutral productions. Importantly, this effect is not due to the fact that the 20 sentences occurred with different frequencies in the ironic and neutral samples. Even when the sentence durations were normalized and recalculated in terms syllables per second, a separate t test (for unpaired samples and homogeneous variances, based on a prior F test) showed that the duration difference between the ironic and neutral sample persisted, also with regard to its magnitude ($t[140]= -10.683$; $p<0.001$). The average syllable per second rate in the ironic sentence set was 5.5. This is about 30% slower than the average rate of 7.1 syllables per second in the neutral set.

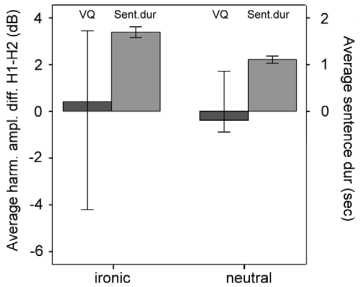


Figure 3: *Means and standard deviations of voice quality (dark gray, i) and sentence duration (light gray, j), measured in ironic (left) and neutral (right) target utterances; n=71 for each bar.*

The voice quality measure only yielded a trend towards higher H1-H2 values in ironic sentences $F[1,140]= 2.796$; $p<0.097$; $\eta_p^2= 0.020$). Next to this trend, however, there is a much more salient and statistically significant difference between the voice-quality measurements of the ironic and neutral samples. As can be seen in Figure 3, this difference concerns the standard deviation of the H1-H2 values, which were much larger

in the ironic than in the neutral sample. It is not surprising in view of this observation that the H1-H2 parameter clearly stood out highly significantly in the Levene tests of the MANOVA ($F[1,140]= 70.073$; $p<0.001$). So, while the overall voice quality in the neutral sentences was mainly constant and modal, the ironic sentences were produced with either breathier or tenser voice qualities. This interpretation of the measurements is in accord with an auditory inspection of the data.

The results of the auditory reduction analysis are summarized in Figure 4. It shows firstly that both the neutral and the ironic target sentences were subject to reduction processes such as assimilation, elision, and lenition. Secondly, however, the degree of reduction in terms of the mean frequency of reduction processes looks overall higher in the ironic than in the neutral sentences. This impression was confirmed in a one-way ANOVA. That is, the ironic set showed on average 4.8 reductions per sentence, which is significantly more than the 3.6 reductions that occurred on average in the neutral set of target sentences ($F[1,38]=10.464$; $p=0.003$; $\eta_p^2=0.216$). However, it must also be noted that this overall difference was not equally present for each target sentence. While sentences like 8, 15, 16, and 19 were realized with about twice as many reductions in the ironic condition, other sentence like 1, 3, 7, 12, and 17 were produced with a similar or even slightly higher number of reductions in the neutral condition.

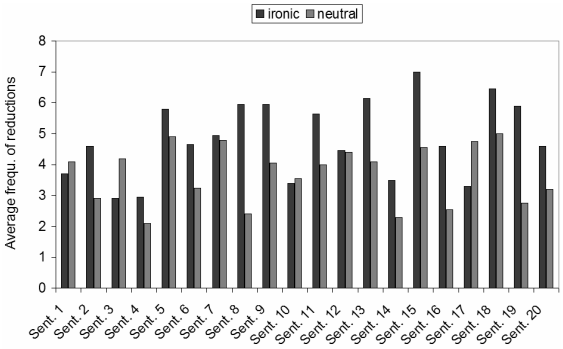


Figure 4: *Mean frequency of reductions found across the 10 speakers in the ironic (dark gray) and neutral (light gray) productions of the 20 target sentences.*

4. Discussion and Conclusions

The present findings support the conclusion of previous studies that sarcastic irony has a separate phonetic profile in comparison to neutral (i.e. sincere) speech. This profile involves extended utterance durations due to a reduced speaking rate, as well as changes in the F0 and intensity patterns. In accord with previous findings for Western Germanic languages, sarcastic irony was characterized in the present study by a lower and flatter F0. Intensity findings for sarcastic irony have always been less consistent both within and across languages. For example, expressing sarcastic irony in varieties of English was associated with an *in*crease in intensity parameters in [9], but with a *de*crease in intensity parameters in [13]. Similarly, for Standard German the intensity level was raised in [12], but lowered in the present study of the same language variety. Moreover, in contrast to [13,16] the present findings do not include a narrowed intensity range, as the intensity minima and maxima were both lowered to the same degree. The lack of a narrowed intensity range is also the only deviation of the results from the hypotheses put forward in the introduction. All other hypotheses were confirmed.

One possible explanation for the inconsistencies in intensity is that the direction of intensity changes – unlike those of F0 and speaking rate – does not reflect sarcastic irony itself, but rather the arousal level on which the expression of sarcastic irony is based. That is, similar to cold and hot anger [26], there may be different degrees or subtypes of sarcastic irony (in addition to higher-level types of irony like kind and sarcastic irony, cf. [15]). It fits in with this explanation that [16] introduced on an impressionistic basis the distinction between dry and dripping sarcasm and found that their intensity patterns differed relative to neutral utterances. This attempt of [16] stresses the need for a separate line of research on the sub-structures of the irony concept, which can then complement and guide our growing phonetic understanding of irony.

As for this growing phonetic understanding, the present study showed for the first time that changes away from modal voice quality and a higher degree of segmental reduction can also be a part of sarcastic irony. It is argued here with reference to [20,21,22] that the higher reduction level is supposed to signal that the speaker distances him/herself from the wordings of the sentences; in the same way as "I'm sorry" or "good morning" are more strongly reduced when they are used (insincerely) for routine matters. Looking at Figure 4, it seems that not all ironic target sentences differed in terms of reduction from their neutral counterparts. Target sentences like 1, 7, and 12 showed about the same absolute number of reductions in the two Sentence Type conditions. However, even for these sentences it can be argued that the *relative* degree of reduction was still higher in the ironic condition, since sentences in the ironic condition were produced (consistently) longer, i.e. at a slower speaking rate. So, speakers had basically more time in the ironic sentences to reach the respective articulatory targets, but obviously made no use of this possibility. This fact, in combination with the significant effect of reduction on Sentence Type, challenge an important assumption in the modelling of speech production. It has often been claimed with reference to the H&H theory [27] that slower speech is automatically also less reduced. The present study shows there is no such automatism. The degree of reduction is not just the passive result of time, speaker economy and listener demands. Rather, reduction is also actively varied to convey different kinds of function in speech communication.

The changes in voice quality went in both directions tense voice and breathy voice. It seems that the direction of the change was determined by the sentence semantics. For example, target sentences like 1, 10, and 15 were more often realized with a tense voice, whereas the majority of sentences like 2, 4, 8, 14, and 20 typically showed a breathier voice. Thus, like for intensity, and maybe even shaping the latter, the heterogeneous voice quality findings may reflect different degrees or subtypes of (sarcastic) irony. For example, it is striking that statements in which the irony is used to point to future alternatives were tenser, whereas ironic statements pointing to the lack of alternatives were mostly breathier.

In any case, it can be concluded from the present study that sarcastic irony is a four-dimensional rather than just a three-dimensional prosodic phenomenon, which moreover extends beyond the prosodic layer into the traditionally separated segmental layer of the speech signal. Both of these new insights must be further investigated and put into a cross-linguistic perspective. It seems also worth looking for speaker-specific differences, even though the only differences found here were those between sentences, probably caused by uncontrolled degrees or subtypes of (sarcastic) irony.

5. References

[1] Partridge, E., "Usage and Abusage: A Guide to Good English", London: Penguin Press, 1969.

[2] Muecke, D.C.,"The Compass of Irony", London: Methuen,1969.

[3] Niebuhr, O., "On the phonetics of intensifying emphasis in German", Phonetica 67: 170-198, 2010

[4] Niebuhr, O. Jarzabkowska, P., Lorenz, U., Schulz, C., Sodigov, F., "Say it again, Sam! Phonetic Forms and Functions of Emphatic Reduplication in German", Proc. 6th Speech Prosody, Shanghai, China, 258-261, 2012.

[5] Carton, F., Hirst, D., Marchal, A., Seguinoit, A., "L'accent d'insistance – Emphatic stress (Studia Phonetica 12)", Montréal: Didier.

[6] Cutler, A., "On saying what you mean without meaning what you say. Proc. 10th Regional Meeting of the Chicago Linguistic Society, Chicago, USA, 117-127.

[7] Haiman, J., "Talk is cheap: Sarcasm, alienation, and the evolution of language". USA: Oxford University Press, 1998.

[8] Bryant, G.A., "Prosodic contrasts in ironic speech", Discourse Processes 47 : 545-566, 2010.

[9] Rockwell, P., "Lower, slower, louder: vocal cues of sarcasm", Journal of Psycholinguistic Research 29: 483-495, 2000.

[10] Attardo, S. Eisterhold, J., Hay, J., Poggi, I., "Multimodal markers of irony and sarcasm". International Journal of Humor Research 16: 243-260, 2003.

[11] Cheang, H.S., Pell, M. D., "The sound of sarcasm", Speech Communication 50: 366-381, 2008.

[12] Sharrer, L., Christman, U., "Voice Modulations in German Ironic Speech", Language & Speech 54: 435-465, 2011.

[13] Cheang, H.S., Pell, M.D., "Acoustic markers of sarcasm in Cantonese and English", Journal of the Acoustical Society of America 126: 1394-1405, 2009.

[14] Lœvenbruck, H., Ben Jannet M., D'Imperio, M., Spini, M., Champagne-Lavau, M., "Prosodic cues of sarcastic speech in French: slower, higher, wider", Proc. 14th Interspeech 2013, Lyon, France, 1470-1474, 2013.

[15] Anolli, L., Ciceri, R., Infantino, M.G., "Irony as a game of implicitness: Acoustic profiles of ironic communication", Journal of Psycholinguistic Research 29: 275-311, 2000.

[16] Bryant, G.A, Fox Tree, J., "Is there an Ironic Tone of Voice?", Language and speech 48: 257-277, 2005.

[17] Campbell, N., Mokthari, P., "Voice quality: the 4th prosodic dimension", Proc. of the 15th ICPhS, Barcelona, Spain, 2417-2420.

[18] Kohler, K.J., "Communicative functions integrate segments in prosodies and prosodies in segments", Phonetica 68: 26-56, 2011.

[19] Muecke, D.C., "Irony markers", Poetics 7: 363-375, 1978.

[20] Niebuhr, O., Kohler, K.J., "Perception of phonetic detail in the identification of highly reduced words", Journal of Phonetics 39: 319-329, 2011.

[21] Graupe, E., Görs, K., Niebuhr, O., "Reduktion gesprochener Sprache - Bereicherung der Berhinderung der Kommunikation?" in O. Niebuhr [Ed], Formen des Nicht-Verstehens, Frankfurt: Peter Lang, 2014.

[22] Hawkins, S., "Roles and representations of systematic fine phonetic detail in speech understanding", Journal of Phonetics 31: 373-405, 2003.

[23] Peters, B., "The database – The Kiel Corpus of Spontaneous Speech", in K.J. Kohler, F. Kleber, B. Peters [Eds], Prosodic Structures in German Spontaneous Speech (AIPUK 35a), Kiel: IPDS, 1-6, 2005.

[24] Kohler, K.J., "Segmental reduction in connected speech in German: Phonological facts and phonetic explanations", in W.J. Hardcastle, A. Marchal [Eds], Speech Production and Speech Modelling, Dordrecht: Kluwer, 69-92, 1990.

[25] Mangold, M., „Duden – Das Aussprachewörterbuch", Mannheim: Bibliographisches Institut, 1998.

[26] Banse, R., Scherer, K.R., "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology 70: 614-636, 1996.

[27] Lindblom, B., "Explaining phonetic variation: a sketch of the H&H theory", in W.J. Hardcastle, A. Marchal [Eds], Speech Production and Speech Modelling, Dordrecht: Kluwer, 403-439, 1990.

[28] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International 5, 341-345, 2001.

[29] Local, J., "Variable domains and variable relevance: interpreting phonetic exponents", Journal of Phonetics 31: 321-339, 2003.

6. Appendix

1. Das ist aber mal interessant
   (That's so interesting)
2. Das Wetter gefällt mir heute mal richtig gut
   (I really like the weather today)
3. Das wird bestimmt spannend
   (This will surely be exciting)
4. Das läuft ja super
   (This works out perferctly)
5. Das riecht richtig lecker
   (It smells really delicious)
6. Er sieht aus wie ein richtiger Herr
   (He looks like a real gentleman)
7. Das nenne ich mal eine richtige Freundschaft
   (That's what I call a real friendship)
8. Er ist besonders wichtig
   (He is a very important person)
9. Der schmeckt mir der Senf
   (I like how the mustard tastes)
10. Mach mal ruhig weiter so
    (Just go on like that)
11. Klar, lad ein, wen du willst
    (Sure, invite who you like)
12. Das ist ja lustig hier heute
    (It is indeed funny here today)
13. Danke für den Aufwand
    (Thanks for all your efforts)
14. Das ist ja fantastisch
    (This is indeed fantastic)
15. Komm ruhig später
    (Just come later)
16. Sie kann so gut kochen
    (She can cook so well)
17. Was für eine schöne Jacke
    (What a wonderful jacket)
18. Da bin ich richtig gut drin
    (I'm really good in that)
19. Na klar kannst du das haben
    (Of course, you can have that)
20.) Das ist ein schönes Leben
    (This is a beautiful life)

Is breathing sensitive to the communication partner?

Amélie Rochet-Capellan[1], Gérard Bailly[1] and Susanne Fuchs[2]

[1] GIPSA-lab, UMR 5216 CNRS/Université de Grenoble –France

[2] Centre for General Linguistics, Berlin – Germany

{amelie.rochet-capellan,gerard.bailly}@gipsa-lab.grenoble-inp.fr, fuchs@zas.gwz-berlin.de

Abstract

This paper investigates breathing profiles in eleven female speakers (subjects) when talking successively with the same two females (partners). Breathing kinematics of the two interlocutors was recorded synchronously by means of two Inductance Plethysmographs. In order to understand the implication of breathing in dialogue, we analyzed changes in breathing pauses according to the main dialogue events (listening, backchannels, turns start and turns continuation). Breathing and syllable rates were also compared among partners and subjects. The duration of inhalations and related pauses was reduced before a turn continuation in comparison to a turn start. The delay between speech offset in a breathing cycle and the onset of the next inhalation increased when a speaker and a listener swap roles as compared to a speaker who continued the turn. This was observed for both partners and subjects. The partners differed in their breathing and articulation rates but the two rates were not clearly correlated. In agreement with previous works, the current study shows that breathing kinematics is strongly linked to dialogue events. However, it doesn't show any clear effect of partner on speaker's breathing. This last result is discussed relative to methodological aspects.

Index Terms: Breathing, Respiration, Spontaneous dialogue, Interpersonal adaptation, Breathing rate, Syllable rate, Inhalation pauses

1 Introduction

Breathing is actively involved in speech production as it provides the airflow required to generate speech sounds. The vital need of air is also a constraint that organizes the discourse into inhalation pauses and speech intervals. Several studies have analyzed breathing in reading tasks and to a lesser extent in spontaneous speech. These studies consistently showed that speech production is achieved by a specific control of breathing, visible in the clear reduction of the inhalation duration relative to the exhalation phase, and as compared with quiet breathing [1-3].

In text reading, the inhalation pauses consistently occur at syntactic boundaries [4]. The duration and amplitude of inhalation are mainly related to the syntactic constituents of the text (e.g. more air is inhaled before a new paragraph than before a sentence inside a paragraph) [5-7]. The properties of inhalation are also related to the length of the upcoming utterance [2, 8-10]. Similar behavior can be found in spontaneous speech, with greater inhalation before a main than an embedded clause, but more inhalation pauses occurring at non-syntactic locations as compared to read speech [11-13].

Analyses of breathing noises showed that they may play a role to indicate continuity between two related speech groups in text reading [14], and could improve the quality of synthetic speech [15]. This indicates that breathing noises are useful for speech perception. Furthermore, when listening to read speech, the listener breathing changes according to the properties of the speech signal, suggesting some influences of the speaker on the listener breathing and interaction between the control of breathing and perceptual processes [16-18].

Few studies have analyzed interpersonal influences of breathing in verbal collaborative tasks. During collaborative reading, readers coordinate their breathing. They breathe in-phase when reading synchronously, and in anti-phase when reading in alternation [19]. During choir singing, singers synchronize their breathing, especially when singing in unison [20].

In dialogue, breathing pauses have to be coordinated with interpersonal constraints and in particular with turn taking events. Despite hypotheses about the implication of breathing in conversation [21], few behavioral studies have analyzed the breathing profiles in spontaneous dialogue. Preliminary analyses of inhalation pauses in spontaneous dialogue were based on short recordings and single dyads [22-23]. They showed adaptation of breathing to dialogue constraints. A more systematic investigation of breathing in scripted and spontaneous dialogue was provided in [24]. This study found evidence for interpersonal alignment of breathing related to turn taking and laughter. Yet, no study addressed changes in breathing profiles during dialogue according to the conversational partner, nor the potential role of breathing in communication.

Several studies found interpersonal adaptations during verbal interactions. These adaptations may occur at different levels. For example, interlocutors involved in a dialogue may converge in speech rate or intensity, f0 or formants values [25-27]. The same speaker may thus behave differently when talking to two different partners. As breathing is specifically involved in speech production and verbal interaction, it could also be involved in interpersonal adaptation. We tested potential changes in speakers' breathing according to their interlocutor by analyzing breathing profiles during spontaneous dialogue. In addition, we tested how breathing pauses durations were linked to the main communicative events of the dialogues (listening, backchannels, turns continuation vs. turns start).

2 Methods

2.1. Subjects

The participants were eleven subjects (S01-S11, age: 31 years (mean) ± 7 (standard deviation), body mass index 21.3 ±1.5) and two partners (P01, age 42, BMI: 20.4, and P02 age 28, BMI: 20.8). Subjects and partners were all native female speakers of German, students or academics with a university degree. The subjects were naïve to the purpose of the study while the partners were not.

2.2. Procedure and data acquisition

Partner and subject were sitting and facing each other with a distance of ~1.5 m. They were instructed to keep their hands on their legs in order to avoid torso and arm movements that could strongly affect the recording of breathing. Subject and

partner's task was to talk with each other about a topic chosen in agreement with one another (holidays, sports, cooking, or movies). In total, each subject had five short conversations with each partner (2.5 min, for each trial), starting with P01 or P02. The topic of the conversation could change or remain the same over the five trials.

The acoustic signals were recorded using two directional microphones (Sennheiser HKH50 P48). The rib cage and the abdominal kinematics were recorded by means of two Inductance Plethysmographs (RespitraceTM). One band was positioned at the level of the axilla (rib cage) and the other band at the level of the umbilicus (abdomen). The acoustic and the breathing signals were recorded synchronously for the two speakers by means of a six channels voltage data acquisition system. All signals were sampled at 11030 Hz.

2.3. Post-processing and labeling

After the recording, the breathing data were sub-sampled at 100 Hz and pass-band filtered (1-40Hz). The onset and offset of inhalation movements were detected automatically from the sum of the rib cage and the abdomen displacements, and manually corrected when required. The breathing cycle was defined from the onset of an inhalation to the onset of the next inhalation (*I+PI* on Figure 1).

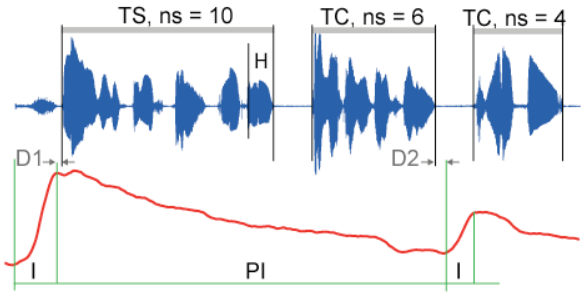


Figure 1. Top: The acoustic signal and the corresponding labeling of speech groups (*TS*: turn start, *TC*: turn continuation, *ns*: number of syllables, *H*: hesitation). Bottom: corresponding breathing kinematics and labeling of the breathing cycle (*I*: inhalation + *PI*: post inhalation). D1: *delayOnset* and D2: *delayOffset* are delays between the main breathing and speech events (see text for details).

Speech productions were labeled in Praat [28] by a trained phonetician. The boundaries of the inter-pausal units (IPUs) were detected. The vocalized hesitations (*uh, uhm mmm*,...) and the non-verbal communicative noises (laughter, mouths noises) were also delimited. Hesitations were distinguished from backchannels (*mhm*) easily as the first ones occurred during speaking and the second ones during listening phases. The spoken productions were transcribed for each IPU. On the basis of this transcription, the number of syllables was derived automatically from the output of the BALLOON toolkit [29]. The vocalized hesitations were considered as one (*uh, uhm*) or two syllables (*mmm*). IPUs were also labeled according to their main function in the dialogue as: turn start, turn continuation or backchannel (Figure 1). A turn start (*TS*) was defined when one interlocutor became the speaker and her interlocutor became the listener. This initial turn could be followed by one or several continuation (*TC*) separated by silent pauses. When the turn holder held the floor, the listener could produce short utterances like "mhm", "okay", "yes", "aha" that did not intend to take the floor, but rather signal to the speaker to continue talking. These utterances were labeled as backchannels (*BC*) [30].

2.4. Data selection and analyses

Each breathing cycle (*I+PI* on Figure 1) was characterized by its total duration (*durC*) and the duration of inhalation (*durI*). For each trial, we computed the breathing rate (as the number of labeled breathing cycles divided by the sum of the durations of these breathing cycles) and the articulation rate (total number of syllables in the speech groups divided by the sum of their durations). Vocalized hesitations were not considered in the computation of articulation rate.

The breathing cycles were classified according to their function in the conversation (*cycle_type*): listening cycle (*LI*, the subject is not speaking while her interlocutor is); backchannel cycle (the cycle included only one or more *BC* units); turn start (the cycle started with a *TS* IPU), and turn continuation (the cycle started with *TC* IPU).

The inhalation noises could be used to signal the continuation of a theme or major thematic breaks during text reading [14] : their amplitudes, durations and phasing relations with speech contribute to encode thematic structure. We investigated if similar strategies could also be found in unconstrained face-to face dialogues to signal turn start or continuation by analyzing pauses related to inhalation. For *BC, TS* and *TC* cycles, we computed the delay between inhalation offset and the onset of the first IPU (*delayOnset*) and between the end of the last speech unit and the onset of the next inhalation (*delayOffset*).

The changes in breathing and syllable rate according to the partner were tested using ANOVAs. The dataset was separated for subjects and partners, with partner as a within subject factor. This allows testing the effect of the partner on subjects' behaviors, and differences between the two partners. The results were considered as significant when $p<.01$.

The effect of partner and task (*TS*, *TC*, *BC*, *LI*) on the duration of inhalation was tested using Linear Mixed Models (lme4 & languageR packages in R version 2.14). The dataset was split between the partners and subjects' data. The subjects and dialogue trials were taken as random factors. The inhalation duration was log-transformed to obtain normally distributed residuals. The results were considered significant with $p<.01$. Additionally we run lmer for the *delayOnset* and *delayOffset*, but the residuals were nonlinear, even when transformed. We will therefore only provide descriptive statistics.

3 Results

3.1. Global description of the dialogue

We first characterized the dialogue at a global level. On average, speech intervals represented 37% of the total duration for P01, 40% for P2, 39% for subjects talking with P01 and 43% for subjects talking with P02. The effect of partner on the duration of subjects' speech failed to reach significance ($F(1, 10)=3.9$, $p=.08$), due to variability in subjects' behaviors. For example, some subjects were talking more than the partners while the reverse was observed for other subjects.

3.2. Relationships between breathing and syllable rates

The breathing and syllable rates were globally smaller for P01 as compared with P02 ($F(1, 10)=18.4$, $p<.01$ and $F(1, 10)=67$, $p<.0001$), see Figure 2. By contrast, subjects' breathing and syllable rates were not different when either talking with P01 or P02. Positive correlations were observed between the average syllable and breathing rates for subjects ($r=.61$ when talking with P01, .43 when talking with P02) and for partners ($r\sim.3$

for P01 and P02). For subjects, the correlation was significant only when talking with P01 ($p<.05$).

No clear correlation was observed between subjects and partners' breathing and syllable rates (Figure 3).

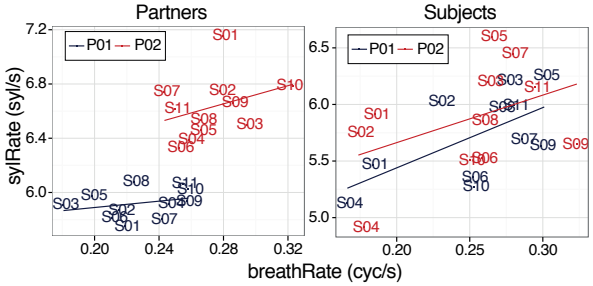


Figure 2. Average syllable rate according to average breathing rate for partners (left) and subjects (right).

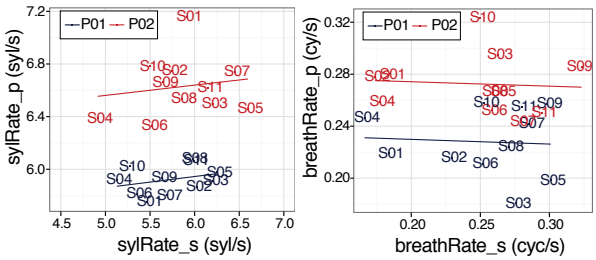


Figure 3. Partner's vs. subject's parameters, left: average syllable rate, right: average breathing rate

3.3. Duration of inhalation

Globally, inhalations were shorter for P02 than P01 (here we used lmer since our dataset was unbalanced, t=-7.3, pMCMC=.0001), while subjects did not show significant changes in *durI* according to the partner (t=-2.3). Inhalations were significantly shorter before turn continuations than before turn starts, backchannels and listening cycles (all $|t|>2.7$, pMCMC<.004). This was observed for both partners and subjects (Figure 4). The difference between *BC* and *LI* was significant for partners (t=2.7, pMCMC=.004) but not subjects (t=2.3).

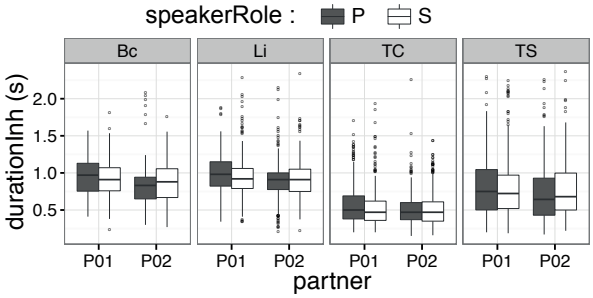


Figure 4. Duration of inhalation (*durI*) according to the function of the first speech group on the breathing cycle (*BC*, *TC*, *TS*) and for listening cycle (*LI*), for dyad involving P01 and P02 and for partners (*P*) and subjects (*S*).

3.4. Delay between inhalation offset and speech onset

The duration of pauses between inhalation offset and speech onset (*delayOnset*) were shorter for turn continuations than turn starts and backchannel cycles (see Figure). P02 may also initiate backchannels earlier on the breathing cycle as compared with P01.

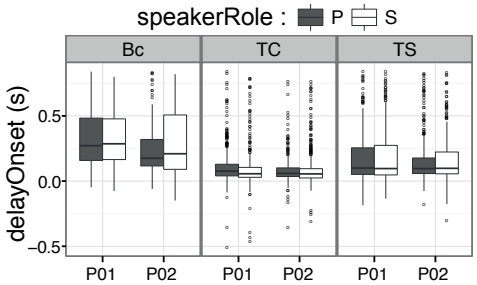


Figure 5. Delay between the offset of inhalation and the beginning of speech according to the dialogue event (*BC, TC, TS*) for the dyads involving P01 and P02 and for the partners (*P*) and the subjects (*S*).

3.5. Delay between speech offset and next inhalation

When speakers talked and then inhaled again, the delay between speech offset and the inhalation onset was shorter when the speaker continued the turn than when she started a new turn. This was observed for both subjects and partners (Figure). *delayOffset* was generally longer when the speaker switched her role and became the listener. This was particularly the case for P02 as compared with P01, while no clear change in subjects' behaviors was observed according to the partner.

The distribution of *delayOffset* is given in Figure 7 when the next breath group started with a turn continuation, a turn start (black and gray bars) or a turn start only (blue bars). The duration of *delayOffset* was in most of the cases shorter than 200 ms. The probability of the next speech group to be a turn start increased with the increase of the duration of *delayOffset*. This was particularly evident for the two partners. Additionally, P02 produced longer *delayOffset* than P01. The distribution was comparable between P01 and subjects who talked to her, but an asymmetry was observed between P02 and subjects talking to her, with a larger number of observations with shorter *delayOffset* for subjects than for P02.

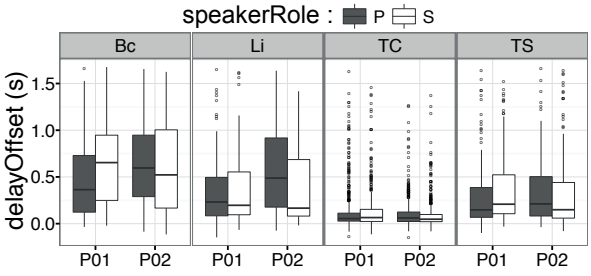


Figure 6. Delay between the offset of the last speech group on the breathing cycle and the onset of the next inhalation according to the dialogue event (*BC, LI, TC, TS*) for the dyads involving P01 and P02 and for the partners (*P*) and the subjects (*S*).

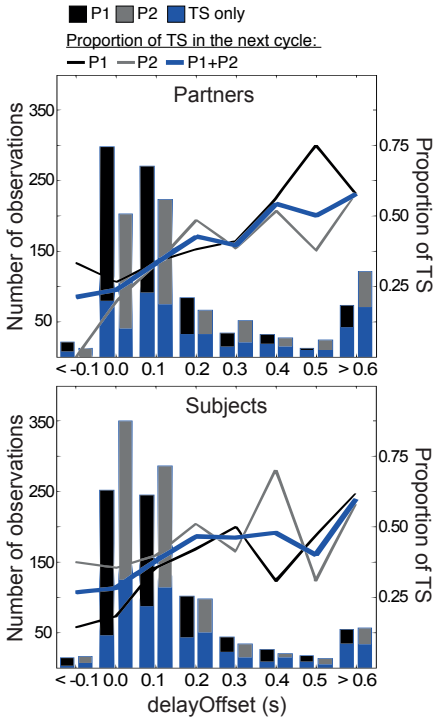


Figure 7 Distribution of the breathing cycles according to the delay between speech offset on the breathing cycle and the onset of the next inhalation. The results are given for the partners (*top*) and the subjects (bottom). The superimposed curves represent the proportion of turn start (*TS*) in the next breathing cycle (*TS*/(*TS*+*TC*)) for each lag.

4 Discussion and conclusion

The aim of the current study was to investigate changes in breathing profiles with respect to the conversation partner during spontaneous dialogue. We observed differences in the two partners' behaviors. P02 was breathing and speaking faster than P01. If breathing would determine dialogue rhythms [21] and if interlocutors were systematically adapting their physiological rhythms to each other through the verbal exchange (as observed before in more controlled tasks [19-20]), we could expect an increase in subjects' syllable and breathing rates when talking with P02 as compared to P01. This was yet not the case: the current analyses of breathing and syllable rate did not provide clear evidence of subjects' adaptation to partners' behaviors. This could be explained by the fact that the recordings were relatively short (2.5 mins * 5 recordings for each dyad) and that longer and continuous interpersonal interactions may be required to generate adaptations in physiological rhythms. Moreover, familiarity with the interlocutor may play a role in conversations. For example, romantic partners or members of a same family may display stronger adaptation of physiological rhythms [31] than persons who did not know each other very well as in the current study.

Our experimental paradigm involved two non-naïve partners. This choice allows testing the same partners for all subjects, similarly as in our previous work on breathing adaptation during listening [18]. However, for dialogue, this choice introduced an asymmetry in the interlocutors' roles: partners were aware they should maintain the dialogue, ending in conversations globally balanced between subjects and partners. Some variability was yet observed according to the subject: some subjects were talking more than the partners, some others less, with some interaction between subjects and partners. This suggests that the partners were not fully controlling the dialogue and that a better understanding of interpersonal adaptation of breathing may require a better description of socio-cultural and human factors [32]. It is also possible that partners were adapting more to the subjects than the reverse, as they were non-naïve, which could have preserved the subjects' rhythms.

The main difference in subjects' behaviors according to the partner was observed in the delay between speech offset and the onset of the next inhalation. Even if the effect was not statistically tested, it seems that subjects reduced this delay when talking with P02 as compared with P01. By contrast this delay was longer when talking with P02. Detailed analyses of turn types in the current dataset [33] showed that P02 was interrupted more often than P01, which could explain the longer delay at the end of the breathing cycle. The reverse could also be true: P02 could be interrupted more often due to longer pauses before inhalation. Despite the variability in subjects and partners' behaviors, and complex profiles due to inter-individual interactions, inhalation pauses during dialogue also showed strong and consistent patterns related to the dialogue events. The duration of inhalation was shorter when inhalation occurred inside a turn than before the start of a new turn. This "compression" of the breathing gap was also observed in the delay before the inhalation onset and after the inhalation offset. It suggests that partners and subjects avoid gaps inside turn by strongly reducing inhalation. This cue could signal to their interlocutor that they want to keep the floor. Such a strategy may limit the chance of the interlocutor to take the turn. It may also have physiological consequences such as hyper- or hypo-ventilation [3].

As readers used their breathing to indicate thematic changes or continuations [14], interlocutors of a dialogue may control the characteristics of their breathing noises – timing, amplitude and duration – to take or maintain their turn. Because the position of the microphone relative to the speakers' mouth was not precisely controlled in the current study, breathing noises could not be reliably analyzed. Faster inhalation may yet be related with louder noises, indicating to the interlocutor that the speaker wanted to keep or take the turn. By contrast, one could expect silent inhalations during listening, as inhalation noises produced by the listener may interfere with the perception of her interlocutor's speech.

Breathing during dialogue is a complex stream that alternates between different control levels of breathing, partially explained by the switching between speaking and listening [24]. Previous work found a tendency of interlocutors to breathe in-phase or in anti-phase at turn taking [24]. A similar analysis of interpersonal coordination in breathing has been carried out using the current dataset. Results showed no clear coordination of breathing profiles at a global level, but specific coordination patterns according to the type of turn [33]. Together with the current analyses, this seems to confirm the idea that during dialogue, breathing is strongly shaped by the communicative constraints [23]. However more studies are now required to understand the participation of breathing to these communicative constraints, the relationship between breathing changes and mutual accommodation in general [34] and the consequences of dialogue constraints on ventilation.

5 Acknowledgements

This work was funded by a grant from the BMBF (01UG0711) and the French-German University to the PILIOS project. The authors want to thanks Caroline Magister, Jörg Dreyer, Anna Sapronova, and Uwe Reichel for their help with data collection and labelling.

6 References

[1] Conrad, B. & Schönle, P. (1979). Speech and respiration. *Arch Psychiatr Nervenkr, 226*, 251-268.

[2] McFarland, D. H. & Smith, A. (1992). Effects of vocal task and respiratory phase on prephonatory chest wall movements. *J Speech Hear Res, 35*, 971-982.

[3] Hoit, J. D., & Lohmeier, H. L. (2000). Influence of continuous speaking on ventilation. *J Speech Hear Res*, *43*(5), 1240-1251.

[4] Grosjean, F. & Collins, M. (1979). Breathing, pausing and reading. *Phonetica*, 36(2), 98-114.

[5] Cutler, A., D. Dahan and W. Van Donselaar (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech* 40, 141-201.

[6] Conrad, B., Thalacker, S., & Schönle, P. (1983). Speech respiration as an indicator of integrative contextual processing. *Folia Phoniatr, 35*, 220-225.

[7] Winkworth, A. L., Davis, P. J., Ellis, E., & Adams, R. D. (1994). Variability and consistency in speech breathing during reading: lung volumes, speech intensity, and linguistic factors. *J Speech Hear Res, 37,* 535-556.

[8] Whalen, D. H. & Kinsella-Shaw, J. M. (1997). Exploring the relationship of inspiration duration to utterance duration. *Phonetica, 54*, 138-152.

[9] Huber, J. E. (2008). Effects of utterance length and vocal loudness on speech breathing in older adults. *Respir Physiol Neurobiol*, 164, 323-330.

[10] Fuchs, S., Petrone, C. Krivokapic, J. & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *J of Phonetics 41*, 29-47.

[11] Winkworth, A. L., Davis, P. J., Adams, R. D., & Ellis, E. (1995). Breathing patterns during spontaneous speech. *J Speech Hear Res, 38*, 124-144.

[12] Wang, Y. T., Green, J. R., Nip, I. S., Kent, R. D., & Kent, J. F. (2010). Breath group analysis for reading and spontaneous speech in healthy adults. *Folia Phoniatr Logop, 62*, 297-302.

[13] Rochet-Capellan, A., & Fuchs, S. (2013) The interplay of linguistic structure and breathing in German spontaneous speech. *Proceedings of Interspeech 2013, Lyon*, paper 1228.

[14] Bailly, G. & Gouvernayre, C. (2012). Pauses and respiratory markers of the structure of book reading. *Proceeding of Interspeech*, 2012, Portland, OR. Paper?

[15] Whalen, D. H., Hoequist, C. E., & Sheffert, S. M. (1995). The effects of breath sounds on the perception of synthetic speech. *J Acoust Soc Am*, *97*, 3147-X.

[16] Ainsworth, S. (1939). Empathic breathing of auditors while listening to stuttering speech. *JSHD* IV, 139–156.

[17] Brown, C.T. (1962). Introductory study of breathing as an index of listening. *Speech Monographs*, 29(2), 79–83.

[18] Rochet-Capellan, A., & Fuchs, S. (2013). Changes in breathing while listening to read speech: the effect of reader and speech mode. *Frontiers in Psychology*, *4*.

[19] Bailly, G., Rochet-Capellan, A., & Vilain, C. (2013). Adaptation of respiratory patterns in collaborative reading. *Proceedings of Interspeech 2013, Lyon*, paper 54.

[20] Müller, V. & Lindenberger, U. (2011). Cardiac and respiratory patterns synchonize between persons during choir singing. *PLoS ONE* 6(9), e24893.

[21] Warner, R.M., Waggener, T.B. & Kronauer, R.E. (1983). Synchronized cycles in ventilation and vocal activity during spontaneous conversational speech. *J Appl Physiol Respir Environ Exerc Physiol*. 54(5), 1324–34.

[22] Autesserre, D., Nishinuma, Y. & Guitella, I. (1989). Breathing, pausing, and speaking in dialogue. *Proceedings of Europeech Paris,* 2433–2436.

[23] Guaïtella, I. (1993). Experimental study of respiration in spontaneous dialogue. *Folia Phoniatrica* 45, 273–279.

[24] McFarland, D. H. (2001). Respiratory markers of conversational interaction. *J. Speech Lang. Hear. Res.,* 44, 128-X.

[25] Babel, M. & Bulatov, D. (2011). The role of fundamental frequency in phonetic accommodation. *Speech and Language* 55(2) 231–248.

[26] Manson, J.H., Bryant, G.A., Gervais, M.M. & Kline, M.A. (2013). Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behaviour* 34(6), 419-426.

[27] Levitan R., Hirschberg J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. *Proceedings of Interspeech 2011*, Florence, 3081–3084.

[28] Boersma, P., & Weenik, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 5.3, retrieved from http://www.praat.org/

[29] Reichel, U.D. (2012). PermA and Balloon: Tools for string alignment and text processing. *Proceedings of Interspeech 2012, Portland,* paper 346.

[30] Beňuš, S., Gavano, A. & Hirschberg, J. (2011) Pragmatic aspects of temporal accommodation in turn-taking. *J of Pragmatics* 43, 3001–3027.

[31] Helm, J. L., Sbarra, D., & Ferrer, E. (2012). Assessing cross-partner associations in physiological responses via coupled oscillator models. *Emotion*, *12*(4), 748-762.

[32] Giles, H., Coupland, N., & Coupland, J. (1991) Accomodation Theory: Communication, context, and consequence. *Contexts of Accomodation: Developments in Applied Sociolinguistics*, 1–68.

[33] Rochet-Capellan, A. & Fuchs, S. (under review).

[34] Lee, C.-C., A. Katsamanis, M. P. Black, B. Baucom, A. Christensen, P. G. Georgiou and S. S. Narayanan (2014). Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions. *Computer, Speech, and Language* 28(2), 518-539.

13 Thursday 2

Yuhuan Wu tone and the role of sonorant onsets

Carlos Gussenhoven [1], *Lu Wang* [2]

[1] Department of Linguistics, Radboud University Nijmegen, Netherlands
[2] Department of English, East China Normal University, Shanghai, China

c.gussenhoven@let.ru.nl l.l.wang@outlook.com

Abstract

Co-occurrence restrictions on tones and consonants in Yuhuan Wu Chinese syllables provide a powerful illustration of the phonetic basis of phonological contrasts, with sonorant contexts allowing more tone contrasts than other contexts. Interestingly, the language also reveals that the phonetic implementation of tones depends on the phonological contrast it is involved in. Such phonetic enhancement may be the opposite of what could be expected on the basis of speech ergonomics. Moreover, the language has a tone deletion rule that exempt tones in the context with the largest number of contrasts, showing a phonological version of enhancement.

Index Terms: tone, enhancement, Wu

1. Introduction

The relation between phonetics and phonology is bidirectional. On the one hand, phonetic considerations determine the structure of phonological systems, because the interacting ergonomics of speech production and speech perception will be reflected in them (e.g. [1],[2],[3]). We will refer to this connection as the ergonomic relation between phonetics and phonology. On the other hand, the articulatory implementation of phonological forms will reflect the phonological contrasts in them [4]. Speech production is biased towards enhancing specific contrasts and the articulation of the same feature will therefore vary across languages as a function of the contrasts it is involved in ('phonetic knowledge' [5]). In turn, phonetic measures to protect contrasts may be incorporated in the language's phonological systems, either as 'transphonologizations', whereby an enhancement feature takes over the contrastive role of a phonological feature, or distributionally, whereby exceptions arise in otherwise general rules [6]. These situations will be referred to as enhancement, where the phonetic behavior has not been incorporated in the phonological structure, and phonologized enhancement [7], when it has, whether as a transphonologized contrast or as a distributional restriction.

2. Segments, syllables and tones

We present the case of the lexical tones of Yuhuan Wu Chinese as one that particularly clearly illustrates these relations. The variety is spoken in Yuhuan County in the southeast of Zhejiang Province and has some 300,000 speakers. Like other Wu Chinese dialects, notably Shanghainese, Yuhuan Wu has a number of co-occurrence restrictions on consonants and tones in the same syllable.

2.1. Segmental structure

The language's consonant system is given in Table I.

Table I. Contrastive consonants of Yuhuan Wu Chinese

	labial	coronal	velar	glottal
plosive	b p p^h	d t t^h	g k k^h	ʔ
affricate		dz ts ts^h		
nasal	m	n	ŋ	
fricative	f v	s z	x	ɦ
lateral		l		

Onsets and codas are optional. All consonants can occur as onsets, except /ʔ/. Only /ʔ/and /ŋ/ can occur in the coda. The vowels that occur in open syllables are given in Table II. In addition, syllabic /z̩/ occurs after onset /s, z/, while syllabic /ŋ/ occurs more generally.

Table II. Contrastive vowels of Yuhuan Wu Chinese in open syllables

	Front unround	Front unround	Back unround	Back round
high	i	y		u
mid	ɛ	ø		o
diphthong	ei			əu
low			ɑ ɑ̃	ɔ ɔ̃

Additionally, the vowel /ə/ appears in closed syllables. Rimes ending in /ʔ/ have /i, ɛ, ø ,o/ or /ə/, while rimes ending in /ŋ/ have /i, o/ or /ə/. Discounting the onset consonant, some rimes may begin with [j, w], which we interpret as prevocalic /i, u/. Prevocalic /i/ appears before /u, ɛ, ø, ɔ, ɑ, ɑ̃/ and /oŋ/, while prevocalic /u/ appears before /ɛ, ei, ɔ̃, ɑ, ɑ̃, əŋ/ and /oʔ/. The syllable structure is thus (C)(G)V(C), where G stands for the prevocalic glide. The labiovelar prevocalic glide appears only after velars and the glottal /ɦ/ as well as syllable-initially; the palatal prevocalic glide appears syllable-initially and after all onset consonants except /f, v, x, ŋ/.

2.2. Tones and consonants

The first distinction to be drawn is that between sonorant rimes, those containing just a vowel or a vowel-plus-/ŋ/, with or without prevocalic glide, and glottal rimes, those ending in /ʔ/, again with or without prevocalic glide. The number of tone contrasts in sonorant rimes is five, H, L, HL, ML, LH. However, the presence of a glottal coda restricts that number to two, H and L, notated as Hq and Lq to indicate their occurrence in glottal rimes. This is shown in Table III. The explanation for this restriction is the short duration of the voiced portion of glottal rimes [7]. The reason for the reduced number of contrasts is the short duration of the sonorant portions of glottal rimes. Over five randomly chosen words with H-tone in each category, sonorant portions of glottal

rimes average 122 ms, against 438 ms for sonorant rimes. This is evidently a connection between phonetics and phonology of the first kind.

Table III. Tone contrasts in sonorant and glottal rimes

li	H	*plum*
li	HL	*surname*
li	L	*plough*
li	LH	*pear*
li	ML	*separate*
loʔ	Hq	*sway*
loʔ	Lq	*fall down*

Sonorant consonants have a constriction in the vocal tract that allows the air pressure levels on either side to be relatively equal [8]. Despite the higher air pressure behind the constriction, the relatively open constriction prevents a build-up of air pressure, and the interference by these consonants with the process of vocal fold vibration as occurring during vowels is therefore small or absent. Voiced obstruents, by contrast, have a constriction behind which the air pressure builds up, and any voicing during the constriction will be made difficult, because of the impedance the constriction offers to the air flow and the resulting reduction in the pressure difference below and above the glottis. As a result, the rate of vocal fold vibration during and after voiced obstruents is reduced. During voiceless obstruents, the vocal folds are abducted to prevent voicing. When voicing sets in after them, the tenser vocal folds will tend to vibrate faster than after the voiced sonorants; for references see [9, p. 8]. These considerations explain two features of the Yuhuan tone system. First, after obstruent onsets the number of tone contrasts is smaller than after sonorant onsets. Second, the tones that appear after voiced obstruents form a lower selection than those that appear after voiceless obstruents. This is shown in Table IV, which should be compared with Table III. Again, we are dealing with a connection between phonetics and phonology of the first kind.

Table IV. Tone contrasts after obstruent onsets

pʰi	H	*fart*
pʰi	HL	*drape over shoulder*
pʰi	LH	*semi-finished product*
pɔ	H	*leopard*
pɔ	HL	*precious things*
pɔ	LH	*bag*
bɔ	ML	*hug*
bɔ	L	*carpenter's plane*
bɔ	LH	*robe*

This restricted distribution not only applies to plosives and affricates, with their three-way laryngeal contrast, but also to fricatives, which have a two-way laryngeal contrast. As predicted, in glottal rimes, voiceless onsets can only be followed by H, while voiced obstruents can only be followed by L.

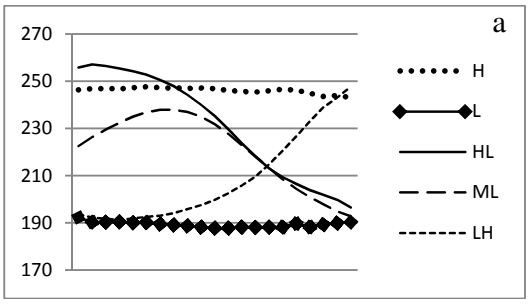
An apparent deviation from the distribution shown in Tables II and IV concerns syllables that begin with phonetic [w] and [j], after which six tones appear. Closer inspection of the beginnings of these syllables, however, indicates that such syllables with H and HL are onsetless, while those with L and ML begin with /ɦ/. The two rises, which differ in pitch range, are to be analyzed as LH occurring after onsetless syllables and syllables beginning with /ɦ/. In the onsetless syllables, a weakly released glottal closure often occurs, which never occurs in the syllables beginning with /ɦ/. Moreover, weak breathy voice may attend syllables beginning with /ɦ/. Strikingly, however, neither of these features appears to be obligatory. Table V presents the contrasts in sonorant rimes after onsetless syllables and syllables beginning with /ɦ/ which contain a prevocalic glide. Two tones appear in glottal rimes, H after onsetless syllables and L after /ɦ/-initial ones.

Table VI. Tone contrasts in onsetless and /ɦ/-initial rimes

uei	H	*fed up*
uei	HL	*bowl*
uei	LH	*hello*
ɦuei	L	*meeting*
ɦuei	ML	*refuse*
ɦuei	LH	*(proper name)*

The distribution in Table VI shows that the prevocalic glide is not an onset. If it were, it would have to be classed as a sonorant consonant. That is, there would have been five tones after a phonetic [w] or [j], not six. As it is, onsetless syllables pattern with voiceless obstruents, while /ɦ/-initial rimes pattern with voiced obstruents. Phonetically, this makes sense, since a glottal closure involves vocal fold tensing. Phonologically, there is a problem with the characterization of voiceless obstruents and zero as a natural class. Analyzing empty onsets as containing /ʔ/ would not work, since the glottal stop is not [-voice], nor are voiceless obstruents [+constricted glottis].

We recorded isolated syllables with all five tones and onset and coda conditions as illustrated in Tables III and IV from three speakers. Average time-normalized f0 tracks for one speaker are given in Fig. 1, pooled over onset conditions, for sonorant rimes (panel a) and glottal rimes (panel b). The pitch range is approximately 60 Hz, small for a female speaker; the values for H and L are fairly constant across tones; ML differs from HL in having a lower starting point.



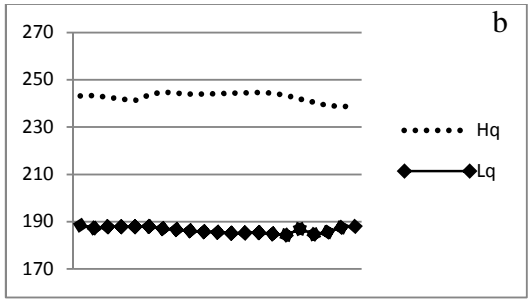


Figure 1: *F0 tracks on normalized times scales for H, L (n=24), HL (n=56), ML, HL (n=36) (panel a) and Hq (n=20) and Lq (n=16) (panel b). Speaker WL.*

3. Phonologized enhancement

Varieties of Wu have prosodic words, aka tone units, whose tonal pattern tends to be determined by the leftmost or rightmost syllable. On the basis of an investigation of disyllabic prosodic words, we find that Yuhuan preserves the tone on the rightmost syllable, and as such provides the mirror image of Shanghainese [10, 11]. That is, tones on non-final syllables are deleted. This is illustrated in Fig. 2, which shows the five (seven) tones before final LH. Thus, the final tone(s) don't spread left; rather, the preceding syllable is pronounced at mid or low pitch, a default pitch for what we take to be toneless syllables. Middish pitch appears in the case of H and LH (approx. 193 Hz), low pitch in the case of underlying L, ML, Hq and Lq (approx. 178 Hz). Differences of that magnitude are found in many cases of phonological identity in the data, and we cannot at this point say if the difference between the two groups of tones is systematic.

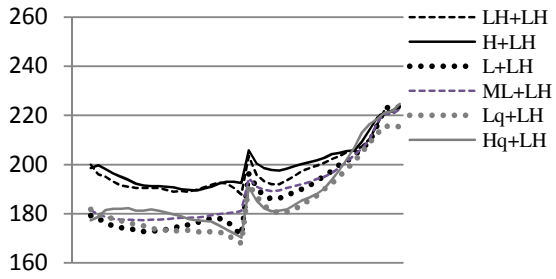


Figure 2. *F0 tracks on normalized times scales for LH, H, L, ML before final LH. Speaker WL.*

There is, however, a striking exception to this tone deletion rule. HL in syllables with a sonorant onset consonant does not delete. This non-final tone deletion rule is given in (1). Fig. 3 presents the fate of HL before LH and L.

(1) Pre-final tones are deleted, except HL in a syllable with a sonorant onset.

There are two questions that this exception raises. One is why HL is preserved after sonorant rather than obstruent onsets. The answer lies in the number of contrasts that are neutralized by the deletion. After obstruents, the reduction is from three forms to one, a loss of two; after sonorants, it is from five forms to two, a loss of three. A loss of four would place a greater strain on the system, which is apparently enough for it to be prevented. The other question is why it is HL rather than one of the other four tones that is preserved. One possible answer is that it provides the clearest phonetic contrast with mid or low pitch, another that it is the most frequent tone in this kind of syllable. In a corpus of 80 random syllables, the highest number is for L (23), followed by ML (20), HL (16), LH (10) and H (9). This suggests that only the first answer is correct. Relative to mid pitch, neither H nor L are very distinct, while because of their smaller pitch range, both ML and LH are closer to mid pitch than is HL. We show HL preservation in syllables with sonorant onsets before LH and L in Fig. 3.

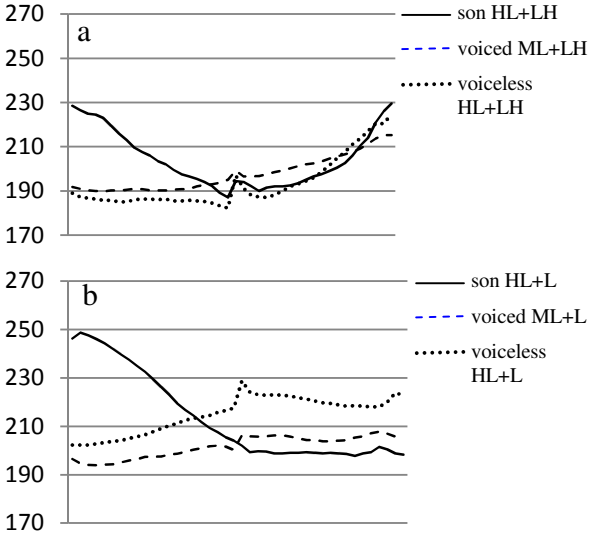


Figure 3: *F0 tracks on normalized times scales for HL before final LH (panel a) and L (panel b) broken down by the phonological specification of the onset (sonorant, voiceless obstruent and voiced obstruent). Speaker WL.*

Summarizing, because onset sonorants have no appreciable physiological effect on the f0 of the rime, Yuhuan Wu has more tone contrasts after sonorant onsets than after obstruent onsets. This distribution presupposes an analysis of initial glides as pre-nuclear glides in the rime, as opposed to onset glides. The observation illustrates the effect of phonetic ergonomics on phonological structure. Second, because there is a larger number of tones after sonorant onsets, an otherwise general tone deletion rules exempts HL from deletion if it occurs in rimes preceded by a sonorant onset. The exception illustrates that phonetic implementation will respect and preserve phonological contrasts, which tendency may in turn lead to distributional restrictions.

4. Phonetic enhancement

Phonetic enhancement can be observed in the way ML is pronounced in syllables with sonorant onsets as compared with voiced obstruents. If things were left to the phonetic implementation without regard for the place of the phonological representation in the system of contrasts, sonorant onsets would have no effect on the f0 of the initial part of the rime, while after voiced obstruents, the f0 of the first part of the rime would be depressed somewhat due to the impedance of offered by the constriction. After voiceless obstruents, the initial f0 after HL would be elevated relative to the situation for HL after sonorant onsets. In reality, the elevation after voiceless obstruents is negligible. Panel (a) of Fig 4 shows that there is no difference in the pronunciation of HL in the two onset conditions. This suggests that the tendency for raised f0 after voiced obstruents is counteracted

by a similar raising of the pitch after sonorants. Since it is only after sonorant onsets that the contrast between HL and ML exists, a comparison for ML should reveal the policy behind the lack of a perturbation effect in panel (a). Panel (b) in fact shows that the initial pitch after sonorant onsets in considerably *lower* than after voiced obstruents, against expectation, if expectation is based on speech ergonomics. By contrast, if we take into consideration that speech behavior is guided by contrast enhancement as well as speech ergonomics, the pattern in panel (b) can be understood as a manifestation of 'phonetic knowledge' [5].

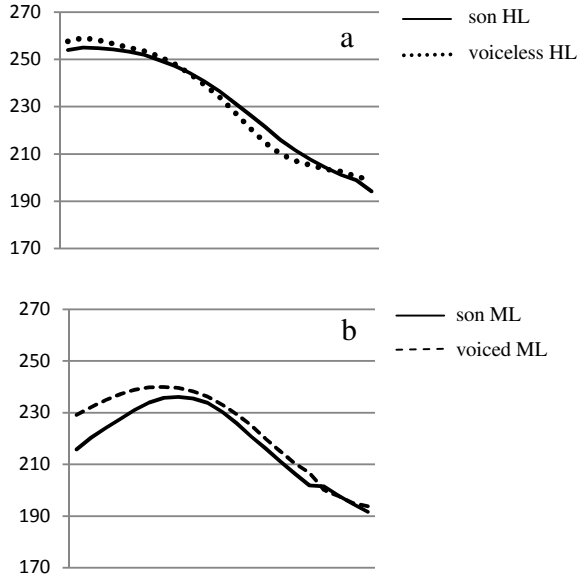


Figure 4: *F0 tracks on normalized times scales for HL with sonorant and voiceless obstruent onset (panel a) and ML with sonorant and voiced obstruent onset (panel b). Speaker WL.*

5. Discussion

Yuhuan Wu appears to be a textbook case for the illustration of the complex relation between phonetics and phonology. On the one hand, phonological systems reflect the opportunities afforded by the speech production and perception systems for creating phonological contrasts, the speech ergonomic connection [1, 2, 3]. In particular, patterns of f0 perturbation by onset consonants explain why there are five tone contrasts after sonorant onsets and three after onsetless or obstruent onsets. Similarly, the large difference in duration between sonorant and glottal rimes explains why glottal rimes have two tone contrasts and sonorant rimes five [7]. On the other hand, contrast enhancement is illustrated by the exceptional retention of HL in pre-final position just in the situation that the onset is sonorant, whereas H, L, ML and LH are deleted in all onset conditions. This suggests that exhaustive deletion of the five-way tone contrast would have jeopardized the maintenance of a comfortable level of lexical distinctions.

Finally, the rationale behind the distributional pattern arising from post-sonorant tone retention is suggested by the phonetic enhancement of the ML tone after sonorant onsets. It consists of a lowering of the initial part of the pitch contour which is more extreme that the physiologically induced lowering due to onsets consisting of voiced obstruents. This shows the enhancement of the contrast between HL and ML in the only context in which it exists, in sonorant rimes after sonorant onsets.

Phonetic enhancement, therefore, may lead to new phonological contrasts ('transphonologization'), but may also lead to restrictions in the distribution of phonological features. Having said that, we cannot be certain about the developmental path that led to the exceptional retention of HL in prefinal sonorant-initial syllables [12]. The deletion may have started gradually, with the retained tones lagging behind as motivated by contrast preservation, which pattern ultimately became phonologized. Alternatively, the deletion may have involved one tone first, which rule then expanded its focus tone by tone, stopping short at HL.

6. Conclusions

The tone grammar of Yuhuan Wu reveals a bidirectional relation between phonetics and phonology. While it has five lexical tones after sonorant onsets in sonorant rimes, lower numbers of tone contrasts appear after obstruent onsets and in glottal rimes. This pattern, variants of which can be found in many languages, in particular Wu dialects of Chinese, is explained by speech ergonomics (f0 perturbation and rime duration). The opposite direction in the relation between phonetics and phonology is shown by a tone deletion rule. Pre-final tone deletion preserves HL if it occurs in a syllable with a sonorant onset. This exception is explained by phonological contrast preservation. Third, the language shows that the contrast between HL and ML, which only appears in sonorant rimes with sonorant onsets, is enhanced, such that the beginning pitch of ML is lower than after voiced obstruents, quite against ergonomic phonetic considerations. After voiced onsets, ML does not contrast with HL, so that less differentiation between the tone sis called for here than om syllables with sonorant onsets.

7. References

[1] Martinet, A., "Elements of General Linguistics". London: Faber and Faber. 1964.
[2] Flemming, E., "Auditory Representations in Phonology". New York: Routledge. 2002. [PhD dissertation MIT 1995.]
[3] Boersma, P., "Functional Phonology: Formalizing the Interactions between Articulatory and Perceptual Drives". The Hague: Holland Academic Graphics. 1998.
[4] Gussenhoven, C. and Kager, R., "Introduction: Phonetics in phonology". Phonology 18: 1-6. 2002.
[5] Kingston, J. and Diehl, R.L., "Phonetic knowledge", Language 70: 419-454. 1994.
[6] Hyman, L.M. "Phonologization". In A. Juilland (ed.), Linguistic studies presented to Joseph H. Greenberg. Saratoga: Anma Libri. 407-418. 1976.
[7] Zhang, J. "The effects of duration and sonority on contour tone distribution: Typological survey and formal analysis". Doctoral dissertation UCLA. 2001.
[8] Chomsky, N. and Halle, M. "The sound pattern of English". Harper & Row. 1968.
[9] Gussenhoven, C. "The phonology of tone and intonation". Cambridge University Press. 2004.
[10] Zee, E. and Maddieson, I. "Tones and tone sandhi in Shanghai: Phonetic evidence and phonological analysis," Glossa 14: 45-88. 1980.
[11] Chen, Y. "Revisiting the phonetics and phonology of Tone Sandhi in Shanghai Chinese". Proceedings of Speech Prosody 2008, 253-256. Campinas, Brazil. 2008.
[12] Hyman, L.M. "Enlarging the scope of phonologization". In A.C.L. Yu (ed.) Origins of sound change: Approaches to Phonologization. Oxford University Press.

An Investigation of Prosody in Hindi Narrative Speech

Preethi Jyothi[1], *Jennifer Cole*[1,2], *Mark Hasegawa-Johnson*[1,3], *Vandana Puri*

[1]Beckman Institute, University of Illinois at Urbana-Champaign, USA
[2]Department of Linguistics, University of Illinois at Urbana-Champaign, USA
[3]Department of ECE, University of Illinois at Urbana-Champaign, USA
{pjyothi,jscole,jhasegaw}@illinois.edu,vanu.p.sharma@gmail.com

Abstract

This paper investigates how prosodic elements such as prominences and prosodic boundaries in Hindi are perceived. We approach this using data from three sources: (i) native speakers of Hindi without any linguistic expertise (ii) a linguistically trained expert in Hindi prosody and finally, (iii) classifiers trained on English for automatic prominence and boundary detection. We use speech from a corpus of Hindi narrative speech for our experiments. Our results indicate that non-expert transcribers do not have a consistent notion of prosodic prominences. However, they show considerable agreement regarding the placement of prosodic boundaries. Also, relative to the non-expert transcribers, there is higher agreement between the expert transcriber and the automatically derived labels for prominence (and prosodic boundaries); this suggests the possibility of using classifiers for the automatic prediction of these prosodic events in Hindi.

Index Terms: Hindi prosody, perception study, automatic labeling of prosodic events in Hindi.

1. Introduction

Hindi is one of the most widely spoken Indo-European languages in the world with over 200 million native speakers in northern parts of India. There have been a sizeable number of studies on intonation in Hindi. Many early works studied the phenomenon of lexical stress in Hindi words which manifests itself as prominence via a designated syllable [1, 2, 3, 4] and acoustic evidence for lexical stress in Hindi [1, 2, 4, 5]. There are two consistent observations regarding prosody in Hindi across previous work (refer to [1, 3, 6, 7] among others): 1) every content word (i.e. a prosodic word), except for the phrase-final one, is associated with a rising pitch contour and 2) focus induces post-focal pitch range compression.

Prior work confirms the presence of pitch accents on content words with a rising pitch contour. However, there are varying opinions regarding the reason these pitch contours are triggered [8, 9, 10]. Most recently, Féry and colleagues [6, 9, 11] claim that Hindi does not have prominence-lending pitch accents and uses only prosodic phrasing to structure an utterance, with edge-marking phrase tones. As evidence for this claim, they note that Hindi speakers do not produce a consistent pattern of pitch movement on a stressed syllable, and in general have very weak intuitions about the location of stress prominence at the word level.

A question of particular interest to us, and one of the objectives of this paper, is to investigate whether ordinary untrained native listeners of Hindi consistently perceive prosodic elements in Hindi speech (specifically, prosodic prominence and prosodic phrase boundaries). This technique of involving ordinary listeners to derive prosodic transcriptions (the latter will henceforth be referred to as *non-expert transcriptions*) was successfully implemented by Cole et al. [12] for English. This is categorically different from most of the previous work on Hindi prosody; the latter is predominantly based on production studies where trained experts analyzed sentences spoken by native Hindi speakers for evidence of various prosodic elements. To our knowledge, there has not been a systematic enquiry into ordinary listeners' perception of prosody in Hindi speech. We use a corpus of Hindi narrative speech to conduct our perception study, described in more detail in Section 2.

This paper also attempts to initiate the discussion about whether we can automatically detect prosodic elements such as pitch accents and prosodic boundaries in Hindi speech. There is a large body of research that studies the identification and classification of prosodic events in English ([13, 14, 15, 16, 17] are a sampling of some of the important works in automatic labeling of English prosody). Automatic prosody labeling is a relatively unexplored area for Hindi. Many of these studies make heavy use of the ToBI Standard [18] – a formalized notation developed to describe the intonation of Standard American English. Recently, a publicly available toolkit called AuToBI [19] has been developed to automatically detect and classify prosodic events (using ToBI labels) in English. As a first step, we use models of prosody trained on English obtained via AuToBI to automatically label prosodic pitch accents and phrase boundaries in Hindi speech. We hope this investigation informs us of what would be needed to build improved models of Hindi prosody. This could also prove to be useful for the design of automatic speech recognition systems in Hindi.

To summarize, the objectives of this paper are two-fold:

1. *Perception of prosody in Hindi by ordinary listeners*: What is the untrained, ordinary listener's perception of prosodic prominence and phrase boundaries in Hindi? How does this compare to the prosodic transcription by a linguistically trained expert Hindi listener? Do native listeners consistently identify pitch accents in Hindi speech? What about phrase boundaries? These are some of the questions we try to address; the experiments are detailed in Section 3.

2. *Automatic labeling of prosody in Hindi*: How do trained models of prosody in English perform when evaluated on Hindi data? Is the automatic labeling of prosodic events more consistent with the non-expert transcriptions or the expert transcription? What can be deduced from the Hindi evaluation task to build better prosody models for Hindi? These questions are discussed further in Section 4.

We conclude this paper with a closing discussion along with scope for future work in Section 5.

Figure 1: A screenshot of the user interface for the study. Prosodically prominent words turn red on being selected.

2. Speech materials

The speech material used in our experiments was drawn from the "OGI Multi-language Telephone Speech Corpus" [20]. This corpus consists of telephone speech in eleven languages, including Hindi; the corpus has recorded speech from 198 Hindi speakers. The Hindi speech was collected in narrative form by asking volunteers to talk about any topic for up to a minute. Sixty-eight of these one-minute audio clips have corresponding hand-labeled phonetic transcriptions.

Out of the sixty-eight audio clips with phonetic transcripts, we selected ten and extracted excerpts, one from each clip, averaging 24.10 secs in length and averaging 59.2 in the number of words per excerpt; there are a total of 592 words over all excerpts. Since the speech in the OGI corpus was collected from volunteers speaking impromptu, it contains many occurrences of conversational elements such as disfluencies, hesitations and repetitions. Our ten excerpts were chosen such that the utterances in each excerpt were relatively free of disfluencies and the usage of English words were kept to a minimum.[1]

3. Perception of prosody in Hindi by non-expert transcribers

3.1. Method and Experimental Setup

Ten adult native speakers of Hindi participated in this study.[2] All the participants were English-speaking students living in the United States. Most of them could speak and write only Hindi and English (and not other languages); three of them could additionally understand other Indian languages such as Kannada, Oriya and Bhojpuri. Information about their language background was retrieved via a questionnaire administered along with the main study.

The entire study was conducted with the help of a web-based software [21]. The interface of the experiment and the instructions for the prosody transcription tasks were worded in Hindi. This was done in order to, hopefully, make the participants more receptive to prosodic elements in the Hindi excerpts.

The non-expert transcribers were shown the ten excerpts described in Section 2 in a randomized order. For each audio file, a participant was asked to complete two tasks – firstly, listen to how the speaker breaks up the text into chunks and mark the location of chunk boundaries (*prosodic phrase boundaries*) and secondly, mark the words that are emphasized or stand out relative to the other words in the utterance (*prosodic prominence*). The participants were explicitly informed that the phrase chunks do not necessarily have to coincide with any punctuation. In order to get acquainted with the two tasks, the experiment was preceded by a training session using one speech excerpt.

On identifying a chunk, the participant was asked to select the final word in the chunk and a "/" delimiter was inserted after the word to mark the phrase boundary. For the second task of identifying emphasized words, the participants' boundary markers from the previous task were kept visible for them to refer to. Figure 1 shows a snapshot of the interface for an excerpt during the prominence marking task. There was no limitation on the response times. The entire experiment was set up such that it would not exceed an hour. However, the participants could listen to each excerpt any number of times and they could choose to devote any amount of time to each excerpt.

As a second set of experiments, this entire run of two tasks per excerpt for all ten excerpts was repeated – only this time the transcripts were displayed without any accompanying audio clip. Each participant was asked to 'listen' to their own inner speech while reading the excerpt text and mark the word chunks and emphasized words. By removing the associated speech clips, the participants would have to rely entirely on lexico-syntactic cues to guide their annotation.

Finally, we also asked an expert in Hindi prosody to transcribe the same data using ToBI.[3] The expert transcriber was provided the audio signals, along with pitch and intensity tracks and phonetic and word alignments (the annotations were done using the Praat [22] toolkit). We note here that the conditions under which the expert interprets prosody is positively different from the conditions for ordinary listeners. Our experimental results also point to this difference, as detailed in Section 3.2.

3.2. Experimental results and discussion

We first compute Cohen's kappa agreement coefficients [23] between the ordinary listeners' and the expert's transcriptions of prominences and boundaries. This is a fairly standard measure of agreement that takes into account the chance probability of agreement. Values of 0.01 to 0.2 indicate slight agreement, 0.21 to 0.4 is fair agreement and 0.41 to 0.6 indicates moderate agreement. Fig. 2 shows the distribution of agreement coefficients across all the participants for both prominence and boundary labels. This shows that ordinary listeners perceive boundaries (with a moderate $\kappa = 0.41$) much more similarly to the expert than prominences (with a slight $\kappa = 0.15$). The slight agreement for prominence marking in Hindi has a parallel in English. In English, prominences that are more closely tied to meaning (e.g., the nuclear prominence that marks focus) are more reliably marked by transcribers than pre-nuclear prominences, which may serve a rhythmic function ([24, 25]).

We also compute Fleiss' kappa statistics (typically used to compute agreement across multiple transcribers) across all listeners (ignoring the expert transcriptions). Fig. 3 shows Fleiss' coefficients for both prominences and boundaries; *With audio* and *Without audio* specify non-expert transcriptions obtained

[1] Most of the volunteers who helped collect Hindi data for the OGI multi-language corpus were graduate students in the United States and often made use of English words while speaking impromptu in Hindi.

[2] One participant listed "Marwari" as their native tongue but identifies themselves as a native Hindi speaker.

[3] The last author served as our expert. She is a native speaker of Hindi and a simultaneous English-Hindi bilingual (much like the non-expert transcribers).

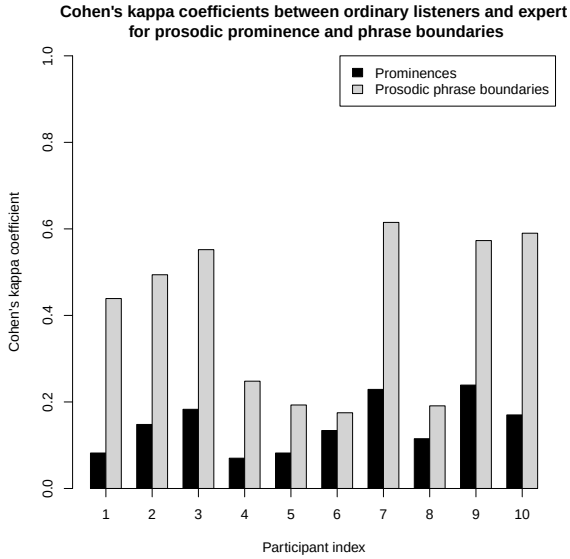


Figure 2: Cohen's kappa agreement coefficients for each participant against the expert for prominences and boundaries.

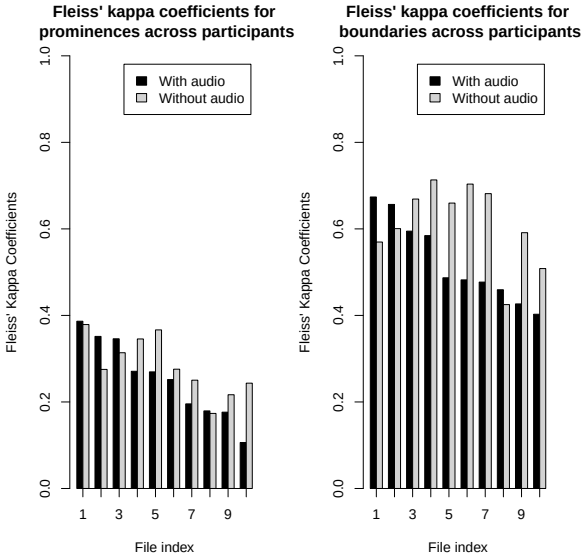


Figure 3: Fleiss' agreement statistics for prosodic prominences and boundaries across all the participants.

with audio and without audio, respectively.[4] We focus first on the non-expert transcriptions with audio. We observe that the non-expert transcribers agree on the location of boundaries well above chance (mean $\kappa = 0.524$) and agree with one another more than they agree with the expert ($\kappa = 0.407$). There is only a fair amount of agreement on prominences ($\kappa = 0.253$)). This partially supports Féry's [9] prediction that Hindi speakers will reliably and consistently perceive prosodic phrases in Hindi utterances and will not reliably perceive prosodic prominence as distinct from phrasing in Hindi utterances.

Comparing the non-expert transcriptions with and without audio, the mean κ values for both prominences and boundaries are comparable (0.25 vs. 0.28 and 0.52 vs. 0.61, respectively). The distributions of non-expert transcriptions with and without audio in Fig. 3, however, suggest that the transcribers may not be getting cues from the same information structure.

Finally, we compute the rate of occurrence of prominences and boundaries. The mean length of intervals (in words) between prominences range from 5.4 – 11.4 for each audio clip, across all transcribers. This indicates the speaker dependent variation of the rate of prominences. Similarly, for boundaries, this range is 5.3 – 8.4. We also compute the mean prominence and boundary intervals for each listener averaged over data across all the clips: 5.0 – 14.0 and 4.6 – 18.5, respectively. This corresponds to listener dependent variation. We note that the listener dependent variation is larger than the speaker dependent variation as previously observed for American English [26].

4. Automatic prosody detection in Hindi

4.1. Method and Experimental Setup

AuToBI [19] is a publicly available toolkit to automatically detect the presence and type of prosodic events, from the ToBI standard, present in a speech sample. The toolkit is accompanied by a number of trained models[5] of pitch accent and phrase boundary detection (and classification using ToBI labels); we use models for pitch accent detection and intonational phrase boundary detection, trained on three spontaneous speech corpora of Standard American English. The classifications are performed using the logistic regression algorithm with a range of pitch, intensity and duration input features [19]. The trained models were evaluated on all ten Hindi excerpts to derive labels indicating pitch accents and prosodic phrase boundaries.

4.2. Experimental results and discussion

Fig. 4 shows the kappa values for the automatically derived labels against the expert for both prominences and phrase boundaries; a confusion matrix with details of the insertion and deletion errors of AuToBI relative to the expert are also shown. We see that AuToBI almost never (only for 2 words) predicts a boundary when the expert does not. However, there are many instances (126 words) where AuToBI does not predict a boundary after the word while the expert does. These errors mainly stem from instances where a new prosodic phrase begins even when there is no preceding silence; this silence is an important feature for AuToBI to detect a boundary. For prominences,

AuToBI\Expert	Accent	No accent
Accent	74% (130/175)	37% (156/417)
No accent	26% (45/175)	63% (261/417)

Kappa coefficient: **0.311**

AuToBI\Expert	Boundary	No boundary
Boundary	43% (95/221)	0.5% (2/371)
No boundary	57% (126/221)	99.5% (369/371)

Kappa coefficient: **0.479**

Figure 4: Confusion matrix of AuToBI predictions against expert predictions, along with the kappa agreements, for both prominences and phrase boundaries.

[4] The speech files were sorted in descending order according to the Fleiss' coefficients of the "with audio" case.

[5] The toolkit and trained models are available at the following website: `http://eniac.cs.qc.cuny.edu/andrew/autobi/`.

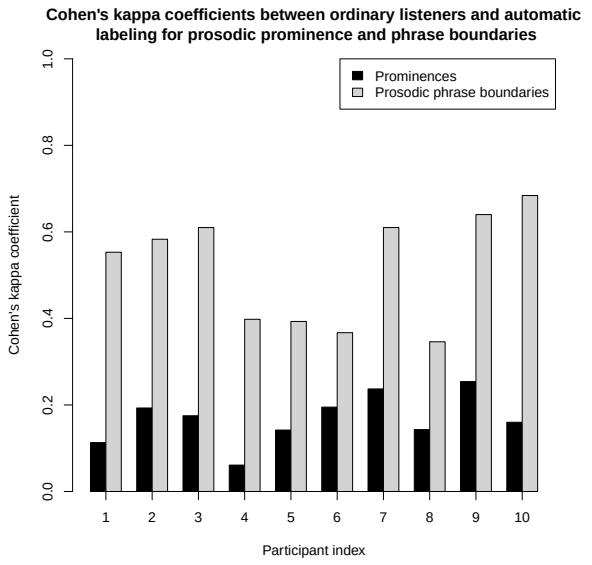


Figure 5: Cohen's kappa agreement coefficients for each participant against the automatically derived transcriptions for prosodic prominence and prosodic phrase boundaries.

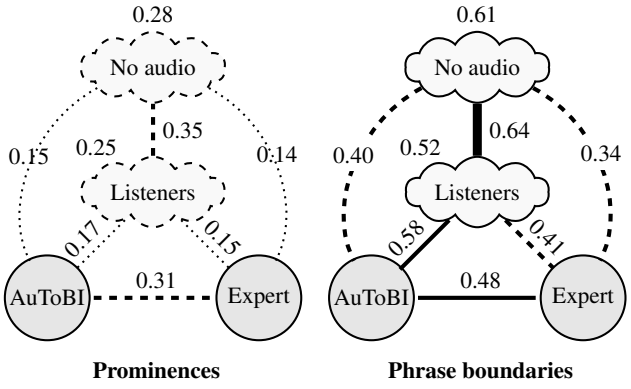


Figure 6: Kappa agreements between the non-expert transcribers, both using audio (*Listeners*) and without audio (*No audio*), AuToBI and the expert, for prosodic prominence and boundaries (shown on the left and right, respectively). The dotted lines indicate no agreement, the dashed lines indicate fair agreement, the bold lines indicate moderate agreement and the thick bold line indicates substantial agreement (according to the interpretation of the kappa statistic in [27]).

the false-positives result from words with a rising pitch accent which get classified as being prominent due to the pitch excursion (but are actually not prominent according to the expert).

Fig. 5 shows Cohen's kappa coefficients for each participant against the AuToBI predictions. As observed in Fig. 2, the listeners show a much higher value of agreement for phrase boundaries (mean $\kappa = 0.582$) than for prominences (mean $\kappa = 0.167$). Fig. 6 summarizes the agreement statistics between the non-expert transcriptions (both with and without audio), the expert transcriptions and the automatically derived transcriptions. We emphasize the following points:

1. The automatically derived labels for both prosodic events show fair to moderate agreement with the expert. This suggests the possibility of using AuToBI in the future for automatic prominence and boundary labeling in Hindi.

2. AuToBI predicts the non-expert transcribers' boundary scores better, but for prominence it is a better prediction of the expert's labels. This reaffirms the claim that ordinary Hindi listeners (unlike experts and machines) do not have a consistent internal definition of prominence.

3. The listeners are more in agreement with each other than with the expert; κ between the listeners and the expert for prominences fall within $[0.07, 0.24]$ while κ of the listeners with each other is in $[0.04, 0.49]$. This suggests that both are possibly tapping into different criteria for prosody perception.

4. In perceiving prosodic boundaries, the *Listeners* and *No audio* groups show moderate and substantial agreement with each other ($\kappa = 0.524$ and $\kappa = 0.612$, respectively). Further, for each participant, there is substantial agreement between the boundaries perceived with and without audio (κ is in the range $[0.55, 0.86]$). This suggests a fairly consistent bias amongst the listeners regarding what is expected of the task.

5. Listeners have much lower agreement for prominence than for boundaries. But the findings also show that, relative to the non-expert listeners, there is higher agreement between the expert transcriber and AuToBI on prominence (κ between listeners and AuToBI fall in the range of $[0.06, 0.25]$ showing that none of the listeners agree with AuToBI as much as the expert does, with $\kappa = 0.31$). This suggests that there are acoustic patterns in Hindi speech that are similar to the acoustic patterns that mark prominence in English, and further, that a trained Hindi speaker can discriminate among words on the basis of these acoustic patterns, as a basis for identifying prominence.

5. Conclusions and future work

We observe that non-expert listeners have much lower agreement for prominence than for boundaries amongst themselves as well as with the expert. On the other hand, AuToBI is more in agreement with the expert on prominence, relative to the non-experts. The fact that non-expert listeners fail to identify prominence on the basis of the same cues used by the expert and the machine suggests that either the patterns of acoustic prominence do not function to mark important linguistic information in Hindi, or they may serve multiple functions that are not easily lumped together in a single percept. Future research on Hindi is needed to investigate prominence under a wider range of pragmatic conditions (beyond contrastive focus), in production and perception.

We have found that automatic models of prosody for English make fairly good predictions about prosody in Hindi. We hope to improve on these models by fine-tuning them using labeled Hindi data; this would allow us to use relatively limited amounts of labeled Hindi data as opposed to building models of Hindi from scratch. We also propose to make use of these models in automatic speech recognition systems for Hindi.

6. Acknowledgements

This research was supported in part by a Beckman Postdoctoral Fellowship for the first author. The second and third author's contributions were supported by NSF BCS 12-51343 and QNRF NPRP 09-410-1-069, respectively. The authors gratefully acknowledge Tim Mahrt at the University of Illinois, Urbana-Champaign for developing the web software used in our perception study, Language Markup and Experimental Design Software (LMEDS).

7. References

[1] P. Moore, "A study of Hindi intonation," Ph.D. dissertation, University of Michigan, 1965.

[2] M. Ohala, "A search for the phonetic correlates of Hindi stress," C. M. Bh. Krishnamurti and A. Sinha, Eds., 1986, pp. 81–92.

[3] J. D. Harnsberger, "Towards an intonational phonology of Hindi," *Ms., University of Florida*, 1994.

[4] R. Nair, "Acoustic correlates of lexical stress in Hindi," in *Linguistic Structure and Language Dynamics in South Asia–papers from the proceedings of SALA XVIII roundtable*, 2001.

[5] L. O. Dyrud, "Hindi-Urdu: Stress accent or non-stress accent?" Ph.D. dissertation, University of North Dakota, 2001.

[6] U. Patil, G. Kentner, A. Gollrad, F. Kügler, C. Féry, and S. Vasishth, "Focus, word order and intonation in Hindi," *Journal of South Asian Linguistics*, vol. 1, pp. 53–70, 2008.

[7] V. Puri, "Intonation in Indian English and Hindi late and simultaneous bilinguals," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2013.

[8] S. Genzel and F. Kügler, "The prosodic expression of contrast in Hindi," in *Proceedings of Speech Prosody*, 2010.

[9] C. Féry, "Indian languages as intonational 'phrase languages'," in *Festschrift to honour Ramakant Agnihotri*, I. Hasnain and S. Chaudhury, Eds. Aakar Publisher, 2010.

[10] A. Sengar and R. Mannell, "A preliminary study of Hindi intonation," in *Proceedings of SST*, 2012.

[11] C. Féry and G. Kentner, "The prosody of embedded coordinations in German and Hindi," in *Proceedings of Speech Prosody*, 2010.

[12] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 1141–1177, 2010.

[13] M. Q. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech & Language*, vol. 6, no. 2, pp. 175–196, 1992.

[14] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 2, no. 4, pp. 469–481, 1994.

[15] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of Interspeech*, 2002.

[16] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model," in *Proceedings of ICASSP*, 2004.

[17] S. Ananthakrishnan and S. S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[18] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "TOBI: a standard for labeling english prosody," in *Proceedings of ICSLP*, 1992.

[19] A. Rosenberg, "AuToBI-a tool for automatic ToBI annotation." in *Proceedings of Interspeech*, 2010.

[20] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multilanguage telephone speech corpus," in *Proceedings of ICSLP*, 1992.

[21] J. Cole, T. Mahrt, and J. I. Hualde, "Listening for sound, listening for meaning: Task effects on prosodic transcription," To appear in *Proceedings of Speech Prosody*, 2014. [Online]. Available: http://prosody.beckman.illinois.edu/lmeds.html

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," Version 5.3.51, retrieved from http://www.praat.org/.

[23] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[24] S. Calhoun, "Information structure and the prosodic structure of English: A probabilistic relationship," Ph.D. dissertation, The University of Edinburgh, 2007.

[25] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the perception of prosodic prominence," *Laboratory Phonology*, vol. 1, no. 2, pp. 425–452, 2010.

[26] Y. Mo, J. Cole, and E.-K. Lee, "Naive listeners' prominence and boundary perception," *Proceedings of Speech Prosody*, 2008.

[27] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

Prosodic focus marking in Bai

Zenghui Liu[1], *Aoju Chen*[1,2] *& Hans Van de Velde*[1]

Utrecht University[1], Max Planck Institute for Psycholinguistics[2]

l.z.h.liu@uu.nl, aoju.chen@uu.nl, h.vandevelde@uu.nl

Abstract

This study investigates prosodic marking of focus in Bai, a Sino-Tibetan language spoken in the Southwest of China, by adopting a semi-spontaneous experimental approach. Our data show that Bai speakers increase the duration of the focused constituent and reduce the duration of the post-focus constituent to encode focus. However, duration is not used in Bai to distinguish focus types differing in size and contrastivity. Further, pitch plays no role in signaling focus and differentiating focus types. The results thus suggest that Bai uses prosody to mark focus, but to a lesser extent, compared to Mandarin Chinese, with which Bai has been in close contact for centuries, and Cantonese, to which Bai is similar in the tonal system, although Bai is similar to Cantonese in its reliance on duration in prosodic focus marking.

Index Terms: focus, prosody, duration, pitch, Bai

1. Introduction

This is a study of prosodic focus marking in Bai. The Bai language is spoken in the Southwest of China by more than one million people of the Bai ethnic group. It has eight lexical tones from three tonal categories: level (55, 44, 33), rise (35), and fall (42, 21, 32, 31) [1, 2, 3]. The term 'focus' refers to the part of a sentence that conveys new information on a topic, following [4] and [5]. Focus can differ in the size of the focused constituent and contrastivity. In terms of size, focus can be on a whole sentence (broad focus), or on a lexical word (narrow focus). If narrow focus also conveys an explicit contrast to alternatives in the context, it is termed as contrastive focus [6]. There is some debate on the position of Bai within the Sino-Tibetan group [2], but this discussion is not crucial for the topic of our paper. Pertinent to the current study is that Bai has been in close contact with Mandarin Chinese for centuries [7].

Pitch is used to encode focus in many non-tone languages [6, 8, 9]. Previous studies have shown that pitch also plays an important role in signaling focus in some tone languages. For example, in Stockholm Swedish, a lexical pitch accent language with two contrasting lexical accents, a separate high tone is added to the lexical accent to mark focus, making pitch relevant for both lexical and post-lexical distinctions [10]. In Mandarin Chinese, a tone language with four lexical tones, pitch is used as a major prosodic cue to realize focus in addition to duration [11]. According to [11], the pitch range at the focus is substantially expanded; the pitch range after the focus is lowered as well as compressed; and the pitch range before the focus does not really deviate much from the neutral focus condition. In addition to that, in Mandarin Chinese, syllable duration increases significantly under focus, regardless of the position of the syllable in the utterance. The data presented in [11] reveals that the pre-focus constituents undergo little change in pitch and duration compared to the same constituents in focus in Mandarin. The same is true for Vietnamese, a tone language with six lexical tones [12, 13]. However, other tone languages do not use pitch to mark focus. For example, in Cantonese, a language with six lexical tones, pitch variation is not systematically modified to mark focus. According to [13], duration and intensity are the main acoustic correlates of focus in Cantonese; both are increased significantly in the on-focus words in any word location for all lexical tones. Besides, no decrease in mean pitch range is found in the post-focus words. In Yucatec Maya, a language with two lexical tones, pitch is only used at the lexical level and focus is not prosodically encoded [15, 16, 17].

Such differences in prosodic focus marking among tone languages suggest that which cues are used to what extent can vary from language to language and is not related to the total number of lexical tones in a language. Against this background, we investigate how pitch and duration may be used to mark focus in the southern variety of Bai by adopting a semi-spontaneous experimental approach. The southern dialect of Bai is chosen because it is well studied at the segmental and lexical level compared to other varieties of Bai [1, 2, 3, 18, 19]. The southern variety of Bai is hereafter referred to as Bai.

2. Methodology

2.1. Experimental materials

The production experiment aimed to elicit SVO sentences in five focus conditions: narrow-focus on the subject NP in sentence-initial position (NF-i), narrow-focus on the verb in sentence-medial position (NF-m), narrow-focus on the object NP in sentence-final position (NF-f), broad focus (BF) and contrastive-focus on the verb in sentence-medial position (CF-m). The focus condition was set up by a WH-question or a statement from the experimenter, as illustrated in examples (1) to (5), where focused constituents appear in square brackets.

Target sentence: *ɕo*^{42} *tɯ*^{21} *kɯ*^{21} *tsɯ*^{33}.

Bear one (quantifier) *sell* *tree.*

The bear sells the tree.

(1) Experimenter: Look! The tree. There is also a price label. It seems someone sells the tree. Who sells the tree?

Participant: [THE BEAR] sells the tree. (NF-i)

(2) Experimenter: Look! The bear and the tree. It seems like that the bear does something with the tree. What does the bear do with the tree?

Participant: The bear [SELLS] the tree. (NF-m)

(3) Experimenter: Look! The bear, it stands behind a shelf. It seems like that the bear sells something. What does the bear sell?

Participant: The bear sells [THE TREE]. (NF-f)

(4) Experimenter: Look! This picture is very blurring. I can't see anything clearly. What has been depicted in the picture?

Participant: [THE BEAR SELLS THE TREE]. (BF)

(5) Experimenter: Look! The bear and the tree. It seems like that the bear does something with the tree. I guess the bear wipes the tree.

Participant: The bear [SELLS] the tree. (CF-m)

Each focus condition was realized in 30 SVO sentences. The lexical tones of verbs were strictly controlled. The verbs were the items for acoustic and statistical analysis, as they can have multiple roles played in the five focus conditions. For example, the verb could be a focused constituent in the BF, NF-m and CF-m conditions; it can also be a pre-focus constituent in the NF-f condition and a post-focus constituent in the NF-i condition. The property of the verbs in the present setting thus provided us with an opportunity to investigate effects of both focus (focused vs. unfocused) and focus type (narrow focus vs. broad focus vs. contrastive focus) on pitch and duration in Bai.

In order to keep the experiment within a feasible length, three lexical tones were included, representing the three tonal categories existing in Bai: level, falling, and rising tones. Tone in Bai can be considered as a complex combination of pitch, phonation type, and degree of tenseness [2]. The present study has selected lexical tones that were well spread over the tonal space of Bai's tone system. Specifically, 55 was selected as a representation of level tones, 21 as a representation of falling tones, and 35 as the rising tone.

The target sentences were constructed in such a way that each was a unique combination of a subject-noun and a VP (verb + object-noun). Six verbs were included, two in each tonal category. In Bai, the noun needs to be followed by a quantifier to form an NP as a subject in a sentence, but the quantifier of the NP can be omitted when the NP is an object in a sentence [20]. Four subject-nouns were selected, which followed by a same low fall-tone quantifier in all the target sentences. Four level-tone object-nouns were selected. The six verbs and four object nouns formed 24 VPs, each of which appeared in each focus condition. This gave us 120 VPs. The subject nouns were evenly distributed over the 120 VPs to form 120 target sentences. To make sure that the duration of the experiment was manageable and reasonable for the participants, the 120 target sentences were split into two lists. Each list contained all the five focus conditions realized on different sentences, and all the six representations of the tones, but only half of the V+O combinations. This results in 60 items per list and participant.

2.2. Data elicitation

In the picture-matching game, three piles of pictures were used: the experimenter and the participant each held a pile of pictures ordered in a certain sequence; the third pile of pictures were scattered around on a table. In the experimenter's pictures (the first pile), there was always something missing, like a subject, an action (verb) or an object. The participant's pictures (the second pile) all contained a complete event. The participant's task was to help the experimenter with sorting out pictures from her own pile and the third pile that went together. Here is a detailed example of a trial eliciting a target sentence in the NF-i condition: First, the experimenter took a picture (e.g. a tree) from her own pile, drew the participant's attention to the picture and established what the picture was by saying, e.g. *"Look! The tree! There is also a price label. It seems like that someone sells the tree."* This was done to make sure that the entity in the picture was referentially given to the participant before the utterance of the question. Second, the experimenter asked a question about the picture (e.g. "*Who sells the tree?"*). Third, the participant took a complete picture from his or her pile and looked at it. The experimenter then repeated the question, followed by an answer from the participant (e.g. "*[THE BEAR] sells the tree.*"). Fourth, the experimenter found the picture containing the missing information in the third pile and pairs it up with her own picture. The participants were explicitly instructed (1) to respond in full sentences and (2) not to show their own pictures to the experimenter. Prior to the picture-matching game proper, the experimenter conducted six practice trials with the participant to familiarize him or her with the game.

In order to ensure the consistency in the participants' word choice, the picture-matching game was preceded by a picture-naming task, which was designed to familiarize the participants with the target words and the entities in the pictures used in the game.

2.3. Participants and procedure

Five native speakers of Bai (four male and one female, aged between 23 and 25) took part in the experiment. The participants all met the following criteria: (1) using Bai on a daily basis with self-estimated daily use exceeding 60%; (2) not having lived outside the Bai speaking community for the past 10 years; (3) not having used Chinese or other languages for a long period on a daily basis; (4) having no self-reported speech and hearing impairments.

Every participant was randomly assigned to one of the two lists. The game lasted 20 to 25 minutes per participant. The participants were tested individually by a female experimenter, who was a native speaker of Bai, in a quiet room in a villager's private home. The experiments were recorded using a portable ZOOM H1 digital recorder at a 44.1 kHz sampling rate and 16 bit accuracy. Each session was also video-taped.

3. Analysis and Results

3.1. Analysis

The auditory recordings from each participant were first orthographically annotated so that the participant's responses could be selected. A strict selection criterion of the usable data was applied, i.e. a sentence was considered usable only if it contained no self-correction and hesitation and was uttered as a response to the target question. In total 80.3% of the obtained responses (N=241) were included in further analysis. The usable sentences were subsequently acoustically annotated in Praat [21]. A textgrid with four interval tiers (word, tone, sentence, comment), and two point tiers (pitch, duration) was created for each target sentence. Every sentence was segmented into words in the 'word' tier, then landmarks demarcating verb onset and offset, and the locations of pitch-maximum and pitch-minimum within the verb were added to the 'duration' and 'pitch' tiers. The landmarks for the onset and offset of verbs were determined according to the information in the waveform and spectrogram.

The pitch values of the pitch landmarks and the time values of the word boundaries were subsequently extracted via Praat scripts. Two measures from these values were calculated: word duration (i.e. offset time minus onset time) and pitch range (i.e. the difference between the maximum pitch and the minimum pitch). In 55 of the usable responses, the pitch values could not be reliably measured. These responses were thus excluded from the analysis on pitch range.

In order to investigate how focus is prosodically realized in Bai, several analyses were done to find out the effect of focus (focused vs. unfocused), focus type, and the interaction between these variables and the tone of the verb on the duration and pitch range of the verbs. To find out the effect of focus and its interaction with tone, we compared the duration and pitch range of the focused verbs with these measures of the verbs in the unfocused conditions: NF-m (focus) vs. NF-i (post-focus); NF-m (focus) vs. NF-f (pre-focus). To find out the effect of focus type and its interaction with tone, we compared the verbs in the NF-m condition with the same verbs in the BF and CF-m conditions.

3.2. Results

3.2.1. Duration

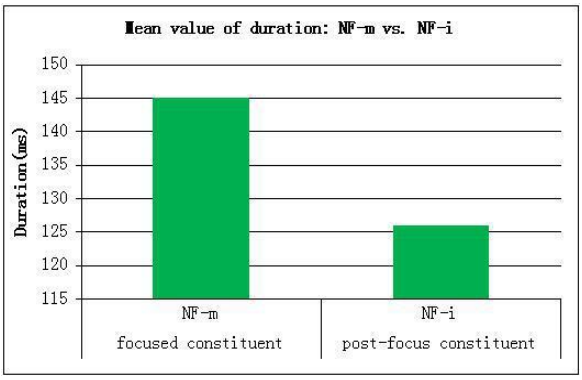


Figure 1: *Mean duration (in ms) of verbs in focus vs. non-focus position. (NF-m=verb in sentence medial focused position, NF-i=verb preceding a focused constituent)*

The duration data obtained from the verbs in the NF-m and NF-i conditions showed that the verbs were on average 19.1 ms longer when focused (NF-m) than when not focused and following a focused constituent (NF-i). To assess the effect of focus (focused vs. unfocused) on the duration of the verbs, we built a mix-effect model with 'focus' as the fixed factor (independent variable), 'speaker' and 'verb' as the random factors; and another mix-effect model with only the random factors. The variation between speakers was corrected in the models per focus condition. A statistically significant difference between the two models was taken as the evidence for a main effect of the fixed factor at issue. In the comparison between NF-m and NF-i, the model including the fixed factor differed significantly from the model with only the random factors ($p<0.01$). This indicated that the speakers used duration to distinguish focus from non-focus when the unfocused verb was in post-focus position. Furthermore, we also used mixed-effect-modeling to assess the effect of the interaction between 'focus' and 'lexical tones of verbs'. The model involving the interaction did not differ significantly ($p=0.31$) from the model without the interaction. Thus, the speakers used duration to distinguish focus from non-focus regardless of the tonal category of the verbs.

The duration data obtained from the verbs in the NF-m and NF-f conditions showed that the verbs were on average 12.5 ms longer when focused (NF-m) than when not focused and preceding a focused constituent (NF-f). Mixed-effect modeling was used to assess the effect of 'focus' on the duration of the verbs, as described above. It did not reveal a main effect of 'focus' ($p=0.13$). Furthermore, the model involving the interaction did not differ significantly ($p=0.31$) from the model without the interaction. This indicated that there was no main effect of 'focus', i.e. the speakers did not use duration to distinguish focus from non-focus (i.e. pre-focus in NF-f). This suggested that the duration in the pre-focus constituents hardly changed relative to the constituents in focus.

With regard to the effect of focus type, mixed-effect modeling revealed no main effect ($p=0.16$) of 'focus type' (referring to the three types of focus). Further, there was no interaction between 'focus type' and 'lexical tones of verbs' ($p=0.49$). Thus, duration was not used to differentiate the three focus types regardless of the tones of the verbs. To find out whether duration was used to distinguish NF-m and CF-m, two focus types with a smaller focus-constituent size, from BF, we grouped NF-m and CF-m and built new models. The models showed that NF-m and CF-m did not differ from BF in the duration of the verbs regardless of the tone of the verbs ($p=0.33$). Finally, models were built to see whether NF-m and CF-m could differ in duration. Again we found no significant difference in duration.

3.2.2. Pitch range

The mean pitch ranges hardly differed across conditions, i.e. 8.78Hz in the BF condition, 7.99Hz in the CF-m condition, 7.89Hz in the NF-f condition, 10.78Hz in the NF-i condition, and 9.12Hz in the NF-m condition. Mix-effect-modeling confirmed that pitch range was not used in any way in focus marking in Bai.

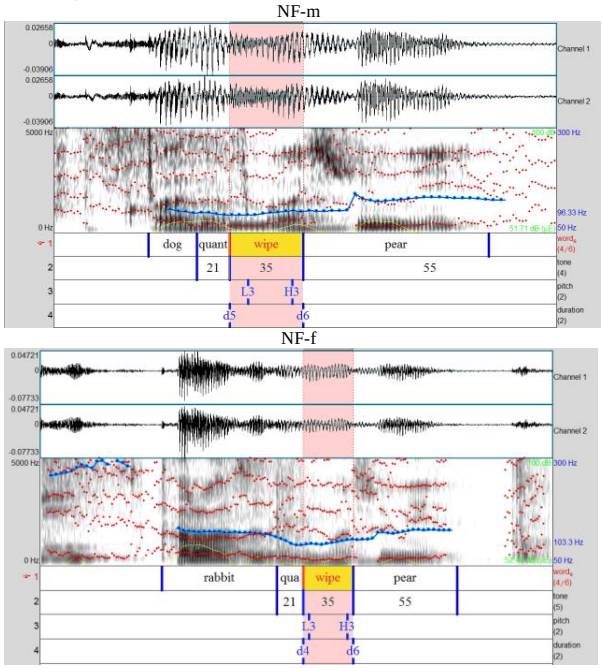


Figure 2: *Pitch contour (in Hz) of verbs in focus vs. non-focus position. (NF-m=verb in sentence medial focused position, NF-f=verb following a focused constituent)*

4. Discussion and Conclusions

The present study shows that speakers of the southern variety of Bai increase the duration of the focused constituent and reduce the duration of the post-focus constituent to encode focus, similar to speakers of Mandarin Chinese [11] and Cantonese [14]. Further, they do not vary the duration of a pre-focus constituent compared to the same constituent in a focused position, again similar to speakers of Mandarin Chinese and possibly Cantonese. However, they do not use pitch variation in any way in focus marking, different from speakers of Mandarin Chinese but similar to speakers of Cantonese. In addition, speakers of this variety of Bai do not use duration distinguish focus types differing in size and contrastivity, different from speakers of Mandarin Chinese [11] and Cantonese [14], who use duration to distinguish focus types differing in the size of the focused constituent. These results suggest that Bai uses duration in prosodic focus-marking to a lesser extent than Mandarin Chinese and

Cantonese. Related to this is the fact that Bai also exploits word order and morphological topic marker to distinguish focal information from non-focal or topical information. Specifically, the canonical word order in Bai is SVO, the word order OSV can be used to highlight the topic status of the object. Further, the topical status of a subject can be optionally marked by topic markers, such as 'nɯ^{55}' and 'lɯ^{44}'[19]. The use of these non-prosodic cues may explain the modest use of prosody in focus marking in Bai. Furthermore, the results add to the existing findings on prosodic focus-marking in tone languages and show that there is no relationship between the cues used to mark focus prosodically and the number of lexical tones in a language.

Our study suggest two topics for future research. Bai has been in close contact with Mandarin Chinese for centuries [7], which has led to a large number of Chinese-loan words in Bai, and deep influence from Chinese syntactic structure on the syntax of Bai [2, 19]. However, in spite of the lexical and syntactic influence from Mandarin Chinese, Bai is by and large more similar to Cantonese in prosodic focus-marking in that both Bai and Cantonese, duration is the major prosodic cue rather than pitch. This puts forward an interesting hypothesis for future research. That is, prosodic focus-marking may not easily undergo changes as a result of language contact. Further, considering the dialectal differences in Bai language, it is not clear whether prosodic focus marking is similar across Bai dialects. Future research on the northern variety of Bai can shed light on this question.

5. Acknowledgements

We are grateful to Wenju He for collecting data in Xizhou County, and to the participants in the village. We also thank Liquan Yang and Yanzhen Zhao for their feedback on Bai, Paula Cox, Anqi Yang, Anna Sara Romøren for their help. This study is supported by a scholarship from the Chinese Scholarship Council to the first author and a VIDI grant (276-89-001) from the Netherlands Organisation for Scientific Research to the second author.

6. References

[1] 赵衍荪, 徐琳, 白语, 中国社会科学院, & 民族研究所. (1996). 白汉词典 四川民族出版社.
(Zhao, Xu, Bai-Chinese dictionary,1996)
[2] Allen, B. (2007). Bai dialect survey. SIL International,
[3] 艾磊, 苏玮雅, & 尹曼芬. (1997). 白语喜洲镇话声调的测试分析. 大理学院学报 (社会科学版), 2, 011.
(Allen, Su and Yin, The experimental study on Xizhou Bai lexical tones, 1997)
[4] Vallduví, E., & Engdahl, E. (1996). The linguistic realization of information packaging. Linguistics, 34(3), 459-520.
[5] Lambrecht, K. (1996). Information structure and sentence form: Topic, focus, and the mental representations of discourse referents Cambridge University Press.
[6] Gussenhoven, C. (2004). The phonology of tone and intonation Cambridge University Press.
[7] Hefright, B. E. (2011). Language Contact as Bilingual Contrast among Bái Language Users in Jiànchuān County, China,
[8] Ladd, D. R. (2008). Intonational phonology Cambridge University Press.
[9] 王蓓, 吐尔逊, 卡得, & 许毅. (2013). 维吾尔语焦点的韵律实现及感知. 声学学报, 38(1), 92-98.
(Wang, Tursun and Xu. Prosodic encoding and perception of focus in Uygur, 2013)
[10] Bruce, G. (1982). Textual aspects of prosody in swedish. Phonetica, 39(4-5), 274-287.
[11] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. Journal of Phonetics,27(1), 55-105.
[12] Jannedy, S. (2007). Prosodic focus in vietnamese. Interdisciplinary Studies on Information Structure, 8, 209-230.
[13] Jannedy, S. (2008). The effect of focus on lexical tones in vietnamese. Experimental Linguistics ExLing 2008, , 113.
[14] Wu, W. L., & Xu, Y. (2010). Prosodic focus in hong kong cantonese without post-focus compression. Speech Prosody 2010,
[15] Kügler, F., & Skopeteas, S. (2007). On the universality of prosodic reflexes of contrast: The case of yucatec maya.
[16] Gussenhoven, C. (2006). Yucatec maya tone in sentence perspective. Poster Presented at LabPhon10, Paris,
[17] Gussenhoven, C., & Teeuw, R. (2008). A moraic and a syllabic H-tone in yucatec maya. Fonología Instrumental: Patrones Fónicos y Variacion, , 49-71.
[18] 邓瑶, & 何稳菊. (2012). 云南大理喜洲白族居民语言生活调查. 民族翻译, 3, 017.
(Den and He. The language attitude survey on Bai in Xizhou, Dali. 2012)
[19] 徐琳. (2008). 大理丛书・白语篇.
(Xu, The Dali series. Bai language. 2008)
[20] 赵燕珍. (2009). 赵庄白语参考语法
(Zhao. Zhaozhuan Bai reference grammar. Doctoral dissertation. 2009)
[21] Boersma, P. (2002). Praat, a system for doing phonetics by computer. Glot International, 5(9/10), 341-345.

Perception of Glottalization in Varying Pitch Contexts in Mandarin Chinese

Maria Paola Bissiri[1], *Margaret Zellers*[2], *Hongwei Ding*[3, 4]

[1] Institut für Akustik und Sprachkommunikation, Technische Universität Dresden, Germany
[2]Department of Speech, Music & Hearing, Kungliga Tekniska Högskolan, Stockholm, Sweden
[3]School of Foreign Languages, Tongji University, Shanghai, China
[4]School of Foreign Languages, Shanghai Jiao Tong University, China

Maria_Paola.Bissiri@tu-dresden.de, zellers@kth.se, hongwei.ding@tongji.edu.cn

Abstract

Although glottalization has often been associated with low pitch, evidence from a number of sources supports the assertion that this association is not obligatory, and is likely to be language-specific. Following a previous study testing perception of glottalization by German, English, and Swedish listeners, the current research investigates the influence of pitch context on the perception of glottalization by native speakers of a tone language, Mandarin Chinese. Listeners heard AXB sets in which they were asked to match glottalized stimuli with pitch contours. We find that Mandarin listeners tend not to be influenced by the pitch context when judging the pitch of glottalized stretches of speech. These data lend support to the idea that the perception of glottalization varies in relation to language-specific prosodic structure.

Index Terms: prosody, voice quality, perception, glottalization

1. Introduction

Glottalization, and in particular creaky phonation, is associated across languages with lowered amplitude, positive spectral tilt, and higher first formant values than modal voice (Gordon & Ladefoged [1]), and with low fundamental frequency (F0), damping, and aperiodicity (Gerratt & Kreiman [2]). Because of its low F0, creak/glottalization has been classically described as giving the impression of "a stick being run along a railing" (Catford [3]: 32). Glottalization, despite its acoustic complexity, can be identified with 95% accuracy by listeners (Blomgren, Chen, Ng & Gilbert [4]), and can therefore be considered a robust tool for phonetic signaling.

Glottalization's acoustic characteristics have often led to it being associated with low pitch, especially in intonation languages. However, associating glottalization with low pitch on the basis of its having low F0 is problematic. Many tone languages associate creakiness with high tones (Gordon & Ladefoged [1]; Gussenhoven [5]), and this can be the case even in intonation languages (Pierrehumbert & Talkin [6]; Pierrehumbert [7]; Dilley, Shattuck-Hufnagel & Ostendorf [8]; Redi & Shattuck-Hufnagel [9]). Dilley et al. [8] found, for example, more frequent glottalizations of word-initial vowels if the target words were pitch-accented. Since in their corpus L* accents were rare, they claimed that glottalization was influenced by pitch-accents per se and not by low F0 ([8]: 439). Pierrehumbert [7] also reported glottalization associated with pitch accents.

In intonation languages, creaky voice and other forms of glottalization have often been associated with prosodic boundaries. Henton & Bladon [10] and Redi & Shattuck-Hufnagel [9] reported that in English glottalization is likely in utterance-final position, while Pierrehumbert & Talkin [6] and Huffman [11] found its presence at prosodic boundaries. Ogden [12, 13] noted that glottal productions characterize possible turn transition locations in Finnish. Creak and/or glottalization have also been shown to occur in other positions. For example, the glottalization of word-initial vowels is a frequent word juncture marker in Czech and in German (Bissiri et al. [14]; Kohler [15]).

Tone languages may demonstrate some of the same intonational features as intonation languages, and this appears to hold true for glottalization features. Ding et al. [16] found glottalization at the beginnings of vowel-initial syllables in Mandarin Chinese, similar to the effects reported above in German, English and Czech. In Mandarin, tone itself can also have an influence on the occurrence of glottalization. Chao [17] observed glottalization in Mandarin's falling-rising Tone 3. Ding & Helbig [18] reported glottalization most frequently in the middle (i.e. associated with the pitch valley) of Tone 3 and sometimes at the end of the falling Tone 4. This would appear to contradict Silverman's [19] and Frazier's [20] proposals that tone and glottalization are perceived sequentially, at least in the case of Mandarin Chinese. It therefore remains the case, as Ní Chasaide & Gobl [21] pointed out, that studying pitch perception without analyzing the influence of glottalization is likely to lead to incomplete results.

Bissiri & Zellers [22] investigated the perception of glottalization in different pitch contexts by German, English and Swedish listeners. They found that long stretches of glottalization tended to be associated with low pitch by listeners in all three of these languages, although listeners could also associate glottalization with mid or rising pitch in a substantial minority of cases. The pitch contour preceding the glottalized stretch influenced English and German listeners' judgments: when the glottalization was preceded by a rising pitch contour, it was more likely to be perceived as falling pitch. Swedish listeners, however, did not show this perceptual effect. Bissiri & Zellers [22] proposed that this was because Swedish listeners' experience with the lexical pitch accent system in Central Swedish made them more sensitive to pitch in general, and therefore more likely to respond based on the actual F0 of the stimuli, than the English and German listeners, who may have relied more on their predictions about the continuing direction of the pitch contour.

If listeners from a pitch-accent language background are more sensitive to F0 than listeners from an intonation language background, it follows that listeners from a tone language should also be more sensitive to F0. The study reported below investigates Mandarin Chinese listeners' perception of glottalized stretches in different pitch contexts.

2. Perception experiment

The current study uses the same AXB forced-choice paradigm as Bissiri & Zellers [22] to investigate whether Mandarin Chinese listeners would consistently associate glottalization of a final syllable with a falling, level, or rising pitch contour in that position. Following Bissiri & Zellers' findings for Swedish versus German and English, we hypothesize that Mandarin listeners, coming from a tone language background, will not be influenced by the surrounding pitch contour.

2.1. Methodology

2.1.1. Stimuli

The stimuli used for this experiment were the same as those used by Bissiri & Zellers [22], and a more detailed account of their creation can be found there. Stimuli consisted of the syllable sequences /da'dada/, /la'lala/, and /na'nana/. Although /da/, /la/ and /na/ are acceptable Chinese syllables, the sequences are nonsense words in Chinese. By means of a Praat (Boersma & Weenink [23]) script, twelve versions of each syllable sequence were created. Nine of them were the Pitch conditions. A flat pitch contour of 220Hz was set over the whole syllable sequence. Then, four pitch points were added: at the onset of the first consonant, the onsets of the middle and final vowels, and the end of the final vowel. The second pitch point (onset of stressed vowel) was either raised 3 semitones above the initial pitch (fall), lowered 3 semitones below the initial pitch (rise), or kept level (mid). Similarly, the fourth pitch point (end of final vowel) was raised 3st, lowered 3st, or kept level. The third pitch point (onset of the final vowel) was held constant. In this way, 9 pitch contours were created (see Figure 1).

For each initial pitch condition (mid, rising, or falling), a further version was created with the final syllable containing glottalization instead of a smooth pitch contour; these stimuli constitute the Glottalization set. The glottalization was taken from naturally-produced tokens, and was matched in length to the final syllables of the Pitch conditions. In order to make sure that the consonant preceding the glottalization remained clear, the glottal stretches were spliced into the stimuli following the formant transitions. Using a Praat script, four additional pitch points were added in the transitions to create the impression of glottalization. These pitch points, with values of 35Hz, 220Hz, 35Hz, and 220Hz respectively, were equally spaced through the duration of the formant transitions. The amplitude of the entire glottalized part of the stimuli was then reduced in order to improve naturalness. All Pitch and Glottalization stimuli were then amplitude-normalized.

2.1.2. Participants

Participants in the experiment were native speakers of Mandarin Chinese (N=21) between 20 and 27 years old, students at the Tongji University in Shanghai.

2.1.3. Experiment procedure

The experiment was presented by means of Praat's ExperimentMFC. Participants heard sequences of three stimuli. Their task was to compare the second stimulus to the first and third, and to decide which it sounded most like. The

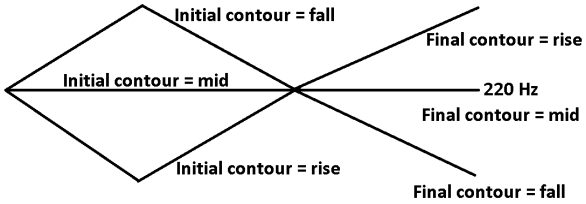


Figure 1: Schematic of pitch contours.

AXB paradigm was chosen instead of ABX because the target stimulus is adjacent to both options, thus putting fewer demands on memory. AXB sets were created from the same syllable set (/da/, /la/, or /na/), and also with the same initial pitch contour. The A and B options were always stimuli from the Pitch set, while the target item X was from the Glottalization set. Since participants in AXB tests tend to choose the more recent and thus better remembered B option, each critical pitch contour comparison was presented in both AXB and BXA format. Thus, each participant heard 126 stimuli in total: 54 test comparisons, in which the X item contained a glottalization, and 72 control comparisons, in which all the items were drawn from the Pitch set. For each participant, the order of presentation of all sequences was randomized by Praat.

The experiment was preceded by a practice session to get participants accustomed to the task. During practice, participants heard 3 control items with the syllables /ba'baba/. According to feedback from a pilot participant, optional breaks after every 32 items were included in the experiment. The time taken by participants to complete the task was between 17-20 minutes.

2.1.4. Data cleanup

Prior to the data analysis, participants' responses to control items were checked to make sure they had accurately matched pitch patterns. The control items always contained two identical contours and one that was different; any participants who failed to identify the identical contours in more than 10% of the control items were considered to have not performed the task correctly. One Chinese listener, who systematically gave the incorrect response in 93% of the control stimuli, was excluded on this basis. Of the 20 Chinese participants whose results were kept, 16 had an accuracy above 95%.

2.2. Results

The cases in which participants had to choose between a) fall and mid, b) fall and rise, and c) mid and rise pitch contours were analyzed separately in three subsets of 360 observations each. We found no statistically significant difference in any of the three data subsets on the basis of the three syllable sequences with /da/, /la/ and /na/; therefore, in the following analysis, the three syllable sets will be collapsed.

Glottalization was most frequently associated with falling pitch, but with a remarkable minority of non-fall responses (see Figure 2). When participants had to choose between fall and mid, fall responses were 65.8%, lower than the percentage of fall responses found by Bissiri & Zellers [22] for German (69.4%), English (73%) and Swedish (66.7%) listeners. Most Chinese listeners identified glottalization with falling pitch in over 50% of the cases; however, two of them gave a majority of mid responses and three gave 50% mid responses. When

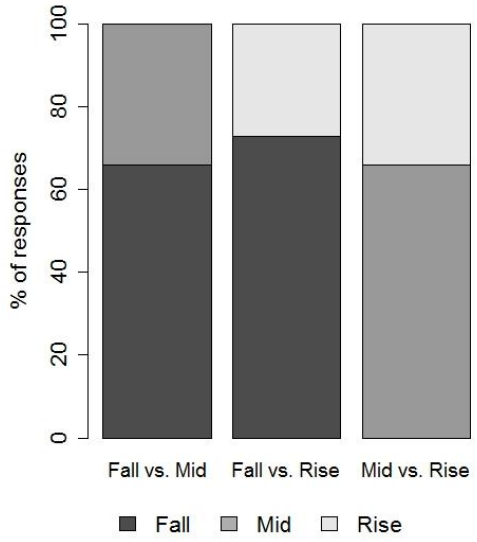


Figure 2: Responses by Chinese listeners.

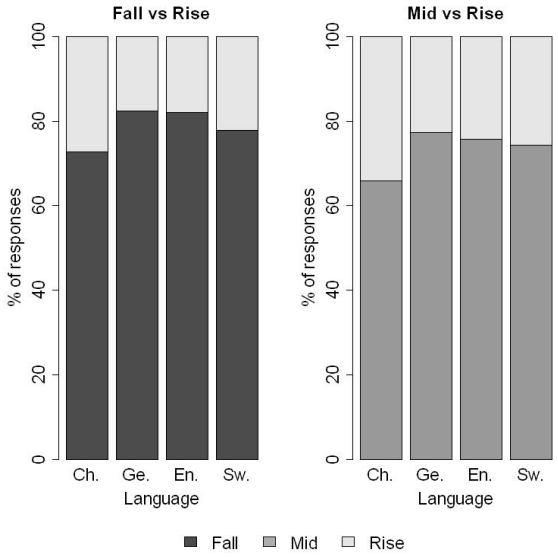


Figure 3: Responses by Chinese listeners and by Bissiri & Zellers' [22] German, English and Swedish listeners when rise was one of the options.

rise was one of the options, there was a substantial minority of rise responses: 27.2% in the fall and rise subset, and 34.2% in the mid and rise subset. The Chinese listeners gave significantly more rise responses than Bissiri & Zellers' [22] German, English and Swedish listeners: subset fall and rise χ^2 (1, N = 1368) = 10.9798, $p < 0.001$, subset mid and rise χ^2 (1, N = 1368) = 14.0703, $p < 0.001$ (see Figure 3). In the fall and rise subset, three Chinese participants had a majority of rise responses. Two of them had 50% rise responses in the mid and rise subset, while three other Chinese participants had a majority of rises.

For Chinese listeners, the initial pitch contour had no influence on the association of glottalization with a final pitch (subset fall and mid: χ^2 (2, N= 360) = 1.556, n.s.; fall and rise: χ^2 (2, N= 360) = 1.374, n.s.; mid and rise: χ^2 (2, N= 360) = 0.9633, n.s.). Figure 4 represents responses by Chinese across initial pitch when they had to choose between fall and mid, in comparison with Bissiri & Zellers' [22] Swedish, German and English participants. German and English listeners had significantly more fall responses when initial pitch was rise, while Swedish listeners, similarly to the Chinese listeners in the present study, did not.

3. Discussion

As with the English, German and Swedish listeners in Bissiri & Zellers' [22] analysis, the Mandarin Chinese listeners in the current study were most likely to associate glottalization with falling pitch rather than mid or rising pitch. However, to an even greater extent than Bissiri & Zellers' listeners, the Mandarin listeners were flexible in their association of glottalization with different pitch contours, with around 30% of responses preferring a mid or rising contour even when a falling contour was also presented. Furthermore, there was no statistically significant effect of the initial pitch contour on listeners' choice of final pitch contour to match the glottalization.

The lack of effect of initial pitch contour on Mandarin listeners' judgments seems most likely to be the result of the Mandarin tone system, in which every syllable is assigned a lexical tone. In the case of the current stimuli, there was no reason for the Mandarin listeners to make predictions about the pitch of the final syllable on the basis of the first two unless they adopted a strategy of assigning lexical meanings that could constrain the last syllable. Six of the Chinese participants reported after the experiment that the stimuli sounded similar to their native language. However, the stimuli did not constitute existing words in Chinese, and we can therefore reasonably exclude the possibility that a mapping of some stimuli to real Chinese words introduced a bias in the responses.

It was possible that an influence of tone sandhi from the preceding pitch contour could constrain perception of the pitch of the glottalized syllable, but this does not appear to have been the case in the current data. There are several possible reasons for this. The first is of course that the pitch percept in the glottalized stretches was too strong to allow for alternative interpretations. However, the substantial minority of non-falling responses suggests that this is not the case, since if the pitch percept was extremely strong, it seems unlikely that those non-falling responses would be possible. Therefore, we must look elsewhere for an explanation.

One possibility is that glottalization may be so firmly associated with Tones 3 and 4 that listeners' interpretation was constrained by the glottalization even if the pitch contours did not match the canonical form. The higher percentage of rise responses by the Chinese compared to Bissiri & Zellers' [22] German, English and Swedish listeners might be explained by an association of glottalization with the falling-rising Tone 3. Liu & Samuel [24] report that Tone 3 is perceived by Mandarin Chinese natives with the same accuracy even if its rising portion is missing. They also found that listeners could still perceive Tone 3 most of the time even if it consisted only of its rising portion, which is surprising given its redundancy for Tone 3 perception. These results indicate a) a possible flexibility in tone perception by Mandarin Chinese listeners allowing them to associate glottalization with rising pitch, if they relate both acoustic cues to Tone 3, and b) a possible relevance of other secondary cues beside F0 for tone perception. Liu & Samuel [24] propose duration and amplitude

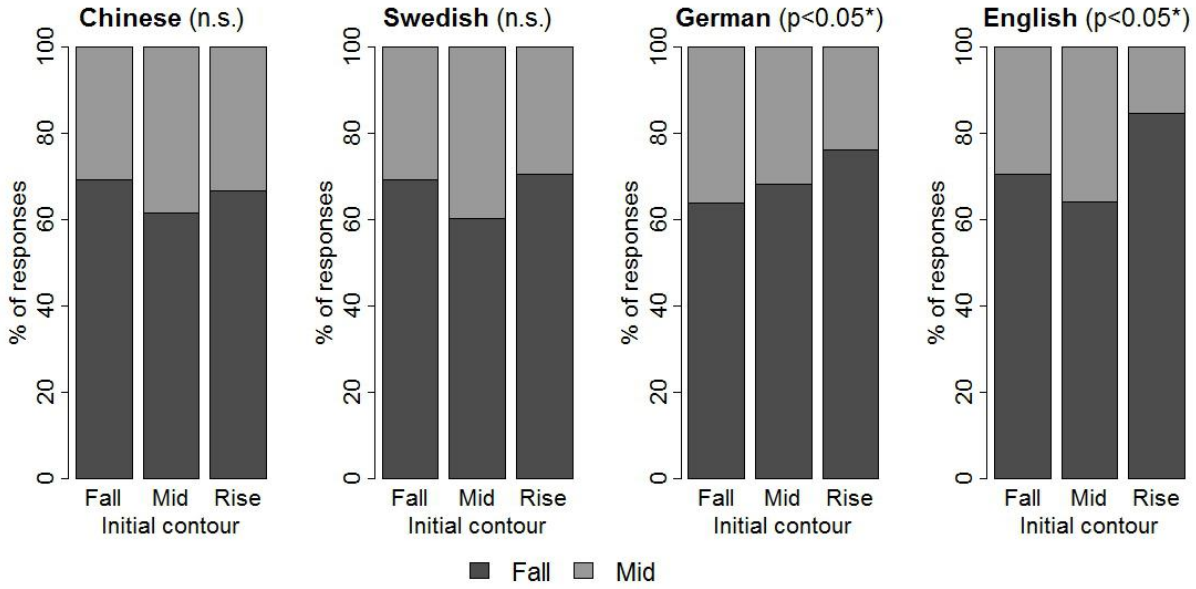


Figure 4: Responses by Chinese listeners and by Bissiri & Zellers'[22] Swedish, German and English listeners across initial pitch when choosing between fall and mid contours (Pearson's χ^2 test).

as possible candidates, and Lee et al. [25] also suggest voice quality or glottalization as a relevant perceptual cue for Mandarin tones.

Another alternative for the interpretation of our results is that listeners did not treat the syllables as Chinese lexical items at all. Since, as mentioned above, some participants treated the items as coming from a nonsense language, they could have potentially interpreted the pitch according to a tone-based system without being completely influenced by the form of the Mandarin tone system.

Combined with Bissiri & Zellers' [22] results, a consistent picture of the perception of glottalization begins to emerge from the data. Specifically, the different functional load of prosodic features across languages leads to different perceptual patterns. In Mandarin Chinese, where pitch is essential to lexical meaning, and every syllable has its own tone assigned, listeners were apparently not influenced by the pitch of preceding syllables but only made judgments about the pitch of the glottalized syllable itself. In Swedish, where pitch also contributes to lexical meaning but glottalization is not so firmly associated with either tone, listeners apparently were strongly inclined to seek a pitch interpretation of the stimuli, with most responses matching glottalization to the lowest available pitch contour in the pair. In German and English, where pitch is assigned on the intonational rather than a pitch-accent or tonal level, larger-scale contours appear to have a stronger influence on listeners' judgments, with listeners apparently making many of their pitch judgments on the basis of prediction of how a pitch contour would continue, rather than on calculation of the pitch of the glottalized stretch alone. It is important to note, however, that the current task intentionally influences listeners towards interpreting glottalization as having pitch, or being related to pitch variation. In other words, the task may have biased listeners towards a categorization of glottalization within the tonal/intonational system of their native language. A task which depended on a function of glottalization not (necessarily) tied to pitch variation – e.g. turn transition, as in Ogden [12, 13] – might lead to a different result.

4. Conclusion

Our data, combined with Bissiri & Zellers' [22] findings, give growing evidence for a language-specific link between perception of pitch and perception of glottalization in long stretches of creaky voice. It appears that listeners' experience with tone or intonation languages can lead them to focus on different prosodic information when judging the pitch of glottalized stretches. In particular, the influence of the surrounding pitch context is different across different languages. Mandarin Chinese listeners, similarly to Bissiri & Zellers' Swedish listeners, were not influenced by the previous pitch contour while associating glottalized syllables to the possible pitch alternatives. Their performance contrasted with Bissiri & Zellers' German and English listeners, speakers of intonation languages, who were influenced by the previous pitch contour in their responses. Creak may lead Mandarin listeners to a falling-rising Tone 3 interpretation as well as a falling Tone 4 one, while the one-syllable size of the tone domain apparently influences them to ignore the preceding pitch context when making their judgments. Thus the different sizes of pitch domains relevant in each language, and even the internal phonological structure itself, appear to influence listeners' expectations about the relevant interpretive context for glottalization.

5. Acknowledgements

The authors are grateful to Xinping Xu for conducting the experiment, and to all of the Chinese participants. This research was supported by the European Union grant MC-IEF GeCzEnGlott, by the postdoctoral research grant 'Perception of prosody in linguistic contexts' (VR-435-2011-6871) from the Swedish Research Council (Vetenskapsrådet) and by the National Social Science Foundation of China (13BYY009).

6. References

[1] Gordon, M. & P. Ladefoged (2001). Phonation types: a cross-linguistic overview. Journalof Phonetics 29: 383-406.

[2] Gerratt, B.R. & J. Kreiman (2001). Toward a taxonomy of nonmodal phonation. Journal of Phonetics 29: 365-381.

[3] Catford, J.C. (1964). Phonation types: the classification of some laryngeal components of speech production. In: Abercrombie, D. et al. (eds.) In honour of Daniel Jones, London: Longmans, pp. 26-37.

[4] Blomgren, M., Y. Chen, M.L. Ng, & H.R. Gilbert (1998). Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. Journal of the Acoustical Society of America 103(5): 2649-2658.

[5] Gussenhoven, C. (2004). The phonology of tone and intonation. Cambridge: Cambridge University Press.

[6] Pierrehumbert, J. & D. Talkin (1992). Lenition of /h/ and glottal stop. In Papers in Laboratory Phonology II. Cambridge: Cambridge University Press, 90-117.

[7] Pierrehumbert, J. (1995). Prosodic effects on glottal allophones. In: Fujimura, O., Hirano, M. (eds.), Vocal fold physiology: voice quality control. Singular Publishing Group, San Diego, pp. 39-60.

[8] Dilley, L., S. Shattuck-Hufnagel, & M. Ostendorf (1996). Glottalization of word-initial vowels as a function of prosodic structure. Journal of Phonetics 24: 423-444.

[9] Redi, L. & S. Shattuck-Hufnagel (2001). Variation in the realization of glottalization in normal speakers. Journal of Phonetics 29: 407-429.

[10] Henton, C. & A. Bladon (1988). Creak as a socio-phonetic marker. In Hyman, L.M. & C.N. Li (eds.) Language, Speech and Mind: studies in honor of Victoria A. Fromkin. London, pp. 3-29.

[11] Huffman, M.K. (2005). Segmental and prosodic effects on coda glottalization. Journal of Phonetics 33: 335-362.

[12] Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. Journal of the International Phonetic Association 31: 139-152.

[13] Ogden, R. (2004). Non-modal voice quality and turn-taking in Finnish. In Couper-Kuhlen, E & Ford, C. (eds.) Sound patterns in interaction: cross-linguistic studies from conversation. Amsterdam: John Benjamins, pp. 29-62.

[14] Bissiri, M. P., M.L. Lecumberri, M. Cooke & J. Volín, (2011). The role of word-initial glottal stops in recognizing English words. Proceedings of Interspeech 2011, Florence, Italy, pp. 165-168.

[15] Kohler, K. J. (1994). Glottal stops and glottalization in German. Phonetica 51: 38-51.

[16] Ding, H., O. Jokisch & R. Hoffmann (2004). Glottalization in inventory construction: a cross-language study. Proceedings of ISCSLP 2004, Hong Kong, pp. 37-40.

[17] Chao, Y.R. (1968). A Grammar of Spoken Chinese. Berkeley, University of California Press.

[18] Ding, H. & J. Helbig (1996). Sprecher- und kontextbedingte Varianz des dritten Vokaltones in chinesischen Silben – eine akustische Untersuchung. Proceedings of DAGA 1996, Bonn, Germany, pp. 514-515.

[19] Silverman, D. (1997). Laryngeal Complexity in Otomanguean Vowels. Phonology 14: 235-261.

[20] Frazier, M. (2008). The interaction of pitch and creaky voice: data from Yucatec Maya and cross-linguistic implications. UBC Working Papers in Linguistics: Proceedings of Workshop on Structure and Constituency in the Languages of the Americas (WSCLA), pp. 112-125.

[21] Ní Chasaide, A. & C. Gobl (2004). Voice quality and f0 in prosody: towards a holistic account. Proceedings of the 2nd International Conference on Speech Prosody, Nara, Japan, pp. 189-196.

[22] Bissiri, M.P. & M. Zellers (2013). Perception of glottalization in varying pitch contexts across languages. Proceedings of Interspeech 2013, Lyon, France, pp. 253-257.

[23] Boersma, P. & D. Weenink (2013). Praat: doing phonetics by computer [Computer program]. Available http://www.praat.org/.

[24] Liu, S. & A.G. Samuel (2004). Perception of Mandarin lexical tones when F0 information is neutralized. Language and Speech 47(2): 109-138.

[25] Lee, C.-Y., L. Tao & Z.S. Bond (2008). Identification of acoustically modified Mandarin tones by native listeners. Journal of Phonetics 36: 537-563.

Cross-linguistic perception of Mandarin intonation

Robert Bo Xu, Peggy Pik Ki Mok

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
xuborobert@gmail.com, peggymok@cuhk.edu.hk

Abstract

This study investigated how phonological knowledge and psychoacoustic mechanism interact in intonation perception. In the experiment, Mandarin and Cantonese listeners identified Mandarin statement and question in both unfiltered and low-pass filtered contexts. The results show that the importance of different perceptual factors varies depending on the perception materials. Language background plays an important role even in processing low-level psychoacoustic materials.

Index Terms: cross-linguistic perception, intonation, Mandarin, the Frequency Code

1. Introduction

1.1. Intonation and the Frequency Code

This study investigates the interaction between psychoacoustic mechanism and phonological knowledge in intonation perception. The most relevant psychoacoustic mechanism to the perception of statement and question is the Frequency Code proposed by Ohala [1]. It explains the cross-linguistic correspondence between intonation contour and (para-)linguistic meaning, stemming from a biological basis [2]. Since the organ that produces a lower sound is usually larger, low-pitched individuals are associated with being dominant and aggressive, whereas individuals that produce a higher pitch are associated with being subordinate and submissive. As a result, the informational interpretations of the Frequency Code relate high or rising contour to "uncertainty", and thus questioning, and relate low or falling contour to "certainty", and thus being assertive, i.e., statements [1], [2].

1.2. Tone and intonation of Mandarin

1.2.1. Lexical tones in Mandarin

There are four lexical tones in Mandarin [3] (Table 1), all differ in pitch shape. There is a tone sandhi rule in Mandarin. In a T3-T3 sequence, the first T3 is produced as a high rising tone [4], perceptually indistinguishable from T2 [5] [6].

Table 1. *Lexical tones in Mandarin.*

	T1	T2	T3	T4
Tone Shape	Level	Rising	Dipping	Falling
Tone Letter	55	35	21/214	51

1.2.2. Production of Mandarin question intonation

While some studies suggested that questioning in Mandarin could be a local event at the final tone [5]–[8], other studies showed that the tone shape of the final tone remains very similar to the citation form acoustically [11]-[13]. Many studies have demonstrated that Mandarin questions were signaled by modification of the F0 contour of the whole utterance [14]-[21]. Yuan [13] showed that the global F0 of a question was raised and the whole sentence had a smaller declination slope compared with that of a statement. The final T2 of a question had a steeper rising and a higher ending, and the final T4 of a question had a higher beginning and a higher ending than those of a statement, and hence had a flattened falling contour. He concluded that a boundary tone is not necessary to model the distinction between statement and question in Mandarin. The essential mechanism of Mandarin questioning is its raised F0 contour over the scope of the whole utterance.

1.2.3. Perception of Mandarin question intonation

Studies on intonation perception of Mandarin revealed several asymmetries in their perception results. First of all, statements are easier to identify than questions. This bias towards statement shows that statement would be an unmarked sentence type [13], [22]. Second, the sentence-final lexical tone does not affect the identification of statements, but influences the identification of intonational questions (yes-no questions that are signaled by intonation only, without any sentence final particles). Questions with a sentence-final falling tone (T4) are easiest to be identified, and a question ending with a final rising tone (T2) is the most difficult to identify [9]. Yuan [13],[23] explained that this was due to the flattening of the falling tone at the end of a question. This is a language/tone specific perceptual pattern; and intonation perception was sensitive to the phonological tone identity at the end of an utterance. Third, question-final lexical tones are rarely affected by intonation perceptually. Each tone in statements or in questions can be easily recognized by listeners [23]. Jiang and Chen [24] showed that cutting off the final tone does not significantly influence perception. What is important to questioning is the last prosodic word of the utterance.

1.3. Intonation of Cantonese

This study includes Cantonese listeners because Cantonese intonation patterns are different from Mandarin. Cantonese yes-no questions are signaled by a final rising that change the canonical tone shape of any tone at the end of the questions [25]-[27]. In addition, global F0 contour does not play an important part in questioning [27], [28]. As a result, Cantonese listeners may have different interpretation of Mandarin intonation patterns from native listeners.

1.4. Cross-linguistic perception of prosodic features

An important theme in cross-linguistic perception is how phonetic forms are shaped in the perception with different language backgrounds [29], [30]. Huang and Johnson [31] showed that language experience was impactful in the perception of tonal contrasts even in low-level psychoacoustic processing. Burnham et al. [32] showed that tonal contrasts in musical sounds were better discriminated than in low-pass filtered speech, which in turn was better discriminated than in

normal speech, indicating that listeners were more sensitive to acoustic differences in non-speech signal than in speech.

While some studies showed that language background played an important role in perceiving intonation [33], [34], Gussenhoven and Chen [35] showed that listeners with different native languages associated question intonation with a later or a higher F0 peak and higher end pitch in a made-up speech, concurring with the Frequency Code that a high or rising pitch contour is associated with questioning. Grabe et al. [29] suggested that experience with a native language was added to the universal auditory mechanism in shaping listeners' perception of intonation.

1.5. Summary

Previous studies have shown that the Frequency Code plays an important part in intonation perception [35], but intonation perception is also shaped by listeners' first language intonation phonology[33], [34]. However, no previous studies have systematically demonstrated the interactions between these factors. This study aims to fill in this research gap.

2. Method

2.1. Materials

Nine-syllable sentences in Mandarin shown in Table 2 were designed for the experiment. The final two syllables of each sentence share the same lexical tone. With the final syllable cut off, the utterances still remained meaningful, and the ending tone remained the same (with the exception of T3 because of unavoidable tone sandhi).

Table 2. *Mandarin utterances used in the experiment*

Finals	Sentences (Chinese, pinyin and English translation)
T1	妈妈今晚炖的是鸡(汤)。 ma1 ma jin1 wan3 dun4 de shi4 ji1 (tang1). 'Mommy cooked chicken (soup) for tonight'
T2	亚马逊是最长的河(流)。 ya3 ma3 xun4 shi4 zui4 chang2 de he2 (liu2). 'Amazon is the longest river.'
T3	他最大的缺点是懒(散)。 ta1 zui4 da4 de que1 dian3 shi4 lan3 (san3). 'His biggest shortcoming is laziness.'
T4	工人在修公园的路(面)。 gong1 ren2 zai4 xiu1 gong1 yuan2 de lu4 (mian4). 'The workers are repairing the road in the park.'

Two native speakers, both professional Mandarin teachers, (male aged 30 from Shaanxi, female aged 26 from Hubei) were recorded reading the sentences for the experiment. The recording took place in a sound-treated room. The speakers were instructed to read the sentences focus-neutrally.

After screening the naturalness of the recordings, four presentation sentence conditions (complete statement and question, statement and question with the final syllable cut-off) were prepared for the perception test for two experiment contexts (normal speech and low-pass filtered speech). All cutting points were at zero crossing following the offset of the penultimate syllables of the utterances. The average amplitude of all the utterances was also normalized. Low-pass filters were applied with 100 Hz bandsmoothing. The cutoff frequency was determined by the highest pitch each speaker produced in their T2 questions, being 250 and 400 Hz for the male and female speakers respectively. Informal tests showed that native speakers of both languages could not understand the content of the low-pass filtered sentences They reported only hearing some low frequency humming.

2.2. Listeners and procedure

20 Cantonese and 20 Mandarin listeners participated in the perception experiment. All the Cantonese listeners were native Hong Kong Cantonese speakers, speaking Mandarin with varying proficiency. All the Mandarin listeners came from Northern China. They all spoke English as a foreign language. None of them have received any systematic training in linguistics when they participated in the experiment. None had a reported history of speech or hearing disorder.

The perception experiment was carried out in a sound-attenuated room. The materials were presented to them using the software E-Prime. After a practice session, the listeners finished the section with low-pass filtered speech before the section with normal speech, to prevent them from guessing the content of the filtered materials. In the filtered section, two blocks of filtered Cantonese utterances were used as fillers to conceal the identity of language the listeners heard. The blocks and the tokens within each block were randomized. Each of the 64 stimuli (4 sentence conditions × 2 contexts × 2 genders × 4 tones) was repeated twice. Listeners listened to each trial only once before they decided whether the utterance they just heard was a statement or a question by pressing a button on the keyboard.

2.3. Data analysis

Average identification accuracy (IA) were calculated. Three-way (Intonation type × Condition × Final tone) repeated measures ANOVAs were conducted on the identification scores. Corrections for violations of sphericity were made, where appropriate, using the Greenhouse–Geisser estimates of sphericity. Bonferroni correction was applied for pairwise comparisons.

3. Results

3.1. Mandarin listeners: Normal speech

Figure 1 shows the identification accuracy for Mandarin listeners listening to normal utterances. The main effects of Intonation type [$F(1,19)=23.02$, $p<0.001$], Condition [$F(1,19)=148.52$, $p<0.001$] and Final tone [$F(3,57)=18.65$, $p<0.001$] are all significant. In the complete utterance condition, statements were well identified across different tones. T3 ($M=78\%$, $SD=0.26$) complete question has the lowest IA, significantly lower than T2 and T4 ($p<0.05$ in each comparison). In the cut-off condition, while T1, T2 and T4 statements remain well identified, T3 statement has the lowest IA ($M=74\%$, $SD=0.29$), probably due to the penultimate high rising pitch contour induced by tone sandhi. The IA of T3 cut-off statement is significantly lower than T2 and T4 statements ($p<0.05$ in each comparison). Among the cut-off questions, T4 question remains almost unaffected by the elimination of the last syllable, while the IA for T1 and T2 questions become much lower than their complete counterparts. The identification accuracy for T4 is significantly higher than all the other tones ($p<0.05$ in every comparison). The results suggest that T3 is a difficult tone to identity in Mandarin

questions in the compete condition. In contrast, T4 may be the easiest ending tone for Mandarin questions. Cutting off the final syllable influences the perception of intonation, but the exact influence depends on different tones.

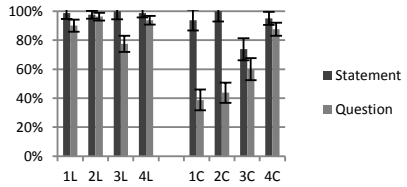


Figure 1: *IA of normal utterances by Mandarin listeners. (L: complete utterances; C: cut-off utterances)*

3.2. Cantonese listeners: Normal speech

Figure 2 shows that Cantonese listeners had lower IA than native Mandarin listeners, especially for questions. The main effect of Intonation type [$F(1,19)=71.31$, $p<0.05$], Condition [$F(1,19)=55.50$, $p<0.05$] and Final tones [$F(3,57)=13.10$, $p<0.05$] are all significant. For the complete utterances, listeners performed well for statements with all ending tones. This is not the case for questions, as T3 complete question is poorly identified (M=43%, SD=0.20, below chance level). The IA is significantly lower than complete questions with the other tones ($p<0.05$ in every comparison). For the cut-off statements, T3 statements is the worst identified (M=71%, SD=0.34), with IA significantly lower than the other tones ($p<0.05$ in every comparison). For the cut-off questions, T1 and T2 questions (both identified below chance level) are identified significantly more poorly than T3 and T4 questions ($p>0.05$ in every comparison). There are no significant differences between the IAs of T1 and T2 ($p>0.05$) or between those of T3 and T4 ($p>0.05$). The result shows that Cantonese listeners also found T3 to be a difficult ending tone. They also found questions ending with T1 and T2 in the cut-off condition confusing.

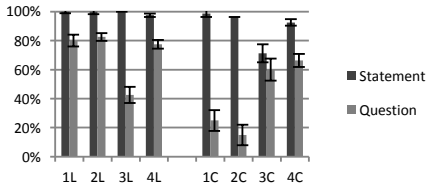


Figure 2: *IA of normal utterances by Cantonese listeners.*

3.3. Mandarin listeners: Filtered speech

Figure 3 shows the identification accuracy for Mandarin listeners listening to low-pass filtered utterances. The main effect of Intonation type [$F(1,19)=57.39$, $p<0.001$], Condition [$F(1,19)=173.57$, $p<0.001$] are significant. Although the main effect of Final tone is not significant [$F(3,57)=1.73$, $p>0.05$], the significant interactions Final tone × Intonation type [$F(2.29,43.59)=15.94$, $p<0.001$] and Final tone × Condition [$F(3,57)=6.74$, $p<0.01$] suggest that tones affect listeners' identification depending on the intonation type and condition. For example, complete question ending with T3 (M=76%, SD=0.23) is the least well identified by listeners among the complete questions, significantly poorer than T1 (M=93%, SD=0.14, $p<0.05$). Among the cut-off utterances, T3 statement is significantly more poorly identified than statements in all other tones, whereas T3 question is significantly better identified than the other tones ($p<0.05$ in all comparisons), both owing to the rising penultimate tone due to tone sandhi. Besides, within the cut-off condition, T4 question (M=51%, SD=0.37) is significantly better identified than T1 question (M=25%, SD=0.26, $p<0.05$), suggesting that for Mandarin listeners, questions ending with T4 are still easier to identify than some other tones, even though the original final tone is cut off.

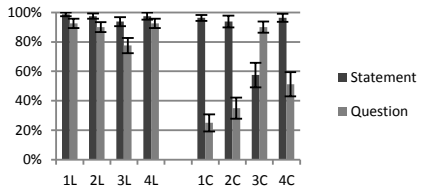


Figure 3: *IA of filtered utterances by Mandarin listeners.*

3.4. Cantonese listeners: Filtered speech

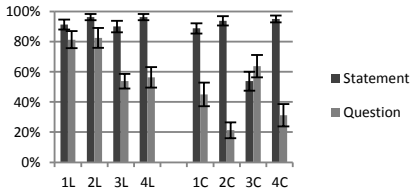


Figure 4: *IA of filtered utterances by Cantonese listeners.*

Figure 4 shows the patterns of Cantonese listeners' identification accuracy for filtered Mandarin utterances. The main effect of Intonation type [$F(1,19)=40.69$, $p<0.001$], Condition [$F(1,19)=69.82$, $p<0.001$], and Final tone [$F(3,57)=5.28$, $p<0.005$] are all significant. In the complete condition, the identification accuracy of T3 (M=54%, SD=0.22) question is significantly lower than T1 (M=81%, SD=0.25) and T2 (M=83%, SD=0.29) questions ($p<0.05$ in each comparison). In the cut-off condition, the T3 utterances end with the sandhi-ed T3 with a rising contour. As a result, the T3 cut-off question (M=64%, SD=0.33) was better identified than the other tones (comparisons with T2 (M=21%, SD=0.23) and T4 (M=31%, SD=0.33) questions were significant, $p<0.05$). The T3 cut-off statement (M=54%, SD=0.28) was identified most poorly among all the tones (significant compared with every other tone, $p<0.05$). Furthermore, unlike the patterns of Mandarin listeners, the T4 complete question (M=56%, SD=0.30) was poorly identified by Cantonese listeners, with the accuracy significantly lower than that of T1 and T2 questions ($p<0.05$ in each comparison). The identification of T2 cut-off question (M=21%, SD=0.23) was also not successful, significantly poorer than T1 and T3 ($p<0.05$ in each comparison). The perceptual patterns suggest that Cantonese speakers mainly use the final tone in their judgment, much like the way they identify intonation type in their own language. Final high level and high rising pitch contours (in the case of complete T1, T2 utterances and cut-off

sandhi-ed T3 utterances) seem to be a strong indicator for question, while a dipping tone and a falling tone (complete T3 and T4 utterances) were associated with statements.

4. Discussion

This study examined the perception of three kinds of materials: native speech as a baseline, non-native speech, and filtered-speech. In general, the listeners from both groups performed well in the complete utterance condition. There is no significant difference in IA among different tones in complete utterances, except that the IA of T3 complete question was significantly lower than the other tones. On the other hand, eliminating the last syllable affects listeners' judgment of questions, but not so much of that of statements. With the exception of sandhi-ed T3, the IA of the cut-off questions is lower than that of complete questions across listener groups and experiment contexts.

Our data show that perception of intonation is influenced by phonological knowledge. For example, Mandarin final T4 benefits question identification because native listeners are sensitive to the identity of final tone in questions [13] (Cantonese listeners also have experiences in speaking Mandarin). This seems to be the case in the normal speech context. In the cut-off condition, T4 question was hardly influenced by the loss of the final tone, and was significantly better identified than questions with all the other tones. However, when the identity of language and lexical tone were concealed from the listeners by the low-pass filter, listeners were unable to use the knowledge of tone-intonation interaction to identify questions. Therefore, the IA of T4 questions decreased for both groups of listeners, especially in the cut-off condition, suggesting that the absence of phonological knowledge affects the identification of questions.

When perceiving non-native speech, Cantonese listeners' identification of Mandarin intonation was facilitated by different tools that interacted with each other. In most cases, they used the phonological knowledge of intonation in Cantonese by locating the perceptual cue for question at the end of the utterances, as this is where the cue for their native language is placed. Then they made use of the Frequency Code to associate a high rising tone with questions and a low tone with statements. For example, they performed well in T3 cut-off questions but performed poorly for T3 complete questions. Finally, the experience in knowing Mandarin also helped them to identify questions with T4.

In the low-pass filtered speech, segmental information was eliminated and F0 became the main resource of information left in the stimuli. Therefore, the use of native phonological knowledge was restricted and the Frequency Code might have become the most, if not the only, available and useful tool for intonation perception. The listeners performed generally quite well in perceiving the low-pass filtered speech, indicating that F0 provided most information for intonation perception. The role of the Frequency Code can be observed from the results. On the one hand, the canonical T3 (low tone) is associated with a statement. On the other hand, the cut-off T3 statement ending with a rising tone was identified significantly less successfully than the other tones; and the cut-off T3 question was identified more accurately than the other tones, especially in the low-pass filtered speech, where the Frequency Code dominated the process of the perception.

To sum up, listeners process different materials differently. When listening to native speech, they primarily use their phonological knowledge of the language. For non-native speech, both the Frequency Code and some phonological knowledge were used. For filtered-speech (or non-speech), the Frequency Code becomes the most useful tool. In other words, the biologically-based universal Frequency Code lays the foundation for perception, while phonological knowledge comes into play when native speech materials are used. When phonological knowledge and the Frequency Code contradicted with each other, the former takes advantage in the perception of native language.

However, although the Frequency Code is shown to have dominated the perception of intonation in the filtered context, the perceptual patterns of the two groups of listeners were not identical. For example, the IA of T1 cut-off question is not significantly different from T3 question for Cantonese listeners, but is significantly lower than T3 questions for Mandarin listeners. T4 complete question was well identified by Mandarin listeners but not so by Cantonese listeners. These differences stem from the different language backgrounds the two groups of listeners have, suggesting that language experience shapes prosodic perception even in the low-level processing of unintelligible filtered speech. This finding echoes previous study in cross-linguistic perception of lexical tone by Huang and Johnson [31].

Finally, the results further the insights into question-statement bias previously observed in individual languages [13]. For both groups of listeners, the IA of questions was constantly lower than that of statements, especially in cut-off utterances. The fact that both groups of listeners showed the same preference towards statements suggests that the bias towards statement being an unmarked sentence type may be universal.

5. Conclusions

In conclusion, the results demonstrate that two interactions in intonation perception. One of them is the interaction between tone and intonation: no matter how intonation modifies F0 contour, at the final tone or on a sentential level, listeners are sensitive to the identity of lexical tone at the end of the utterance. Therefore, the processing of tone and intonation in tone languages is interdependent. More importantly, phonological knowledge and the Frequency Code co-direct listeners' perception of intonation, especially when listeners are given a less familiar language or filtered speech. Listeners follow the Frequency Code when phonological knowledge is not applicable. When the linguistic resources are rich in the speech signal, and when phonological knowledge and the Frequency Code conflict with each other, phonological knowledge would override the Frequency Code. As a result, listeners could perceive intonation patterns that seemingly contradict the Frequency Code in their native/familiar language. Furthermore, language background is an important and robust factor in intonation perception even with filtered speech materials, as different patterns arose from the two groups of listeners because of their language background. The results also confirm that statements are generally identified better than questions. This tendency applies to intonation perception cross-linguistically.

6. References

[1] Ohala, J. J. “Cross-language use of pitch: an ethological view”, *Phonetica*, 40, 1–18, 1983.

[2] Gussenhoven, C. *The Phonology of Tone and Intonation*, Cambridge University Press, 2004.

[3] Chao, Y. R. “Tone and intonation in Chinese”, *Bull. Inst. Hist. Philos*., 4, 121–134, 1933.

[4] Shih, C. “The Prosodic Domain of Tone Sandhi in Chinese”, Ph.D dissertation, University of California, San Diego, 1986.

[5] Wang, W. S.-Y., and Li, K. “Tone 3 in Pekinese”, *J. Speech Hear. Res*., 10(3), 629–636, 1967.

[6] Peng, S. “Lexical versus ‘phonological’ representations of Mandarin sandhi tones”, in M. B. Broe and J. Pierrehumbert [ED], *Acquisition and the lexicon: Papers in Laboratory Phonology V*, 152–167, Cambridge: Cambridge University Press, 2000.

[7] Ho, A. T. “Intonation variation in a Mandarin sentence for three expressions interrogative, exclamatory and declarative”, *Phonetica*, vol. 34, pp. 446–457, 1977.

[8] Wu, Z. “Variation of lexical tones in Mandarin intonation”, *Studies of The Chinese Language*, vol. 6, pp. 439–449, 1982.

[9] Lin, M. “On Production and Perception of Boundary Tone in Chinese Intonation”, in *Proceedings of the International Symposium on Tonal Aspects of Languages*, 2004.

[10] Peng, S., Chan, M. K. M., Tseng, C., Huang, T., Lee, O. J., and Beckman, M. E. “Towards a Pan-Mandarin system for prosodic transcription”, in S.-A. Jun [ED], *The Phonology of Intonation and Phrasing*, 1-63. 2006.

[11] He, Y. and Jin, S. “Experimental investigation of intonation in Beijing Mandarin.” *Language Teaching and Research*, 2, 71–96, 1992.

[12] Yuan, J., Shih, C. and Kochanski, G. P. “Comparison of declarative and interrogative intonation in Chinese”, in *Proceedings of Speech Prosody*, 711–714, 2002.

[13] Yuan, J. “Intonation in Mandarin Chinese: Acoustics, perception, and computational modeling”, PhD Dissertation, Cornell University, 2004.

[14] Shi, P. “Intonation variations in four types of Mandarin sentences”, *Language Teaching and Research*, 2, 71–81, 1980.

[15] Gårding, E. “Chinese and Swedish in a Generative Model of Intonation”, in C.-C. Elert and I. Johansson, [ED], *Nordic Prosody III*, Stockholm: University of Umeå, 79– 91, 1984.

[16] Gårding, E. “Speech act and tonal pattern in Standard Chinese: constancy and variation”, *Phonetica* 44 13-29, vol. 44, pp. 13–29, 1987.

[17] Shen, J. “The tone range and intonation of Beijing Mandarin”, in T. Lin and L. Wang, [ED], *Exprimental Phonetic Studies of Beijing Mandarin*, 73–130, 1985.

[18] Shen, J. “Construction and types of Chinese intonation”, *Dialect*, 3, 221–228, 1994.

[19] Shen, X. S, *The prosody of Mandarin Chinese*, University of California Press, 1990.

[20] Yuan, J. “Mechanisms of question intonation in Mandarin”, in, Q. Huo, B. Q. Ma, E.-S. Cheng, and H. Li [ED], *Chinese Spoken Language Processing*, Springer, 19–30, 2006.

[21] Lee, O. J. “The prosody of questions in Beijing Mandarin”, PhD dissertation, The Ohio State University, 2005.

[22] Yuan, J. “Perception of intonation in Mandarin Chinese”, *J. Acoust. Soc. Am.*, 130(6), 4063–4069, 2011.

[23] Connell, B. A., Hogan, J. T., and Rozsypal, A. J., “Experimental evidence of interaction between tone and intonation in Mandarin Chinese”, *J. Phon.*, 11, 337–351, 1983.

[24] Jiang, P., and Chen, A., “Representation of Mandarin intonations: Boundary tone revisited”, in *Proceedings of the North American Conference on Chinese Linguistics*, 1, 97–109, 2011.

[25] Ma, J. K.-Y., Ciocca, V., and Whitehill, T. L., “Effect of intonation on Cantonese lexical tones”, *J. Acoust. Soc. Am.*, 120(6), 3978–3987, 2006.

[26] Gu, W., Hirose, K., and Fujisaki, H., “Modeling the Effects of Emphasis and Question on Fundamental Frequency Contours of Cantonese Utterances”, *IEEE Trans. audio, speech Lang. Process.*, 14(4), 1155–1170, 2006.

[27] Lee, H.-Y. D., “The Intonation of Cantonese Declarative Questions”, MPhil dissertation, Cambridge, 2008.

[28] Xu, B. R., and Mok, P., “Final rising and global raising in Cantonese intonation”, in *Proceedings of International Congress of Phonetic Sciences*, 2173–2176, 2011.

[29] Grabe, E., Rosner, B. S., and Garcia-Albea J. E., “Perception of English Intonation by English, Spanish, and Chinese Listeners”, *Lang. Speech*, 46(4), 375–401, 2003.

[30] Vaissiere, J., “Perception of intonation”, in D. B. Pisoni and R. E. Remez [ED], *The Handbook of Speech Perception*, Blackwell Publishing, 236–263, 2005.

[31] Huang, T., Johnson, K., “Language specificity in speech perception: perception of Mandarin tones by native and nonnative listeners”, *Phonetica*, 67(4), 243–67, 2010.

[32] Burnham D., Francis, E., Di, W., Sudaporn L., Chayada A., Lacerda, L. F. and Peter K., “Perception of lexical tone across languages: Evidence for alinguistic mode of processing”, in *Proceedings of the International Conference on Spoken Language processing*, 2514–2517,1996.

[33] Handding-Koch, K. and Studdert-Kennedy, M., “An experimental study of some intonation contour”, *Phonetica*, 11, 175–185, 1964.

[34] Jiang, P. and A. Chen, “Tonal effect on Mandarin intonation perception : A comparative study of listeners with different L1 backgrounds Research questions”, in *International Symposium on Bilingualism and Comparative Linguistics*, 2012.

[35] Gussenhoven, C. and Chen A., “Universal and language-specific effects in the perception of question intonation”, in *Proceedings of International Conference on Spoken Language Processing*, 2000.

[36] Grabe, E., Rosner, B. S., and Garcia-Albea, J. E., “Perception of English Intonation by English, Spanish, and Chinese Listeners”, *Lang. Speech*, 46(4), 375–401, 2003.

Segmental and prosodic cues to vowel identification: The case of /ɪ i iː/ and /ʊ u uː/ in Saterland Frisian

Wilbert Heeringa[1], *Jörg Peters*[1], *Heike Schoormann*[1]

[1]Institute of German Studies, Carl von Ossietzky University, Oldenburg, Germany

wilbert.heeringa@uni-oldenburg.de, joerg.peters@uni-oldenburg.de, heike.schoormann@uni-oldenburg.de

Abstract

Saterland Frisian has a complete set of closed short tense vowels. Together with the long tense vowels and the short lax vowels they constitute series of phonemes that differ by length and/or tenseness. We examined the cues that distinguish the front unrounded and the back rounded series of short lax and short and long tense vowels in triplets by eliciting 'normal speech' and 'clear speech' in a reading task from two speakers. Short and long vowels were distinguished by vowel duration, and lax and tense vowels by their location in the F1-F2 space. The durational difference between short tense and long tense vowels, however, was largely restricted to the 'clear speech' condition. In 'clear speech', f0 dynamics and centralization in the F1-F2 space were used as additional means to make short tense vowels more distinct from long tense vowels. These results suggest that length and tenseness are used as distinctive features, while f0 dynamics and centralization in the F1-F2 space were optionally used to enhance the contrast between short and long tense vowels.

Index Terms: f0 dynamics, f0 excursion, formants, Saterland Frisian, tenseness, vowel duration

1. Introduction

Saterland Frisian is spoken in three small villages – Strücklingen, Ramsloh and Scharrel – in the north-western corner of the district of Cloppenburg in Lower Saxony. It is the only remaining living variety of Old East Frisian, which was spoken along the coasts of the Netherlands and Lower Saxony. Saterland is believed to have been colonised by Frisians from the coastal areas in the eleventh century. According to the most recent count, Saterland Frisian is spoken by 2250 speakers [1].

Saterland Frisian has a complete set of closed short tense vowels: /i y u/ [2], [3], [4]. Together with the short lax vowels /ɪ ʏ ʊ/ and the long tense vowels /iː yː uː/ they constitute series of phonemes that differ by length and/or tenseness. Potential acoustic cues which distinguish the vowels in a triplet may be vowel duration, spectral features (F1, F2) as well as the timing and scaling of f0.

In this paper we investigate which acoustic cues distinguish the sounds within two triplets containing /ɪ i iː/ and /ʊ u uː/ respectively. For each of the two triplets we conducted a traditional reading task and a listener-directed task, which maximizes the discrimination between words and is an effective way to reveal potential segmental and prosodic cues.

Several studies show that vowels with stronger f0 dynamics are perceived as being longer. Lehiste [5] found that listeners perceived a falling-rising or rising-falling f0, as opposed to a flat f0 pattern, to be longer even when the stimuli have the same acoustic duration. Yu [6] found the same effect for dynamic versus flat f0 and also showed that syllables with higher f0 are heard as longer than syllables with lower f0. Cumming [7] was able to show the perceived lengthening effect of dynamic f0 for native speakers of Swiss German, Swiss French and French. This effect is likely language-specific [8].

The acoustic cues, which add to the distinction of vowels in Saterland Frisian triplets have not yet been fully studied. Siebs [9] distinguishes between tone accents in Saterland Frisian (*Stoßton* versus *Schleifton*) which suggests that f0 might play a role. In a more recent study Tröster-Mutz [10] [11] investigated the phonetics of Saterland Frisian vowels, but did not find any evidence for tone accent differences in present-day Saterland Frisian. In this paper, we focus on the question whether f0 dynamics have any systematic effect which may help to discriminate between the vowels of each triplet.

2. Method

2.1. Material

The two triplets used are still known by a restricted number of Saterland Frisian speakers. For the closed front vowels /ɪ i iː/ we used *Smitte* 'forge', *smiete* 'to throw' and *Smíete* 'throws' (pl.). The closed back vowels /ʊ u uː/ were elicited by the triplet *ful* 'full', *fuul* 'rotten' and *fúul* 'much'.

2.2. Procedure

For each of the triplets we conducted two experiments, one eliciting 'normal speech' and another eliciting 'clear speech'. The experiments were carried out by two female native speakers, aged 78 and 66 years, henceforth referred to as subject 1 and subject 2 respectively. The two speakers are born and raised in Ramsloh and have lived in this village most of their lives. We chose Ramsloh since it is located in the center of Saterland and its Saterland Frisian variety is considered to be the most conservative [3].

2.2.1. Normal speech

In this experiment Saterland Frisian words were presented in written form to the two native speakers on a computer screen, one word at a time. We used twelve different words: six triplet words (*ful*, *fuul*, *fúul*, *Smitte*, *smiete*, *Smíete*), and six filler words (*Pot*, *Paad*, *Kat*, *leet*, *Täk*, *Poot*).

A session consisted of four blocks in which each of the 12 words was presented four times. Within each block the order of the words was randomized, so that a word was never followed by the same word or by a word belonging to the same triplet. Three of the six filler words (*Pot*, *Paad* and *Kat*) were also used

as a short practice, preceding the first block. In sum, 195 words were presented in one session.

We obtained 16 samples per subject per triplet word. Looking for cues concerning the f0 dynamics, only samples with a clear f0 peak in the vowel can be used for the analysis. The number of word samples which satisfy this condition is given per triplet word and per subject in Table 1.

2.2.2. Clear speech

Saterland Frisian words were presented in written form to the two native speakers on a computer screen. In this condition, only the six triplet words were used. For maximum discrimination, a triplet word was always presented together with the two other triplet words. The word to be pronounced was encircled and displayed in blue (the other words were black). The three words of a triplet were located on the screen so that they were imaginary vertices of a triangle. Each 'triangle' was rotated over an arbitrary angle.

One session consisted of four blocks. Each of the triplet words was presented eight times per block. Within a block, 24 words of the /ʊ u uː/ triplet were presented first, followed by 24 words of the /ɪ i iː/ triplet. Thus, in each block 48 words were pronounced. In each part the words were presented in a randomized order so that a word was not followed by the same word.

In this experiment, the subjects were either speaker or listener and changed roles after each block. When one subject read the words aloud, the other subject marked the triplet word she thought she heard. The reader and the listener were separated by a screen during the experiment.

We obtained 16 samples per speaker per triplet word. Just as for 'normal speech' we limited the analysis to word samples with a clear f0 peak when looking for cues concerning the f0 dynamics. The number of word samples with f0 peak are given per triplet word and per speaker in Table 1.

Table 1: *Number of word samples with a clear f0 peak per triplet word and per subject.*

	Normal speech		Clear speech	
	sub1	sub2	sub1	sub2
Smitte	14	16	15	14
smiete	15	16	16	15
Smíete	13	16	16	16
ful	1	13	6	5
fuul	10	16	5	11
fúul	7	15	15	10

2.3. Acoustic variables

Segmental and prosodic variables were measured with PRAAT [12]. For each word belonging to the /ɪ i iː/ triplet we measured the duration of each of the segments: /s/, /m/, V, /t/ and the final schwa. The duration of /t/ was split into two parts: the time from the beginning of the segment to the burst (t1), and from the burst to the end of the segment (t2). We also measured spectral variables F1 and F2 at the center of the vowel.

We measured f0 variation in the interval from the beginning of /m/ to the end of V. When the f0 peak was somewhere in this interval, we considered a rise and a fall. F0 dynamics was operationalized by calculating the relative f0 excursion in semitones per millisecond, which was obtained by calculating the sum of the f0 rise size (pitch of f0 peak minus pitch at the beginning of the interval) and the f0 fall size (pitch of f0 peak minus pitch at the end of the interval) divided by the duration of the interval. A related approach was used by Grabe [13].

For each word belonging to the /ʊ u uː/ triplet we likewise measured the duration of each of the segments, /f/, V, and /l/, and F1 and F2 at the center of the vowel. When measuring relative f0 excursion, we focused on the interval starting at the beginning of V and ending at the end of /l/.

2.4. Statistical processing

We looked for acoustic cues that distinguish all of the sounds within a triplet. Per acoustic variable we used a Generalized Linear Model (GLM) where the stimulus was the independent variable and the acoustic variable the dependent variable. Since not all of the acoustic variables are normally distributed, GLM was the fitting test. The stimuli were the triplet words, therefore the stimulus variable was a three level factor. Two-tailed p-values obtained from pairwise comparisons were adjusted using the Bonferroni correction.

The variables were initially analyzed per subject and per triplet. However, in this paper the results of the subjects are combined by showing the consensus. For example, when we found $p < 0.01$ for the one subject, and $p < 0.001$ for the other subject, then the consensus is considered to be $p < 0.01$. When we do not find a significant effect for both of the subjects for a particular variable or a significant difference in opposite direction, no significance is reported. In the tables below, levels of significance are indicated by asterisks: $*< 0.05$, $**< 0.01$ and $***< 0.001$.

3. Results

3.1. Normal speech

3.1.1. Results /ɪ i iː/ triplet

Duration values for the /ɪ i iː/ triplet are shown in the upper panel of Figure 1. Vowel plots are found in Figure 2 and f0 contour plots in Figure 3. Statistical results are summarized in Table 2. *Smitte* is distinguished from *Smíete* by a shorter vowel duration, a higher F1, a lower F2, and smaller f0 dynamics. *Smitte* differs from *smiete* by a shorter vowel duration, a higher F1, a lower F2, a smaller fall size and smaller f0 dynamics. The triplet word *smiete* is distinguished from *Smíete* by larger f0 dynamics. Hence, tense short vowels have largest f0 dynamics.

3.1.2. Results /ʊ u uː/ triplet

Duration values for the /ʊ u uː/ triplet are shown in the lower panel of Figure 1. Vowel plots are found in Figure 2 and f0 contour plots in Figure 4. Statistical results are summarized in Table 3. In Table 1 we find that *ful* is represented by just one sample in the normal speech experiment. Therefore comparisons between *ful* and *fuul* and between *ful* and *fúul* are based on subject 2 only. The triplet word *ful* has a smaller vowel duration and higher F1 and F2 values than the other triplet words. Additionally, *ful* has a smaller /f/ duration then *fúul*. The triplet words *fuul* and *fúul* are not distinguished, i.e. short tense and long tense vowels are not distinguished.

3.2. Clear speech

3.2.1. Results /ɪ i iː/ triplet

Duration values for the /ɪ i iː/ triplet are shown in the upper panel of Figure 1. The clear speech graph shows a stronger contrast

between short vowels and long tense vowels than the normal speech graph. The vowel plots are found in Figure 2. Unlike in normal speech, short tense and long tense front vowels are separated in the acoustic vowel space in clear speech. Figure 3 shows that the difference between rise size and fall size in clear speech has decreased compared to normal speech. Statistical results are summarized in Table 2. The number of variables that distinguish triplet words is much larger in clear speech than in normal speech. When comparing *Smitte* and *Smíete* we find smaller values for the duration /m/, V, t2 and /ə/, a higher F1, a lower F2 and a smaller rise size. We get the same findings when comparing *smiete* and *Smíete*, but rise size is not significant and *smiete* has larger f0 dynamics than *Smíete*. *Smitte* has a smaller vowel duration, a higher F1, a lower F2, a smaller fall size and smaller f0 dynamics than *smiete*. V duration, F1 and F2 distinguish all of the triplet words.

3.2.2. Results /ʊ u uː/ triplet

Duration values for the /ʊ u uː/ triplet are shown in the lower panel of Figure 1. Again, the clear speech graph shows a stronger contrast between short vowels and long tense vowels. The vowel plots are given in Figure 2. Eventhough there is still some overlap in the acoustic vowel space for the long and short tense back vowels, the spread and the mean formant values show a clearer separation for the two sounds compared to the normal speech conditon. The f0 contour plots in Figure 4 show a longer f0 fall duration for the long tense vowel compared to the two other vowels. Statistical results are presented in Table 3. *ful* is distinguished from *fúul* by a smaller vowel duration and a higher F1 and F2. *fuul* has a smaller vowel duration, a higher F2, and larger f0 dynamics than *fúul*. *ful* is distinguished from *fuul* by a higher F1.

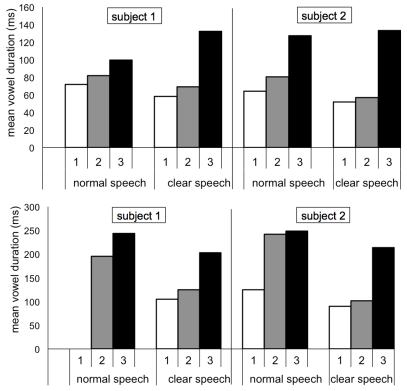


Figure 1: *Duration values for normal speech and clear speech. On top the /ɪ i iː/ triplet (1=Smitte, 2=smiete, 3=Smíete) and at the bottom the /ʊ u uː/ triplet (1=ful, 2=fuul, 3=fúul).*

3.3. Consensus of triplets

In Table 4 we show consensus results of the two triplets. For normal speech we find that lax vowels are distinguished from tense vowels by a higher F1 and a lower (/ɪ i iː/ triplet) or higher (/ʊ u uː/ triplet) F2. Additionally, short lax vowels are distinguished from long tense vowels by a smaller vowel duration. Short tense and long tense vowels are not distinguished by any variable in the normal speech data.

We find that lax and tense short vowels are distinguished only by a higher F1 in clear speech. Short lax vowels are

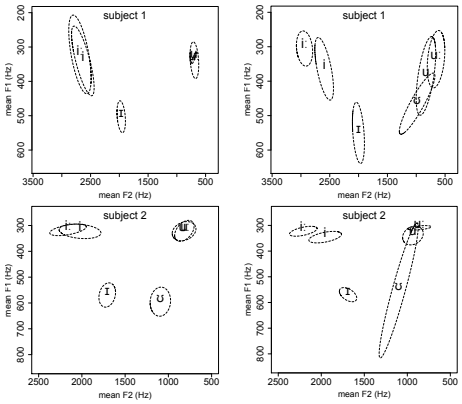


Figure 2: *Vowel plots show the mean formant values of the six triplet sounds for normal speech (left) and clear speech (right). Ellipses enclose two standard deviations.*

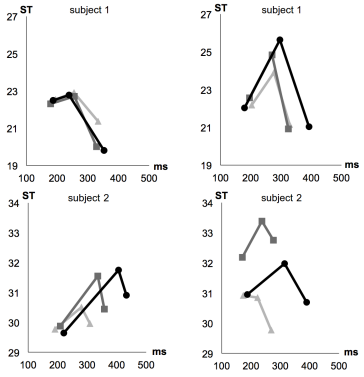


Figure 3: *f0 contours for the /ɪ i iː/ triplet in normal speech (left) and clear speech (right) for subject 1 (top) and subject 2 (bottom). Lighter gray lines with triangles represent short lax vowels, darker gray lines with squares represent short tense vowels, and black lines with circles represent long tense vowels.*

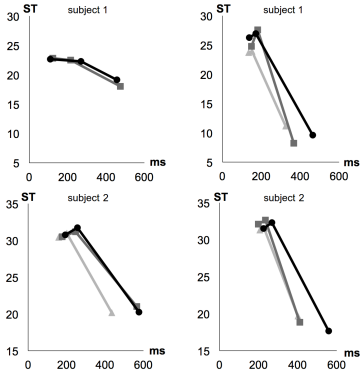


Figure 4: *f0 contours for the /ʊ u uː/. For graphical conventions see Figure 3.*

distinguished from long tense vowels by the same variables as for normal speech. In contrast to normal speech, short tense vowels and long tense vowels are clearly distinguished. Short tense vowels have a smaller vowel duration, a lower (/ɪ i iː/ triplet) or higher (/ʊ u uː/ triplet) F2, and larger f0 dynamics than long tense vowels.

Table 5 shows percentages of stimuli that were correctly predicted on the basis of V duration, F1, F2, and f0 dynamics, and which were obtained by Linear Discriminant Analysis. As expected, percentages for clear speech are higher than for

Table 2: *Results for the /ɪ i iː/ triplet. 1=Smitte, 2=smiete, 3=Smíete.*

	Normal speech			Clear speech		
	sig. 1·2	sig. 1·3	sig. 2·3	sig. 1·2	sig. 1·3	sig. 2·3
/s/ duration						
/m/ duration					**	***
V duration		**		*	***	***
t1 duration						
t2 duration					**	**
/ə/ duration					**	***
F1	***	***		***	***	***
F2	***	***		***	***	***
f0 rise size					**	
f0 fall size	**			*		
f0 dynamics	*	*	*	**		*

Table 3: *Results for the /ʊ u uː/ triplet. 1=ful, 2=fuul, 3=fúul.*

	Normal speech			Clear speech		
	sig. 1·2	sig. 1·3	sig. 2·3	sig. 1·2	sig. 1·3	sig. 2·3
/f/ duration		*				
V duration	***	***			***	***
/l/ duration						
F1	***	***		**	***	
F2	***	***			***	*
f0 rise size						
f0 fall size						
f0 dynamics						**

normal speech. In the case of /ɪ i iː/, all members of the triplet can be distinguished by using vowel duration, F1, and F2 as acoustic cues.

Table 4: *Consensus results of subjects and triplets. For normal speech and clear speech those variables are listed which play a role for both subjects and both triplets. 1=short lax vowel, 2=short tense vowel, 3=long tense vowel.*

speech type	pair	V dur	F1	F2	f0 dyn.
normal	1 · 2		***	***	
	1 · 3	**	***	***	
	2 · 3				
clear	1 · 2		**		
	1 · 3	***	***	***	
	2 · 3	***		*	*

4. Conclusions

We found that vowel duration, spectral variables, and f0 dynamics are cues for the distinction of vowels in two Saterland Frisian word triplets.

Vowel duration. From Table 4 we conclude that vowel duration implements the phonological feature [± long]. In normal speech short lax and long tense vowels are distinguished by vowel duration, whereas in clear speech short tense and long tense vowels are also distinguished, suggesting a division in short and long vowels.

F1 and F2. Table 4 suggests that the feature [± tense] is implemented by spectral features. Lax vowels have a higher F1 and a lower (/ɪ i iː/ triplet) or higher (/ʊ u uː/ triplet) F2 than tense vowels, i.e. lax vowels are more centered than tense vowels (cf. Figure 2). Gussenhoven [14] observed that closed vowels are perceived as relatively longer than open vowels, which suggests that lowering of F1 may increase perceived vowel duration, which can be explained by a tendency of listeners to compensate for the intrinsically shorter duration of closed vowels. In view of this finding, the slight additional centralization of the short tense vowels (/i/ and /u/ cf. Figure 2) relative to /iː/ and /uː/ may be used to enhance the perceived durational contrast between the short and long tense vowels.

Table 5: *Discriminant percentages per triplet word and per subject. For each subject important variables contributing to discrimination are marked by 1 and/or 2.*

triplet	sp. type	V dur	F1	F2	f0 dyn.	sub1 %	sub2 %
/ɪ i iː/	norm.	1,2	1,2	1,2	2	90.5	91.7
	clear	1,2	1,2	1,2		100	100
/ʊ u uː/	norm.		1,2	1	2	55.6	65.9
	clear	1,2	1,2	1,2	1	92.3	96.2

F0 dynamics. Short tense vowels were found to have larger f0 dynamics than long tense vowels and, in case of [ɪ], short lax vowels (cf. Figures 3 and 4). According to [5] and [7], increased f0 dynamics may enhance perceived vowel duration. As shortening of /i/ and /u/ in clear speech resulted in vowels that were hardly longer than the short lax vowels [ɪ] and [ʊ], increased f0 dynamics may increase the perceived durational difference between short lax and tense vowels, at least in the case of [ɪ] and [i]. We found that f0 dynamics have a systematic effect which contributes to the tripartite vowel contrast. The most likely interpretation of this contribution is phonetic feature enhancement. The possibility that variation of tonal structure is involved as suggested by Siebs' terms *Stoßton* and *Schleifton* [9] cannot be determined on the basis of our current data, but we will investigate this in a later study.

Overall, our data suggest that the phonological contrasts between [ɪ i iː] and between [ʊ u uː] can be accounted for by the combination of two distinctive features, [± long] and [± tense]. The slight centralization of /i/ and /u/ relative to /iː/ and /uː/ in clear speech may be used to enhance the durational contrasts between the short and long tense vowels. Increased f0 dynamics of tense short vowels relative to lax short vowels, which was found for [ɪ] and [i], may be regarded as a means to make the short tense vowels sound more different from the short lax vowels.

Kohler [15] investigated triplets of closed vowels in High German and in Low German dialects spoken in Schleswig-Holstein and Lower Saxony. In some dialects he found that lax and tense vowels within a triplet differ qualitatively, and short and long tense vowels differ quantitatively. Just as Kohler we found that short lax and tense vowels are distinguished by spectral features only. The clear speech experiment revealed that short and long tense vowels are dinstinguished by vowel duration, F2, and f0 dynamics, being a combination of qualitative and quantitative variables.

5. Acknowledgements

We would like to thank the two Saterland Frisian informants for participating in our experiments. We are grateful to Darja Appelganz and Nicole Mayer for labeling the recordings in PRAAT. The research reported in this paper has been funded by the *Deutsche Forschungsgemeinschaft* (DFG), grant number PE 793/2-1.

6. References

[1] Stellmacher, D., “Das Saterland und das Saterländische”, Florian Isensee, 1998.

[2] Sjölin, B., “Einführung in das Friesische”, Metzler, 1969.

[3] Fort, M.C., “Saterfriesisches Wörterbuch mit einer grammatischen Übersicht”, Buske, 1980.

[4] Kramer, P., “Kute Seelter Sproakleere – Kurze Grammatik des Saterfriesischen”, Ostendorp, 1982.

[5] Lehiste, I., “Influence of fundamental frequency pattern on the perception of duration”, Journal of Phonetics 4, 113–117, 1976.

[6] Yu, A. C. L., “Tonal effects on perceived vowel duration”, in C. Fougeron, B., Kuehnert, M., D’Imperio, M. and N. Vallée [Eds], Laboratory Phonology 10, 151–168, Mouton de Gruyter, 2010.

[7] Cumming, R. E., “The effects of dynamic fundamental frequency on the perception of duration”, Journal of Phonetics 39, 375–387, 2011.

[8] Lehnert-LeHouillier, H., “A cross-linguistic investigation of cues to vowel length perception”, Journal of Phonetics 38, 472–482, 2010.

[9] Siebs, Th., “Zur Geschichte der englisch-friesischen Sprache”, Max Niemeyer, 1889.

[10] Tröster-Mutz, S., “Phonologie des Saterfriesischen”, MA Thesis University of Osnabrück, 1997.

[11] Tröster-Mutz, S., “Untersuchungen zu Silbenschnitt und Vokallänge im Saterfriesischen”, Theorie des Lexikons 120, Heinrich-Heine-University, Düsseldorf, 2002.

[12] Boersma, P. and Weenink, D., “Praat: Doing Phonetics by Computer”, available at http://www.praat.org, 1992-2013.

[13] Grabe, E., “Pitch accent realization in English and German”, Journal of Phonetics 26, 129–143, 1998.

[14] Gussenhoven, C., “Vowel height split explained: Compensatory listening and speaker control”, in J. Cole and J.I. Hualde [Ed], Laboratory Phonology 9, 145–172, Mouton de Gruyter, 2007.

[15] Kohler, K.J., “Überlänge im Niederdeutschen?”, in R. Peters, H.P. Pütz and U. Weber [Eds], Vulpis Adolatio. Festschrift für Hubertus Menke zum 60. Geburtstag, 385–402. C. Winter, 2001.

An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based on yes/no questions

Marilisa Vitale [1,2], *Philippe Boula de Mareüil* [2], *Anna De Meo*[1]

[1] Università degli Studi di Napoli "L'Orientale", Naples, Italy
[2] LIMSI-CNRS, Orsay, France

vitalem@unior.it, philippe.boula.de.mareuil@limsi.fr, ademeo@unior.it

Abstract

The present study investigates the prosody of yes/no questions (in comparison with statements) in Chinese learners and native speakers of Italian. Acoustic analyses and a perceptual test were performed, in order to identify the main trends in non-native productions. Results show the relevance of prosody, which differentiates elementary, intermediate and advanced Chinese learners of Italian. Listening tests based on prosody transplantation also suggest that non-native segments with a native Italian prosody are rated as less accented than are native Italian segments with a non-native prosody. Similar trends were found, overall, in terms of question/assertion discrimination, confirming the relative importance of prosody. These findings could be helpful for teachers and learners of Italian as a foreign language.

Index Terms: L2 Italian, yes-no questions, Chinese learners, prosody acquisition

1. Introduction

To acquire a good command of prosody in a second language (L2), it is crucial to identify the most salient patterns of the target language and to determine the main tendencies observed in the performance of L2 speakers. In L2 Italian, especially, the teaching of prosody often focuses on questions: this speech act is very important in everyday communication and it is introduced early in language classes [1].

A large body of research has concentrated on problems caused by question intonation. Ullakonoja [2] as well as Santiago-Vargas and Delais-Roussarie [3], for instance, were interested in the challenges posed by yes/no questions to Finnish learners of Russian and Mexican Spanish learners of French, respectively. The present study investigates problems faced by Chinese learners of Italian as an L2, when asking yes/no questions. It intends to better understand how speakers of a tone language perform in a non-tone language. Chinese questions have to preserve, to some extent, lexical tone pitch variations in order to keep the utterance meaning unchanged [4], whereas the Italian language distinguishes between questions and statements by intonation alone. Questions are usually marked by a rising-falling pitch movement on the last stressed syllable of the utterance (most often the penultimate syllable), at least in southern varieties of Italian [5] [6] [7]. Even though a common strategy is observed in the two languages under consideration in the present study, namely an overall higher pitch range in interrogative utterances [8] [9], interferences between the Italian and Chinese systems may affect the acquisition of prosody in the L2. For an overview on the prosody of yes/no questions, see [5] [7] for Italian, and [10] for Chinese.

This study compares statements and yes/no questions produced by both Chinese learners (of various proficiency levels in Italian) and native speakers of Italian. It combines acoustic analyses and a perceptual test using prosody transplantation. This paradigm [11] consists of copying prosodic parameters from a native Italian utterance to a non-native one and vice versa. Unlike previously used techniques such as low-pass filtering [12], the method preserves comprehensibility: it thus yields a more ecological speech material. It was applied to various language pairings [13] [14]: in particular, it was used so as to disentangle the contribution of prosodic features (melody and global timing) and segmental features to the perception of accentedness and intelligibility in German-accented English and English-accented German. Here, the same methodology was used in order to evaluate the contribution of prosody to the perception of Chinese foreign accent and to the discrimination between questions and statements in Italian. The method, materials and results are described below.

2. Corpus

2.1. Speakers and material

For this study, 4 native Italian speakers and 12 Chinese learners of Italian (all females, aged 24 on average) were recorded. The learners, all speakers of Mandarin Chinese, belonged to three groups (each composed of 4 students) according to their level of competence in Italian (elementary, intermediate, advanced). All speakers lived in the Naples region (southern Italy).

Native and non-native speakers were recorded while reading a short dialogue in Italian, including 7 yes/no questions. These questions varied in length (between 5 and 11 syllables each), but they all ended with a paroxytone (that is, a word stressed on the penultimate syllable). This choice was determined by the need to avoid stress-related variation and by the frequency of this stress pattern in Italian: the Italian vocabulary is mostly composed of paroxytone words (in over 75% of cases [15]). Speakers were then instructed to read the same questions in the declarative modality — word order does not change in Italian. The corpus analysed below was therefore made up of 112 questions and 112 statements. Examples of sentences are given in § 3.2.

2.2. Acoustic analysis

The 224 resulting utterances (as many questions as statements) were analysed to identify prosodic differences between native and non-native speakers of Italian, as well as across the three learner groups. For each utterance, mean pitch, the number and duration of inter-pause speech intervals, the number of syllables per inter-pause speech interval, the duration of silent pauses and disfluencies were computed. The speech articulation rate (excluding pauses), phonation rate (including pauses), tonal range (between maximum and

minimum defined pitch values) and the percentage of disfluencies were calculated. Furthermore, fundamental frequency (f0) values were measured at the midpoint of each of the last three syllables, for each utterance. These measurements were taken using the Praat software [16].

Mean pitch differences in semitones (ST) between questions and statements were averaged for each speaker group. In each group, questions are higher than statements: by 0.5 ST for elementary learners, 0.2 ST for intermediate learners, 1.7 ST for advanced learners and 0.8 ST for native speakers. Hence, no clear tendency seems to emerge.

Results of the other acoustic analyses are shown in Table 1. As expected, the speech articulation rate and phonation rate of elementary learners are lower than those of the other two learner groups and those of native speakers. Note that already at an intermediate level of language proficiency, Chinese learners are able to properly manage duration-related variations. Intermediate learners, together with the advanced learner group, show values very close to the native model, partly due to our native speakers' tendency to hyperarticulate.

Table 1. *Acoustic measurements: articulation rate (syll./s), phonation rate (syll./s), tonal range (semitones), % dis. (percent time of disfluencies) within questions (Q) and statements (S).*

Speaker group	Q	S
	Articulation rate	
Elementary L2	3.8	3.8
Intermediate L2	5.2	5.1
Advanced L2	5.5	5.0
Native Italian	5.4	5.3
	Phonation rate	
Elementary L2	3.3	3.7
Intermediate L2	5.1	5.0
Advanced L2	5.5	5.0
Native Italian	5.4	5.3
	Tonal range	
Elementary L2	12.6	8.3
Intermediate L2	7.2	7.9
Advanced L2	7.7	7.2
Native Italian	9.4	13.7
	% dis.	
Elementary L2	3.2	0.0
Intermediate L2	0.1	0.2
Advanced L2	0.0	0.0
Native Italian	0.0	0.0

As for the tonal range of both questions and statements, the three groups of Chinese learners seem to be unable to produce varied pitch movements as native Italian speakers do. On average, non-native speakers' tonal range is more reduced than that of native speakers — by 4 ST. Interestingly, the only exception stems from questions asked by elementary learners. Because of their linguistic insecurity (their uncertainty and lack of confidence in how to pronounce words properly), these speakers tend to produce creaky-voiced vowels and sometimes a high-rising terminal tune, resulting in a tonal range even wider than that of native speakers.

As far as disfluencies are concerned, they mainly come from elementary learners of Italian. Due to the use of read speech, they are very few (only 3% of utterance duration). They mainly consist of self-repairs and repetitions.

In addition to this, the final pitch contours of the 224 sentences of the corpus were analysed in detail. Results are displayed in Figure 1 in semitones calculated with respect to the minimum f0 measure (172 Hz). The comparison of Fig. 1a and 1b shows that the involved native Italian speakers' terminal pitch range is higher in questions than in statements (with a 3.6 ST difference). The average f0 range difference between questions and statements is 3 ST in the case of advanced learners. For the other two groups of Chinese speakers, f0 values are quite similar between the speech acts considered: the measured increase in questions is only 1.1 ST for intermediate learners and 0.7 for elementary learners.

In question-final syllables, the involved native speakers produced a slight upward-downward curve, which is the typical pitch movement of the Italian language (or at least its Neapolitan variety), as evidenced by the literature (e.g. [7]). The group of elementary learners, on average, displayed no movement at all but instead produced a flat pitch curve. It should be noted however that, in contrast with this general trend, several beginners' questions exhibit a rising pitch movement on the very last syllable, which sounds foreign as will be confirmed by the perception test. A flat pitch pattern can also be found in the intermediate speakers' productions, whereas the advanced learners' pitch rise on the penultimate (stressed) syllable is closer to the native Italian model (see Figure 1a).

In statements, a sharp pitch fall is noticeable only in native Italian speakers. The three groups of Chinese learners, on average, show gradual downstep towards the utterance-final syllable (see Figure 1b).

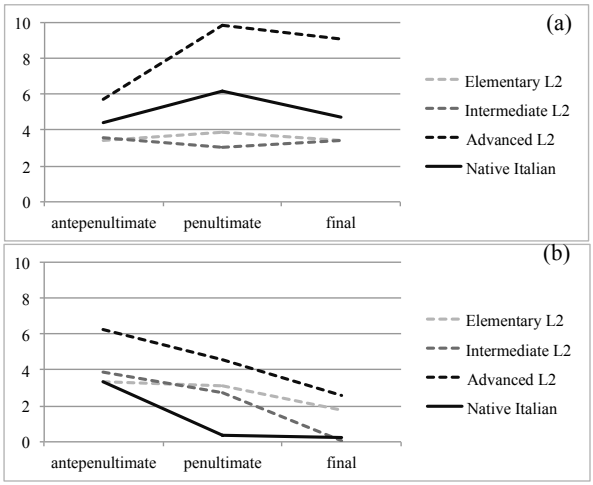


Figure 1: *Last three syllables (average values in semitones) of questions (a) and statements (b).*

3. Perceptual test

3.1. Protocol and listeners

A perceptual test was then conducted to check whether and to what extent there was an audible improvement of prosody from Chinese elementary, intermediate and advanced learners to native Italian speakers, in terms of native-likeness and communicative effectiveness. More precisely, this test aimed at evaluating speakers' degree of foreign accent and

possible difficulties encountered by Italian listeners in properly identifying questions and statements. These difficulties are expected to be primarily related to prosody, rather than to the segmental level. As regards the perception of a foreign accent, it may be due to both prosodic and segmental levels [11] [14]. For this purpose, the prosody transplantation technique was used so as to exchange native and non-native acoustic-prosodic features, phoneme by phoneme. It relied on the PSOLA algorithm implemented in Praat, which allows phoneme durations to be matched, between two realisations of a given utterance, before grafting the f0 curve of version 1 onto version 2 and vice versa.

The perceptual test was administered to 40 native Italian listeners (18 males and 22 females, aged 34 on average) with no reported history of hearing impairment, coming from various regions of Italy. Almost all of them (34 out of 40) were experts in the field of language and phonetics research. However, none of them could speak an Asian or a tonal language.

3.2. Material and tasks

Out of the 84 questions produced by Chinese learners of Italian (7 × 4 per group), 12 questions were used for the perceptual test. Nine of these sentences, uttered by 3 different speakers from each of the 3 learner groups, had the following structures: *È una cosa che si usa spesso?* 'Is it something often used?', *Questo oggetto si trova in casa?* 'Is this object in the house?', *Si usa quando si è stanchi?* 'Is it used when somebody is tired?'. Another 3 questions, uttered by another speaker from each of the 3 learner groups, had the following structure: *È un(a)* 'Is it a' + Noun, with Noun being *divano* 'sofa', *persona* 'person' or *oggetto* 'object'. In addition, 12 questions were selected among the native Italian speakers' corresponding renditions, in such a way that each learner group was associated with 4 different Italian voices producing each of the 4 selected questions. Prosody transplantation was then performed, thus generating 24 additional stimuli. The same type of selection and prosody transplantation procedure was carried out for statements. The question + statement samples finally used in the perceptual test were therefore composed of:

- 12 Q + 12 S original non-native Italian utterances (4 + 4 elementary, 4 + 4 intermediate and 4 + 4 advanced);
- 12 Q + 12 S original native Italian utterances, selected so as to never have the same voice repeating the same utterance and to have different voices associated with each of the three learner groups;
- 12 Q + 12 S utterances with native Italian segments and a non-native prosody;
- 12 Q + 12 S utterances with non-native segments and a native Italian prosody.

The 96 resulting stimuli were administered to the 40 listeners in different random orders, through an online interface (http://www.audiosurf.org/test_perceptif_marilisa/). Participants were informed that the experiment dealt with the Italian language spoken by native and non-native subjects, and that they would listen to excerpts of original or acoustically-modified speech. They were advised to use headphones or earphones.

Subjects first provided autobiographical information (age, education, place of residence, etc.). Also, they were asked very general questions, before a short familiarisation phase with the types of stimuli. They first listened to examples of native/non-native, original/manipulated statements and questions (not used in the actual test). For each utterance, they were then asked to:

- assess the degree of foreign accent on a continuous 0–5 scale (0 = no foreign accent; 5 = very strong foreign accent);
- identify the correct speech act, discriminating between "question" and "statement".

To accomplish the first task, a slider was provided (by default located in the middle of the scale); for the second task, participants had to click on buttons. They could listen to each stimulus as many times as they needed, but it was not possible to correct previous answers once a new stimulus was displayed.

At the end of the test, listeners were asked other questions (1) to indicate the speakers' most salient and characteristic linguistic cues and (2) to identify the native and non-native speakers' geographical and linguistic backgrounds.

3.3. Results

The results of the perceptual test are reported in Table 2. On the basis of original stimuli, analyses of variance (ANOVAs) were carried out separately on listeners' responses, in terms of rating and correct speech act identification, with the two fixed factors Modality (question or statement) and Level group (elementary, intermediate, advanced or native). On the basis of prosody-transplanted stimuli, a second series of ANOVAs was performed on listeners' responses, with the two fixed factors Modality (question or statement) and Type of stimulus (6 levels: see Table 2b).

The degrees of foreign accent attributed by listeners to the original stimuli gradually decrease from elementary learners to native speakers of Italian (Table 2a). The Italian pronunciation of Chinese speakers improves linearly, in both questions and statements, and L1 Italian speakers (with a 0.1 degree of foreign accent) are properly identified as native speakers. Differences related to Modality are not significant, but differences across Level groups are significant [$F(3, 1960) = 2387$; $p < 0.001$] — the interaction between the two factors is marginal. Tukey's HSD post hoc test shows that all differences across Level groups are higly significant.

Responses in terms of question/statement discrimination display slightly different patterns, with significant differences related to Modality [$F(1, 3928) = 10$; $p < 0.001$], significant differences across Level groups [$F(3, 3928) = 123$; $p < 0.001$] and a significant interaction between the two [$F(3, 3928) = 12.7$; $p < 0.001$]. As in accent ratings, a linear progression is observed in the speech act identification of statements. In contrast, beginners' questions are better recognised than are intermediate students' questions, probably due to the fact that the elementary level speaker group sometimes produces high-rising terminal tunes. These pitch rises on utterance-final syllable rather than on the last stressed syllable of the question do sound foreign but they may be effective from a communicative point of view.

The use of acoustically modified stimuli allowed us to tease apart the influence of prosodic (suprasegmental) level and segment articulation in terms of both foreign accent rating and question/statement discrimination. As shown in Table 2b,

the stimuli with non-native segments and a native Italian prosody are perceived as having a slight foreign accent, whereas the stimuli with native Italian segments and a non-native prosody are perceived as more strongly foreign accented. Differences related to Modality are not significant, but the Type of stimulus has a major effect [$F(5, 1908) = 101$: $p < 0.001$] — the interaction between the two being marginal. Post hoc comparisons (Tukey's HSD test) show that stimuli containing native Italian prosody form a homogeneous subset, but differ significantly from the other stimuli.

A similar trend is observed concerning the speech act identification of questions vs. statements. Differences related to Modality are still not significant, but the Type of stimulus has a major effect [$F(5, 1908) = 87.1$: $p < 0.001$] and the interaction between the two is here significant [$F(5, 1908) = 2.23$; $p < 0.05$]. Subsequent post hoc comparisons (Tukey contrasts) show that stimuli with non-native segments and a native Italian prosody are discriminated significantly better than stimuli with native Italian segments and a non-native prosody. Both questions and statements with the native Italian prosody turn out to be identified almost perfectly by the listeners, regardless of possible mispronunciations at the segmental level. By contrast, the superimposition of the non-native rhythm and intonation on native Italian productions dramatically reduced question/statement discrimination.

Table 2. *Average degree of perceived foreign accent and speech act identification of the original (a) and manipulated (b) stimuli Q = questions; S = statements.*

(a) Original stimuli		
Speaker group	Q	S
	Degree (/5)	
Elementary L2	4.3	4.1
Intermediate L2	3.7	3.6
Advanced L2	2.6	2.6
Native Italian	0.1	0.1
	% Q/S Id	
Elementary L2	59	48
Intermediate L2	41	72
Advanced L2	73	82
Native Italian	100	99

(b) Manipulated stimuli			
Prosody donor	Prosody receiver	Q	S
		Degree (/5)	
Elementary L2	Native Italian	3.6	3.5
Intermediate L2	Native Italian	2.4	2.7
Advanced L2	Native Italian	1.8	1.8
Native Italian	Elementary L2	1.6	1.4
Native Italian	Intermediate L2	1.8	1.6
Native Italian	Advanced L2	1.7	1.5
Prosody donor	Prosody receiver	Q	S
		% Q/S Id	
Elementary L2	Native Italian	59	56
Intermediate L2	Native Italian	62	73
Advanced L2	Native Italian	69	76
Native Italian	Elementary L2	98	98
Native Italian	Intermediate L2	97	98
Native Italian	Advanced L2	99	96

In their final comments, most listeners (about 30) mentioned intonation and rhythm among the features which helped them make their decisions. Also, 30 subjects properly identified the regional origin of native speakers (i.e. Campania) and the linguistic background of non-native speakers (i.e. Chinese). Most of them reported an average familiarity with Chinese-accented Italian, which suggests that their evaluations were relevant.

4. Conclusions

This study based on Chinese learners of Italian as an L2 included acoustic analyses and a perception test in which suprasegmentals took precedence over segmentals, as in a previous study focusing on Spanish-accented Italian [11]. It highlighted the importance of focusing on prosody in order to improve both native-likeness and communicative effectiveness, at least as far as the perception of questions and statements is concerned. Overall, the more proficient the speakers, the better their questions are judged. However, results suggest that Chinese learners of Italian, a non-tonal language, succeed in producing yes/no questions in an appropriate manner only if they have acquired an advanced level of proficiency in the L2. It therefore seems important to point out the main problems faced by Chinese speakers of Italian, in order to give them feedback when learning the Italian yes/no question prosody.

Primarily, the pitch range of questions and statements need to be differentiated, since it is usually more extended in the case of questions. When asking questions ending with paroxytone words (a frequent pattern in Italian), Chinese learners of Italian should also be taught to realise a rising-falling pitch movement on the penultimate, prominent syllable of the utterance. Particular attention must be paid to beginners' tendency to produce a pitch rise on the utterance-final syllable: even if this intonational movement is communicatively effective in the sense that it is prone to be interpreted as a question contour, it is far from the native Italian model and sounds foreign.

The Chinese speakers under investigation here were also recorded in their L1. Acoustic analyses of their native productions are currently in progress and provide interesting comparisons. Finally, the results presented here need to be validated by further studies on spontaneous speech, to provide a clearer picture of question intonation in Italian as a foreign language. The understanding of questions in conversational speech will undoubtedly be another challenge.

5. Acknowledgements

This work was done during the first author's stay at LIMSI-CNRS, financed by an Erasmus Placement grant. We are grateful to all the speakers and listeners who volunteered to take part in the experiments.

6. References

[1] Spinelli, B. and Parizzi, F., Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2, La Nuova Italia, Florence, 2010.

[2] Ullakonoja, R., "How do native speakers of Russian evaluate yes/no questions produced by Finnish L2 learners?", Rice Working Papers in Linguistics, 94(2):92–105, 2010.

[3] Santiago-Vargas, F. and Delais-Roussarie, É., "The acquisition of question intonation by Mexican Spanish learners of French", in É. Delais-Roussarie, M. Avanzi, S. Herment [Eds], Prosody and languages in contact. L2 acquisition, attrition, languages in multilingual situations, Springer Verlag, Berlin, 2014 (to appear).

[4] Visceglia, T. and Fodor, J. D., "Fundamental frequency in Mandarin and English: Comparing first- and second-language speakers", in C. Lleao [Ed.], Interfaces in Multilingualism: Acquisition and Representation, John Benjamins, Amsterdam, pp. 27–59, 2006.

[5] D'Imperio, M., "Italian intonation: An overview and some questions", Probus, 14:37–49, 2002.

[6] Grice, M., D'Imperio, M., Savino, M., Avesani, C., "Strategies for intonation labelling across varieties of Italian", in S.-A. Jun, [Ed.], Prosodic typology: the phonology of intonation and phrasing, Oxford University Press, Oxford, pp. 55–83, 2005.

[7] Savino, M., "The intonation of polar questions in Italian: Where is the rise?", Journal of the International Phonetic Association, 42(1):23-48, 2012.

[8] Bolinger, D., "Intonation across languages", in J. Greenberg [Ed.], Universals of human language (Volume 2: Phonology), Standford University Press, Standford, pp. 471–524, 1978.

[9] Gussenhoven, C., "Intonation and interpretation: phonetics and phonology", in Proceedings of the 1st International Conference on Speech Prosody, pp. 47–57, 2002.

[10] Yuan, J., "Mechanism of Question Intonation in Mandarin", in Q. Huo, B. Ma, E.-S. Chng, H. Li [Eds], Chinese spoken language processing, Springer Verlag, Berlin, pp. 19–30, 2006.

[11] Boula de Mareüil, P. and Vieru-Dimulescu, B., "The contribution of prosody to the perception of foreign accent", Phonetica, 63(4):247-267, 2006.

[12] Munro, M. J., "Nonsegmental factors in foreign accent: Ratings of filtered speech". Studies in Second Language Acquisition, 17:17–34, 1995.

[13] Holm, S., "Intonational and durational contributions to the perception of foreign-accented Norwegian: An experimental phonetic investigation", PhD thesis, Norwegian University of Science and Technology, Trondheim.

[14] Winters, S. and O'Brien, M.G., "Perceived accentedness and intelligibility: The relative contributions of F0 and duration", Speech Communication, 55(3):486–507, 2013.

[15] D'Imperio, M. and Rosenthall, S., "Phonetics and phonology of main stress in Italian", Phonology 16(1):1–28, 1999.

[16] Boersma, P. and Weenink, D., "Praat: doinghonetics by computer" [Computer program], Version 5.3.15 retrieved 22 May 2012 from http://www. praat.org/, 2012.

Final Lowering Effect in Questions and Statements of Chinese Mandarin Based on a Large-scale Natural Dialogue Corpus Analysis

Wei Lai [1, 2], *Ya Li* [2], *Hao Che* [2], *Shanfeng Liu* [2], *Jianhua Tao*[2], *Xiaoying Xu* [1, 2]

[1] School of Chinese Language and Literature, Beijing Normal University, Beijing, China
[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China

laiwei_0508@126.com, {yli,hche,sfliu,jhtao}@nlpr.ia.ac.cn, xuxiaoying2000@bnu.edu.cn

Abstract

To support text-to-speech with detailed prosody rules and to generate natural prosody, the paper studied the pitch variation near the end of sentences based on a Chinese Mandarin natural dialogue corpus. An additional lowering effect on the last prosodic word was found in both questions and statements, and proved to be independent of tone influence. Nevertheless, this effect, which is referred to as final lowering in other languages, was claimed to be absent in Chinese by some previous experimental studies. Such a contradiction is very likely to be caused by the difference between experimental speech versus natural speech. Based on this observation, the paper proposed a combination of the two methods in intonation studies, in which experimental speech served as an entry point to develop new topics, while natural speech served as a necessary extension to revise and apply prosody rules.

Index Terms: intonation, questions, final lowering, spontaneous speech

1. Introduction

In many languages, pitch within an utterance tends to drift down, especially in statements. There are many factors accounting for the pitch downtrend. One is the declination effect, which refers to the general tendency of pitch declination over the course of an utterance [4][6][9][12][14]. Another is final lowering, indicating an additional lowering effect near the end of the sentence [5][7]. Besides, there is also a downstep effect, which means that pitch lowering can be triggered by former syllables that carry low tones/accents [1]. Experimental studies of Mandarin showed the presence of declination and downstep [16][17][18], but the absence of final lowering [12]. In Chinese Mandarin, the downstep effect is supposed to be triggered by both low tone T3 and neutral tone T5 [12][16][17]. When it comes to questions, the downtrend is considered suppressed [14].

There have already been a considerable number of intonation studies on questions of Chinese Mandarin. Features of questions in Mandarin were put forward as: a) trends of top/base-lines, Shen, Jiong adjusted Gårding's grid model and suggested a gentle fall on the top-line and a slight rise on base-lines for questions [3][10]; b) starting point, Shen, Xiaonan suggested that all types of questions begin at a register higher than statements [11]; c) boundary tone, Lin, Maocan adopted AM theory and underlined the role of boundary tone in the distinction of interrogative and declarative moods [8]; d) phrase curve and strength, Yuan, Jiahong thought interrogation to be expressed by an overall higher phrase curve and higher strength on final syllables in Chinese [19].

However, the above conclusions cannot fully satisfy the needs of speech synthesis. For one thing, these studies mainly focused on the general intonation trend of questions [10][11][19], but neglected the concrete details of pitch variation within sentences. Is pitch variation distributed evenly over the whole sentence, or is it mainly realized by certain parts of the sentence? Such questions still remain disputable. For another, in order to generate natural prosody, conclusions from experimental works should be checked and revised in natural speech. Since Chinese Mandarin is a tone language, previous intonation research tends to adopt designed stimuli to arrange tone types [12][16][17][18][19][20], while research on spontaneous speech is relatively few. To make up for these inadequacies, we particularly studied the pitch variation near the end of questions and statements by using a large-scale natural dialogue corpus, with the purpose of providing speech synthesis with more detailed prosody rules, and thereby making contributions to generating natural prosody.

2. Corpus

Our research is based on a large-scale Q&A conversation corpus. The corpus contains more than 4000 sentences, i.e., more than 2000 turns of questions and answers, selected from interview programs. A wide range of topics are involved and sentences with less normal style were rewritten. These conversations were transliterated and then re-read by a trained male speaker in a professional recording studio to make sure that F0 values of different sentences come from the same speaker and are comparable. During the recording process, the speaker was required to maintain a natural speaking style without act or exaggeration. The corpus contains an annotation of four-level prosodic units, i.e., syllables, prosodic words, prosodic phrases and intonation phrases, which were manually checked by the first author.

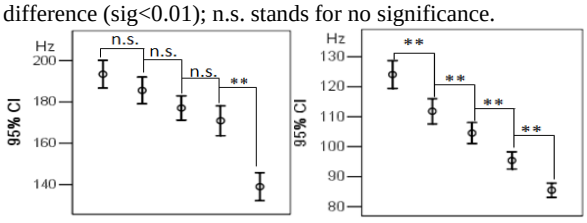
All the questions in the corpus can be divided into 6 types according to syntax structure and pragmatic function. There are 1405 wh- questions, 184 v-neg-v questions (a kind of particular Chinese question syntax), 101 alternative questions, 382 yes-no questions, 114 tag questions, and 2179 statements. Yes-no questions can be further divided into 41 unmarked yes-no questions and 341 particle yes-no questions (including interrogative particles "吧"/ba/ and "吗"/ma/). Since the contrast focus of alternative questions and the tag utterance of tag questions will make the situation much more complicated, they are not discussed in this paper.

3. Experiments and Results

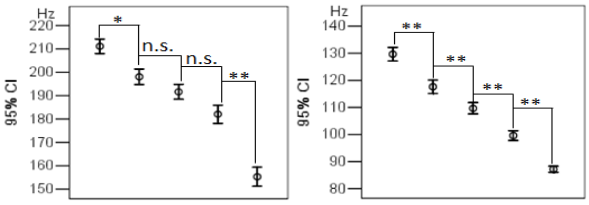
The present experiment was designed to study the F0 variations between prosodic words near the end of the sentences. Peaks and valleys of the last five prosodic words in each type of sentences were extracted. Sentences made of less than five prosodic words were removed from the samples. T5 in Chinese is a neutral tone, and the syllables of T5 are always

unstressed. Therefore, only sentences that end up with full lexical tones (T1-T4) were selected for wh- questions, v-neg-v questions, unmarked questions and statements.

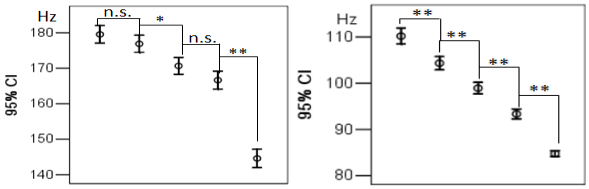
Error bars in Fig. 2 directly describe the pitch variations of the last five prosodic words in different types of sentences. Further ANOVA post-hoc test was done to compare pitch of adjacent prosodic words. Hereinafter, * stands for a significant difference (sig<.05); ** stands for an extremely significant difference (sig<0.01); n.s. stands for no significance.



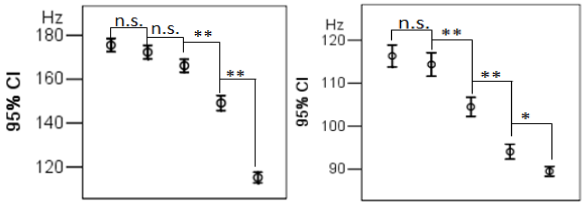
(a) The Topline (left) &Baseline (right) in V-neg-v Q



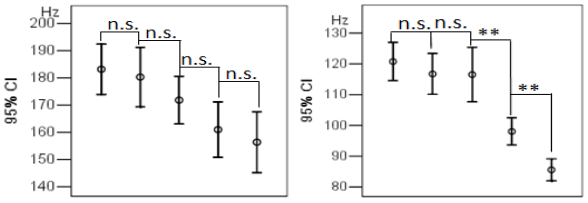
(b) The Topline (left) &Baseline (right) in Wh- Q



(c) The Topline (left) &Baseline (right) in Statement



(d) The Topline (left) &Baseline (right) in Particle Q



(e) The Topline (left) &Baseline (right) in Unmarked Q

Figure 2. *Pitch Variation of the Last Five Prosodic Words in Five Types of Sentences*

As shown in Fig. 2, pitch falls on top-lines are not evenly distributed over the whole sentence. An addition lowering effect appears at the end of all types of sentences except unmarked question, which is usually referred to as final lowering. Unlike top-lines, base-lines are encoded by a general pitch fall throughout the selected parts in most types of sentences. Therefore, the final lowering effect is supposed to be realized mainly by top-lines.

3.1. Top-lines: Presence of Final Lowering

Final lowering can be obviously detected in wh- questions, v-neg-v questions and statements. Pitch drops much more rapidly when it comes to the last prosodic word (around 30~40 Hz) than the former ones (around 10 Hz). Since sentences with neutral tone endings were removed from the sample, the above result was not influenced by the final unstressed syllables.

In particle questions, unstressed final particles were not removed and possibly made a contribution to the final pitch drop. However, a rapid fall also takes place between the last 2nd prosodic word and the last 3rd one. Since there is no evidence suggesting that neutral tone would affect its former syllables, the last second drop might be due to the effect of final lowering.

The presence of final lowering in v-neg-v questions and particle questions is supported by statistic results, for extremely significant differences are detected between the last (and last 2nd) pairs of prosodic words on top-lines. As for wh-questions and statements, there are also significant differences in other positions. Perhaps that is because pitch from natural speech varies more unstably due to the complication of tone. Although an extra pitch drop near the end of the sentence is observed in error bars of the two types of sentences, it has not yet been completely confirmed by statistical data. Thus, a supporting experiment is needed to eliminate tone differences and to narrow pitch gaps.

3.2. Base-lines: Evenly Distributed Pitch Falls

Fig. 2 shows two types of base-line encodings. One is a general decliantion throughout the selected part of sentences, which goes for wh-questions, v-neg-v questions and statements. The other is an additional final drop near the end of the sentence, which goes for yes-no questions. In unmarked yes-no questions, the drop takes place in the last two prosodic words. Coincidentally, in particle yes-no questions, if the last T5 particles were excluded, the extra drop also occurs on the last two prosodic words.

Statistical results on base-lines make a good fit with pitch variations in error bars. The difference between each pair of adjacent prosodic words is significant in wh- questions, v-neg-v questions and statements, suggesting a general smooth fall on base-lines. For unmarked questions and particle questions, the pitch differences stay insignificant except for the last two or three prosodic words. But in general, pitch falls on base-lines are distributed more evenly compared to top-lines.

3.3. Unmarked Q: Suppression of Final Lowering by Boundary Tone Effect

Final lowering is absent in unmarked questions, but the absence can be explained. Unlike other kinds of questions, unmarked questions do not depend on interrogative marks (such as wh- words, interrogative particles or syntactic means) to convey interrogative mood. Instead, boundary tone effect on the last syllable, which brings bring higher pitch as well as steeper pitch curse, is claimed to play an important part in the expression of interrogation [8]. According to our observations, final lowering in unmarked questions is partly set off by pitch rise caused by the boundary tone effect on the last syllable.

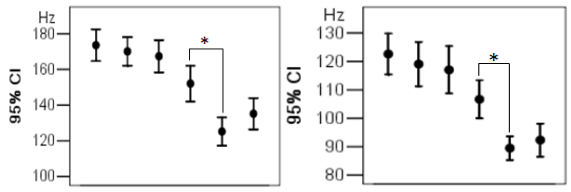


Figure 3: *Pitch Variation of the Last Six Syllables on Top-lines (left) and Base-lines (right) in Unmarked Q*

In Fig. 3, significant differences can be detected between the last 2nd &3rd syllables on both lines while in Fig. 2*(e)*, the differences becomes insignificant. This change is partly due to the pitch increase of the last syllable on top-lines. By contrast, the final pitch is not raised much on base-lines, so the significance is preserved. It is supposed that in unmarked questions, the pitch ascension of the last syllable on top-lines narrows the pitch gap between the last two prosodic words and suppresses the occurrence of final pitch drop to some extent.

3.4. The Basic Unit of Final Lowering: PW vs. SYL

In the light of Herman's data, the scope of final lowering is up to the last three syllables [5]. Also, we found that different syllables are influenced by different sentence types. According to Fig. 4, in wh- questions, pitch difference between the last 2nd&3rd syllables reaches its maximum; in v-neg-v questions, the biggest pitch falls take place between the last 1st&2nd and 3rd&4th syllables; in statements, the last three syllables are all influenced and share an even pitch fall. The conclusion contains much randomness and will cause extra troubles in the process of application.

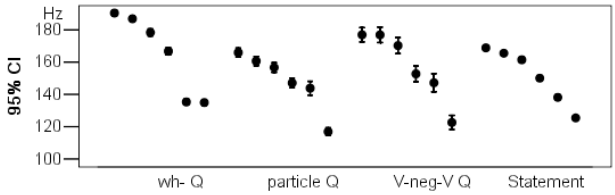


Figure 4. *Final Lowering Carried by Different Syllables on Top-lines of Different Sentences*

Once we changed the basic unit of final lowering from syllable into prosodic word, conclusions of higher consistency were obtained. As illustrated in Fig. 2, final lowering affects exclusively the last prosodic word in all sentences. This conclusion can reflect the prosody rules more precisely, and is more adaptable to intonation modeling and speech synthesis.

4. Supporting Experiment

The complication of tone combination in prosodic words brings about many troubles to our research. First, the pitch information of tone and intonation is hard to be decomposed. Second, many other tone-related effects will take place, such as downstep and neutral tone effect (also considered as a kind of downstep in some works) respectively triggered by low tone T3 and neutral tone T5. Besides, pitch difference between prosodic words could be enlarged. The above factors make the identification of final lowering more difficult.

To make up for such disadvantages, a supporting experiment was conducted to control tone effects. Pitch of the last six syllables was extracted and compared. Section 4.1 aims to discard tone difference by comparing syllables tone by tone. Section 4.2 is designed to evaluate whether our result is influenced by downstep and neutral tone effect. Since final lowering is mainly realized on top-lines according to Section 3, only maximal pitch of syllables is discussed in this section.

4.1. Elimination of Tone Difference

Tone-by-tone comparisons were done to neighboring syllables in order to rule out the differences of tone. Five pairs of syllables (last 1st&last 2nd~last 5th&last 6th) × 5 tones =25 times of comparison were done.

Table 1. *Syllable Pairs with Significant Pitch Difference*

Syllable Comparison	Last 6&5	Last 5&4	Last 4&3	Last 3&2	Last 2&1
Wh-	T2	T2	T1,T4	T1,T2 T4,T5	T1,T2 ,T4
Statement			T1,T4	T1,T2 T4, T5	T1,T3 T4,
V-neg-v				T2,T4	T3
Particle			T1,T3	T2	
Unmarked	n.s.				

According to Table 1, very few significant pitch falls are detected between the former syllables (last6&5, last5&4): only T2 syllables make a difference in wh- questions. Significant differences mainly occur between the last three pairs of syllables. Among them, the highest significance rate lies between the last second pairs, namely between the last 2nd&3rd syllables. To make things clearer, the rate of significant lowering between each pair of adjacent syllables is worked out.

$$Lowering\ Rate = \frac{Numbers\ of\ Significant\ Lowerings}{Five\ Times\ of\ Comparison} \quad (1)$$

Table 2. *Lowering Rate between Each Pair of Syllables*

Syllable Comparison	Last 6&5	Last 5&4	Last 4&3	Last 3&2	Last 2&1
Wh-	0.2	0.2	0.4	0.8	0.6
Statement	-	-	0.4	0.8	0.6
V-neg-v	-	-	-	0.4	0.2
Particle	-	-	0.4	0.2	-
Unmarked	-	-	-	-	-

According to Table 2, significant pitch lowering intensively occurs between the last three pairs of syllables, which approximately corresponds to the last prosodic words. In wh- quesitons, rapid pitch falls exist throught the whole sentence, but they are still most likely to appear between the last 2nd pairs of syllables, with a lowering rate of 0.8. The result proves that the occurrence of final lowering detected in this paper is independent of tone combinations.

4.2. Elimination of Downstep Effect

Since syllables of T3 and T5 mixed in the prosodic words may also result in pitch drops, in this section, sentences involving such syllables were excluded from the sample. Wh- questions and statements were chosen to be tested in this section because of their big quantity (1405 and 2179 respectively).

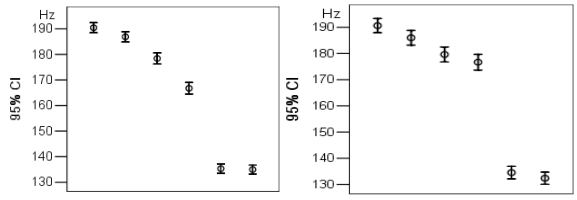


Figure 5. *Pitch Fall Original (left) and after the Exclusion of Sentences with T3 & T5 (right) in Wh- Q*

The left error bar in Fig. 5 shows that final lowering in wh-questions originally occurs between the last 2nd & 3rd syllables. Then, sentences with their last 3rd syllable carrying T3 or T5 are removed. Pitch variation of the qualified sentences was illustrated on the right. It turns out the only change is a pitch rise on the last third syllable, which is due to the delectation of low tones. Nothing happened to the following two syllables. In this case, the abrupt lowering of the last 2nd syllable is not caused by T3 or T5 of the last 3rd syllables in wh- questions.

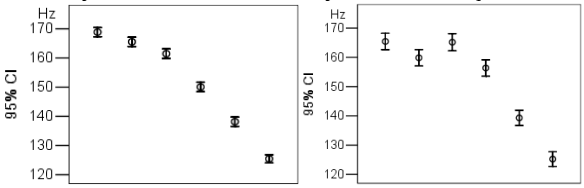


Figure 6. *Pitch Fall Original (left) and after the Exclusion of Sentences with T3 & T5 (right) in Statement*

In statements, significant pitch fall originally occurs at the last 3rd~1st syllables, instead of the last 2nd syllable in wh-questions. Accordingly, the exclusion of sentences was done with more syllables taken into account (the last 4th~2nd syllables). In consequence, a pitch rise appears on the last 4th~2nd syllables due to the exclusion of low tones. The extra pitch fall still occurs in the absence of T3 and T5. Based on the above analyses, we can deduce that our result is not affected by downstep effect triggered by low tones.

5. Discussion

The study brings about a disagreement on the presence vs. the absence of final lowering effect in Chinese Mandarin. Then, a more fundamental problem remaining to be solved is: what leaded to such a disputation? In previous research, designed sentences formed by syllables of the same high tone were used to get a direct observation of intonation, which indicated the absence of final lowering [12]. However, in our study of natural speech, when tone effects were controlled, an additional pitch lowering still appeared on the last prosodic words (or the last two to three syllables). By comparing the two methods, we noticed that the most obvious and substantial factor possibly leading to this disputation is the stimuli/corpus. In fact, the results suggested a potential difference between experimental speech vs. conversational speech, which is instructive to prosody research.

It is reasonable that different methods lead to different conclusions. Sometimes it is hard to label them as absolutely right or wrong, for both materials have their own pros and cons. On the one hand, it is true that designed stimuli can help us get minimal contrast pairs, and high-tone sentences do facilitate our observation of intonation. On the other hand, however, how likely are we to put together syllables of the same tone under natural conditions? Conversely, we tend to combine different tones in our speech intentionally or unintentionally to make it more melodious. That is why experiment stimuli are sometimes accused of being unnatural. Such problem can be fixed by natural corpora, for they contain a large quantity of sentences, extensive topics, diverse tones and flexible syntactic forms. Natural speech has its problems too. Since it involves the information of tone, intonation, stress, prosodic boundary, etc., extra efforts must be made to decompose all kinds of influents and to get what we want.

The experimental control has long been an indispensable part of phonetic research. It serves as an entry point in prosody research, owing to its convenience of simplifying complicated situations and developing new topics. However, we should not be complacent with the existing conclusions from experimental speech. In the next step, these conclusions should be extended into natural speech to get checked, revised or applied. Through this study, we propose a balanced consideration and a combination of the two methods. In our opinion, experimental speech is very essential for developing general topics into specific rules, and then the rules need an additional extension into natural speech for further check-up, revision and application. The combination of experimental control and natural corpus analysis will definitely bring new research topics, novel methods and promising conclusions

6. Conclusion

To support text-to-speech systems with plentiful prosodic details, we analyzed the pitch variation near the end of sentences in a Chinese natural dialogue corpus. A disagreement arises about the presence vs. the absence of final lowering in Chinese, and our results support the presence of final lowering in both questions and statements. Specifically, the effect occurs in wh- questions, v-neg-v questions, particle yes-no questions and statements, but is suppressed by boundary tone in unmarked yes-no questions. Besides, in the light of our data, final lowering is realized mainly on top-lines and affects exclusively the last prosodic word. At the end of the paper, we proposed a combination of experimental speech and natural speech in prosody research, with the former as an entry point of new topics and the latter as an extension of the existing prosodic rules.

In our future work, we will continue to explore the essence of final lowering in Chinese. More linguistic determinants such as accent, boundary tone, and position in the discourse are to be taken into account. Hopefully, the further research would bring out a full discussion on the physiological explication versus the phonological explanation.

7. Acknowledgements

The work was supported by the Beijing Normal University (11-30, 10-12-02), the Fundamental Research Funds for the Central University (2010105565004GK), the State Language Commission 12th Five-Year language research programme (YB125-41), the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027), and partly supported by the Major Program for the National Social Science Fund of China (13&ZD189) and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

References

[1] Beckman, Mary, and Janet Pierrehumbert. "Intonational structure in Japanese and English." Phonology yearbook 3.1 (1986): 5-70.

[2] Cooper, William E., and John M. Sorensen. *Fundamental frequency in sentence production*. New York: Springer-Verlag, 1981.

[3] Gårding, Eva. "Speech act and tonal pattern in Standard Chinese: constancy and variation." *Phonetica* 44.1 (1987): 13-29.

[4] Cohen, Antonie, Rene Collier, and Johan t HART. "Declination: construct or intrinsic feature of speech pitch?" *Phonetica* 39.4-5 (1982): 254-273.

[5] Herman, Rebecca. "Final lowering in Kipare." *Phonology* 13 (1996): 171-196.

[6] Ladd, D. Robert. "Declination: a review and some hypotheses." *Phonology yearbook* 1 (1984): 53-74.

[7] Liberman, Mark, and Pierrehumbert, Janet. "Intonational invariance under changes in pitch range and length." *Language sound structure* 157 (1984): 233.

[8] Lin, Maocan. "Yiwen he chenshu yuqi yu bianjiediao." [Interrogative and Declarative Mood and Boundary Tone]. *Zhongguoyuwen* 4 (2006): 364-376.

[9] Pierrehumbert, Janet. "The perception of fundamental frequency declination." *The Journal of the Acoustical Society of America* 66 (1979): 363.

[10] Shen, Jiong. "Hanyu yudiao gouzao he yudiao leixing." [Intonation structure and intonation types of Chinese]. *Fangyan* 3 (1994): 221-228.

[11] Shen, Xiaonan. *The Prosody of Mandarin Chinese*. Vol. 118. University of California Pr, 1990.

[12] Shih, Chilin. "A declination model of Mandarin Chinese." *Intonation*. Springer Netherlands, 2000. 243-268.

[13] Thorsen, Nina. "Intonation and text in Standard Danish." *Annual Report Institute of Copenhagen* 18 (1984): 185-242.

[14] Umeda, Noriko. "F Declination is situation dependent." *The Journal of the Acoustical Society of America* 68 (1980): S70.

[15] Vaissière, Jacqueline. "Language-independent prosodic features." *Prosody: Models and measurements*. Springer Berlin Heidelberg, 1983. 53-66.

[16] Wang, P., et al. "Putonghua chenshuju zhongde yingao xiaqing he jiangjie." [Declination and Downstep Effect in Declarative Sentences of Chinese Mandarin]. *Zhongguoyuyinxuebao* 3(2012):54-60.

[17] Xu, Yi, and Q. Emily Wang. "What can tone studies tell us about intonation?." *Intonation: Theory, Models and Applications*. 1997.

[18] Xu, Yi. "Principles of tone research." *Proceedings of International Symposium on tonal aspects of languages*. 2006.

[19] Yuan, Jiahong, Chilin Shih, and Greg P. Kochanski. "Comparison of declarative and interrogative intonation in Chinese." *Speech Prosody 2002, International Conference*. 2002.

[20] Yuan, Jiahong. "Perception of Mandarin intonation." Chinese Spoken Language Processing, 2004 International Symposium on. IEEE, 2004.

Intonation Unit Size in Spontaneous Hebrew: Gender and Channel Differences

Vered Silber-Varod [1], *Tal Levy*[2]

[1] The Research Center for Innovation in Learning Technologies, The Open University, Israel
[2] Department of Electrical and Computer Engineering, Rutgers University, USA
vereds@openu.ac.il, tal.tl.levy@rutgers.edu

Abstract

In this corpus-driven research, the question of whether there is a tempo at the Intonation Unit (IU) level, and whether defined IUs differ not only with regard to their pitch contour and boundary tones but also with respect to their phonological size. For this reason, the inventory of syllable size (in terms of segments (phonemes)) and word size (in terms of syllables) was examined, and then each IU category (mainly Terminal vs. Continuous) was measured with respect to the number of syllables and words it contains. Moreover, terminal IU size was also measured with regard to the amount of embedded continuous IUs. Results showed that terminal IUs in spontaneous Israeli Hebrew (IH) do not necessarily consist of embedded continuous IUs. This can be explained due to their massive use as short feedback units in spontaneous speech. Statistical measurements for gender and channel (Face-to-Face vs. telephone conversations) variables were carried with no significance for gender, but with statistically significance for several channel aspects. Last, estimated durational measurements of the IU size are presented.

Index Terms: prosodic unit size, Israeli Hebrew, gender, channel, duration

1. Introduction

When studying speech rate, or speech rhythm, the overall goal is to characterize speech in terms of how many speech units (syllables, prosodic words, etc.) are uttered within a certain time unit (milliseconds, seconds, minutes, etc.). This measure takes into account silent pauses and other types of disfluencies. For example, Syllables-per-second (SPS) is a well know measure for this purpose. For the same reason, analyzing speech unit size in terms of amount of syllables, words, and duration per a certain prosodic unit higher in the hierarchy, can reflects language, or speaker, characteristics. Indeed, since the duration of syllables varies according to their inner-structure [1, p. 55], speech rate is known to be language-dependent, or at least language-type dependent (syllable-timed vs. stress-timed; syllable-rhythm vs. word-rhythm [1, pp. 54-57]. Moreover, in order to characterize prosodic unit at all levels, from syllable to utterance, the prosodic representation "should take into account the heterogeneity and the variations in prosodic constructions encountered in ordinary speech" [2]. There are many applicative aspects of speech rate, such as diagnosing speech disorders, or developing natural speech synthesis. Thus "Obtaining normative data on speaking rate for various groups of speakers is required." [3, p. 131]. For Hebrew, [3] reported that "presently there are no published studies that have directly examined speaking rate among adult Hebrew speakers." [3, p. 131]. Indeed, [3] is a preliminary attempt to provide such data, by quantifying speaking rate within a specific subgroup of speakers – radio newscasters. Nevertheless, it is suggested that speech rate is affected by the communicational setup, specifically, the number of conversation partners was shown to influence rate, such that talking to a single interlocutor is typically performed at a relatively slower rate [4], [5]. Therefore, an examination of the channel aspect as a comparison between Face-to-Face dialogues and Telephone conversations serves this research end. As for gender differences, although there is no linguistic reasoning for such a difference, and indeed no gender differences were found before in Hebrew speech [6], albeit [6] examined the articulation rate and not the speech rate (the latter includes durations of pauses, disfluencies etc.), it was decided to look at this aspect as a baseline examination (and also to use it as a counter-evaluator for the prosodic annotation and segmentation). In the method section (section 2), the prosodic units under investigation are described. In section 3 the analyzed data is described, not only in terms of the corpus type, but also in terms of word and syllable structures. In section 4 the results of gender and channel comparisons are given, and section 5 examines the results and unfolds future research plans.

2. Method

2.1. The prosodic units under investigation

The present paper leans on the prosodic data documented in [7], where the main prosodic unit under investigation is the intonation unit ((IU). For IU definition, see [8, pp. 8-15]. The prosodic annotation process imposes two main types of IU boundary tones that can be defined according to the communicative value of intonation: Terminal (T-) boundary tones and Continuous (C-) boundary tones. A boundary tone was annotated as *T* when the surface intonation signaled that the speaker had "nothing more to say". A boundary tone was annotated as *C* whenever the final tone of the intonation unit signaled "more to come". The recordings were perceptually annotated with a set of prosodic boundaries. Additional validity was achieved using acoustic measurements (see details in [7, pp. 46-55]. This resulted in a back and forth annotation method between perceptual annotation and acoustic measurements, with priority given to major perceptual differences. However, it should be noted that prosodic boundaries were annotated independently of the syntactic structure. The label inventory of prosodic boundaries is as follows. Within the above two types (T- and C- boundary tones), there were two T-boundaries: Terminal Finality (T.), and Terminal Question or Appeal (T?), and five C-boundaries, which were determined according to their tone at the last syllable of the IU: Neutral (C→), Elongated (C:), Rise (C↑), Rise-Fall (C↑↓) and Fall (C↓). Fragmented, or truncated (TR), boundaries were also used in the annotation process.

2.2. The parameters under investigation

In the current research, the following data was extracted for each speaker:

1. No. of words per speaker
2. No. of syllables per speaker
3. No. of pauses (#) per speaker

4. No. of IUs per speaker
5. IU size (word): no. of words per IU
6. IU size (syllable): no. of syllables per IU
7. T- unit size: no. of C-units per T-unit

This will be demonstrated on (1):

heXlaft et ze **T? #** lo ani **C→** lo jodaat ma jihje **T.** (1)

changed-PAST.2SG.F Acc. this **T? #** no I **C→** no know.PARTICIPLE.F what be-FUT.3SG **T.**

'you have changed it**?** no I**,** don't know what is going to be**.**'

For the chunk presented in (1) the following data was extracted:

1. No. of words: 9
2. No. of syllables: 14
3. No. of pauses: 1
4. No. of IUs: 3 (T?, C→, T.)
5. IU size (word): 3/T?, 2/C→, 6/T.
6. IU size (syllable): 4/T?, 3/C→, 10/T.
7. T-unit size (by C-units): 0/T?, 1/T.

The first four parameters are general ones and reflect the speech activity of each speaker.

1. The word parameter counts how many words were transcribed per speaker. This included clitics (function words) that were transcribed separately of their host (the following NP). For example, [ha yeled] 'the boy' was calculated as *two* words and not as a single word, הילד, which is the way the Hebrew orthographic system does.
2. The syllable parameter counts the number of vowels per speaker, following the obligatory presence of a nucleus in the syllable and fact that in Hebrew the nucleus is only a vowel (vowel hiatus, such as in [Raa] see.PST3M 'he saw', was considered as two syllables, but not in cases of word-final diphthongs, where a vowel occurs before the sequence [aX], as in [RuaX] 'wind' (32 such diphthong cases in the corpus)).
3. The pause (marked #) parameter counts pauses above 100ms, per speaker.
4. The IU parameter counts how many IUs (Ts and Cs) were annotated per speaker.
5. IU size (word) parameter counts words per IU, as follows: T-unit size was defined as the number of words from one T-boundary to the next T-boundary. On the other hand, C-unit size was defined as number of words from *any* prosodic boundary to the given C-boundary.
6. IU size (syllable) parameter counts syllables per IU, as follows: T-unit size was defined as the number of syllables from one T-boundary to the next T-boundary. On the other hand, C-unit size was defined as number of syllables from *any* prosodic boundary tone to the given C-boundary.
7. T-unit size parameter refers to the number of minor C-units in major T-units. This parameter reflects rhythm at the intonational unit level.

3. Data

The corpus of this research consists of 19 spontaneous Israeli Hebrew dialogues extracted from The Corpus of Spoken Israeli Hebrew [9]. The recordings were made during 2001-2002. The total duration of analyzed speech is almost five hours. Total number of word types: 4,374; Total number of word tokens: 32,175. The 19 recordings consist of private spoken dialogues in two channels: direct, face-to-face (F2F), dialogues, and distance, telephone (TEL) conversations. Each recording consists of conversations between one core speaker (the "informant", who had the recording equipment on his body for 24 hours) and various interlocutors with whom the speaker interacted on that day. The TEL sub-corpus contains spontaneous phone conversations recorded by the same19 informants. These conversations were part of the 24 hours routine during which the participants recorded themselves. It should be noted that the TEL sub-corpus consists only of the informants' speech, and not their interlocutors'. The total duration of the F2F sub-corpus is 206 minutes; the total duration of the TEL sub-corpus is 83 minutes. Within these 19 recordings, a total of 62 speakers (28 men and 34 women) were transcribed and annotated. The women speech consists of 19,903 words, while the men speech encompasses 12, 272 words. The corpus is heterogeneous in terms of amount of speech per speaker, ranging from 3 to 2,074 words per speaker in the Women sub-corpus; and from 2 to 1,531 words in the Men sub-corpus. The channel groups are also varied. There are only 9 TEL conversations (4 men; 5 women; total of 7,230 words), with speech material ranging from 233 to 2,074 words per speaker; and 61 F2F dialogues (33 women, 28 men; total of 25,107 words), with speech material ranging from 2 to 1,531 words per speaker.

4. Results

The aim of this preliminary research was to investigate if there are: 1. Gender differences; or 2. Channel differences regarding IU size. The data was statistically measured by Mann-Whitney Test, which is a method which has more efficiency on data with non-normal distribution.

4.1. Syllable and word size

In order to measure IU size in terms of words and syllables, it is important to know what the *word* unit size is (in terms of syllables), and what the *syllable* unit size is (in terms of segmental (phoneme) content), in the corpus.

4.1.1. Word size

The statistics of the *word* lengths (in terms of the number of syllables) for the whole database is as follows: The minimum word length is 0 syllables (25 word types; 0.34% in the corpus). This includes cases of interjection such as [m:] 'mm' or truncated words (i.e., false starts). The maximum word length is of 6 syllables. This occurred only once in the loan word [otobijogRafja] 'autobiography'. The average syllables per word (SPW) is 2.3 (standard deviation 0.8). Figure 1 summarizes these statistics, which are close to other studies on word structure in Hebrew child directed speech [10]. Note that in Figure 1 it is demonstrated that monosyllabic words are mostly used in spontaneous IH speech (47.96% tokens), but disyllabic words are four times more varied than monosyllabic ones (black histograms reflects word types). This is an indication to the transcription method carried in the current research, where monosyllabic clitics (i.e., function words such as preposition, definite article, subordinate article) were transcribed as a single orthographic unit, and thus are the most frequent in quantity, but less varied in terms of word-types. Word length statistic was also carried with regard to phoneme per word (PPW). This can be an indication to types (and their relative frequency) of syllable structures in IH. The minimum word-length in terms of number of segments is 1 (18 word-types; 2.47% of corpus). These words include lexemes, such as [o] 'or', interjections and false starts which consist of consonants only. The maximum word length is 13, which

occurred only once in the loan word [otobijogRafja] 'autobiography'. The average word length is 5.6 PPW (standard deviation 1.7).

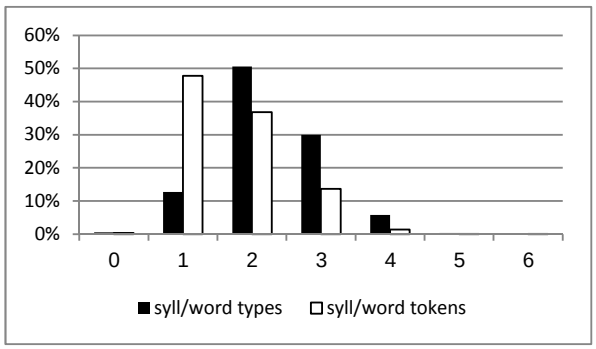


Figure 1: *Syllables per word (SPW) ratios (Types vs. Tokens).*

4.1.2. Syllable size

The average *syllable* length was also calculated using the following method: number of segments per word divided by the number of vowels (i.e., syllables) of that word. For example, in the word [otobijogRafja] 'autobiography' there are 6 syllables. According to the syllable length calculations the syllable length is 2.166 (phonemes per syllable). This is very close to the real average calculation, which is 2, as can be shown from the syllabic division [o.to.bi.jo.gRaf.ja]. There are 4 syllables with 2 phonemes; 1 syllable with 3 phonemes and 1 syllable with a single vowel. By excluding the 25 cases of consonant-based false-starts, the minimal syllable length is 1, meaning the syllable consists of a single vowel, and is related to five different vowel-only words (including [a] and [e] which are interjections). These vowel-only words constitute 2.2% of the corpus. The maximal syllable length (in terms of segments) is 6, and is referred to a single case of the word [dZoRdZ] 'George', which reflects the transcription method of 2 consonants to symbolize the affricate [dʒ]. The average syllable length is 2.6 (phonemes per syllable. standard deviation is 0.7). To sum up, IH speakers speak on average 2.3 syllables per word, an average of 5.6 phonemes per word.

4.2. IU size: gender and channel comparison

In this section, comparison between the two categories: gender and channel, will be analyzed as follows: a comparison with regard to the distribution of IU types; a comparison with regard to IU size in terms of word; Finally, an estimated durational measurements of IUs will be presented.

4.2.1. Gender

In the gender category, no statistically significant differences with regard to IU size were found. The distribution of prosodic boundary tones among females and males is shown in Figure 2. Both groups use the (T.) tone widely, and (T?) to a lesser extent. The T-units' sizes vary from 1-58 words (0-94 syllables). No significant gender differences were found with regard to number of words or syllables, but statistically significant results were found with regard to the gender use of T-boundaries (T and TQ) and C-boundaries ($x^2 = 31.8631; p < .0001$). Figure 2 demonstrates that men use more T(.) and T(?) units than women, who use relatively more C-units (32% of all units, compared to 26% in men's speech). This can imply more intonational variations in women's speech.

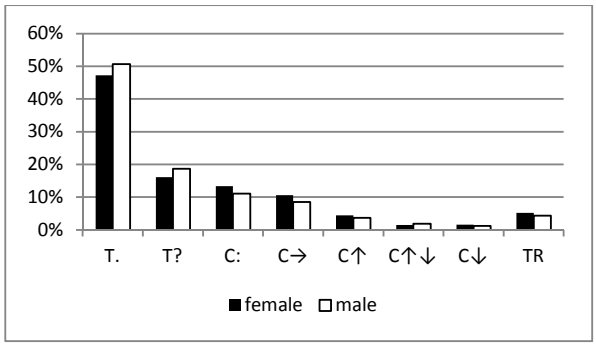


Figure 2: *Distribution of prosodic boundary tones among females and males.*

4.2.2. Channel

The distribution of the prosodic boundary tones in F2F and TEL are shown in Figure 3. The two variables, channel and boundary type (T- or C-) were tested in Chi-square statistics and the results are statistically significant ($\chi^2 = 98.116; p < .0001$), meaning that the two variables show contingency. Moreover, C:-units are relatively more common in TEL.

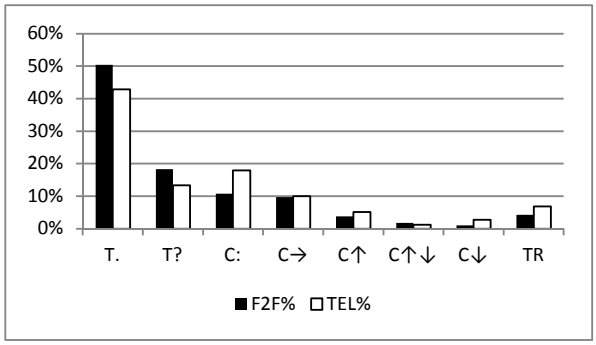


Figure 3: *Distribution of prosodic boundary tones in F2F and TEL.*

Statistically significance difference was found in the channel comparison, where the mean size values of several terminal units were found relatively much more frequent in TEL than in F2F. For example, 1-word T?-unit ($p=0.05$) and 3-words T?-unit ($p<0.01$) were found significant (among others); also 1-word T.-unit ($p<0.01$) and 2-words T.-unit ($p<0.05$). This can be explain as a discourse characteristic of TEL, where feedback are more frequent due to the lack of facial and other visual feedback. The question remains open for much larger units (8, 9, 10, 13 and 15 syllable units) that were found statistically significant. As to the T-size in terms of C-units, at the channel test this parameter was found significant in several cases. First, a difference was found in T-units with no C-units ($p<0.05$). Zero-C-units reflect mostly the short, feedback responses that were explained before as more common in TEL. It is interesting to see that also the T?-unit has some unique use in TEL: 1-C-boundary as well as 3, 8, and 9-C-boundaries were found significantly more frequent ($p<0.01$) in TEL. Perhaps this can be attributed to the opportunity to process longer stretches of speech in TEL, where the interlocutor is more attentive, with comparison to

more lively conversations in F2F dialogues. Significant was found also in the channel comparison of C-unit size. Each of the five C-units had at least one unit-size that demonstrated statistical significance. This correlates with the explanation of longer stretches of speech in TEL, which was mentioned before with regard to terminal units. Two more measurements were carried out on small sub-set of the corpus. First, since the 19 informants of each of the 19 recordings were the main speakers, a subset of Main speakers only was extracted. This subset consists of 8 men and 11 women. An Anova test with repeated measures was carried out, and no significant gender differences were found. For example, p=0.752 on T-unit size. This means that men and women use the same pattern of T-unit sizes (Large amount of 1-word T-units, much less 2-words T-units and declination in use till 15-words T-unit size). It seems that even when reducing the corpus to a less varied speech material, no gender differences are found. Another sub-corpus was extracted for channel pairs of the same speaker. The motivation behind this sub-corpus was to investigate if there is an intra-speaker channel difference. In this sub-corpus only eight speakers out of the 19 main speakers had both TEL and F2F speech material. A T-test was carried out and no significant channel difference was found (p>0.05). This means that speakers use the same patterns, with regard to IU size, in TEL and F2F conversations.

4.3. Syllable duration in various prosodic environments

In order to compare the durational parameter of the syllables in the five continuous boundaries, a pilot study of 22 minutes from TEL (female speaker) and 12 minutes from F2F (male speaker) were segmented manually into syllables. These two recordings were chosen since their acoustic quality meets basic acoustic measurement standards, such as a clear voice and an absence of background noises or speech overlaps. Moreover, in both recordings there is only one interlocutor (compared to other F2F recordings where normally consists of more than two interlocutors). Only the speech of the informant speaker was measured. In TEL, the number of IUs is 610; in F2F, 177. Figures 4 and 5 summarize the durational measurements taken. In this study, the duration of the entire syllable was measured. A threshold to the C: boundary tone was set on a minimum of 230ms. Therefore, the syllables that carry the C: tone were found in both recordings to have the highest mean values (rightmost B&W histograms in Figure 4), while the fluent syllables were found in both cases with the lowest mean values (leftmost B&W histograms in Figure 4).

4.4. Estimated IU size in IH

The estimated unit size was measured as a combination of two variables: 1. In order to avoid the "long tail" bias, we calculated the mean value (syllables) of the most frequent IU sizes; and 2. We used the durational measurements of syllables mentioned in subsection 4.3 above. The mean values of most frequent IU sizes are as follows (almost no difference between the two channels):

T. = 3 (TEL)-3.5 (F2F) syllables per IU
T? = 2.3 (TEL)-3 (F2F) syllables per IU
C→ = 5 syllables per IU
C: = 2.5 syllables per IU
C↑ = 4.5 syllables per IU
C↑↓ = 2 syllables per IU
C↓ = 6.6 syllables per IU

It is demonstrated in the above that the mean C-unit is *longer* than the average T-unit. This is exactly because the calculations took the frequency of use into consideration. Since above 75% of T-units are with no internal C-units, their mean size seems shorter. Second, it should be stressed here that due to the limited size of durational measurements, and to the mixed variables recordings, the results shown on Figures 4 and 5 are within the realm of estimation only. Again, the estimated duration of the two T-units does not include duration of C-units, since above 75% of T-units were without internal C-units.

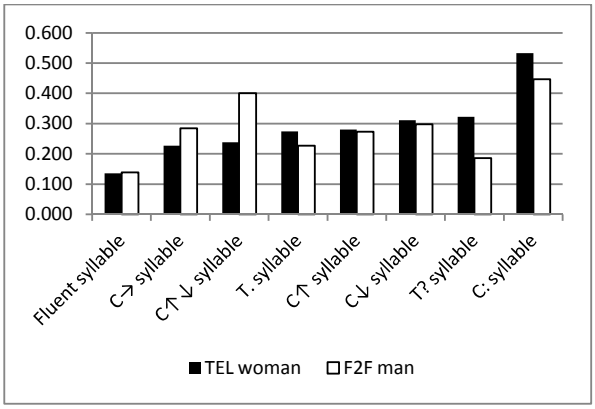


Figure 4: *Syllable duration (seconds) in TEL (woman) and in F2F (man).*

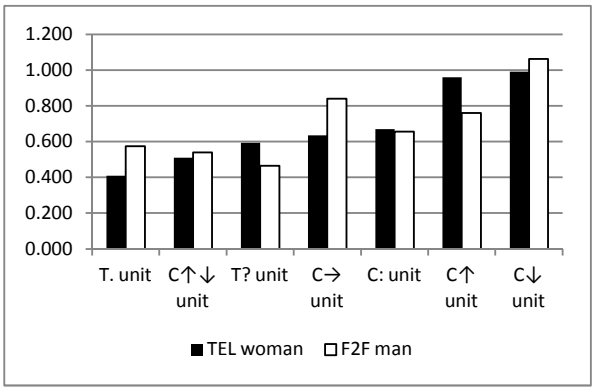


Figure 5: *Estimated IU size (seconds) in TEL (woman) and in F2F (man).*

5. Discussion

This preliminary study highlights the need to examine prosodic unit size in spontaneous IH. Prosodic unit size was examined with regard to syllable and word structures, and sub-units (C-units) in higher prosodic units in the hierarchy (T-units). TEL conversations are characterized by *single-word coherent contour* (T.)-units, while (T?)-units mostly consist of several IUs (with at least two C-units). The syllable durational measurements suggest that in IH, as in AE, "filled pauses differ dramatically from (…) other instances in duration" [11]. With regard to gender and channel differences, in both variables the use of T-boundaries versus C-boundaries was found to be statistically significant. Nevertheless, the duration measurements were carried out on a relatively small portion of the corpus, and syllable duration in spoken IH still needs to be investigated in future research.

6. References

[1] Schmid, S., “Phonological typology, rhythm types and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects”, in A. Ender, A. Leemann and B. Wälchli, (Eds), Methods in contemporary linguistics, 45-68, Berlin: Mouton de Gruyter, 2012.

[2] Lacheret, A., Bordal, G., and Truong, A. “Ch. 11: The prosodic structure”, in A. Lacheret, S. Kahane, and P. Pietrandrea, (Eds.). Rhapsodie: a prosodic syntactic treebank of spoken French. Amsterdam-Philadelphia: John Benjamins Publishing Company, 2014.

[3] Finkelstein, M. and Amir, O., “Speaking Rate among Professional Radio Newscasters: Hebrew Speakers”, Studies in Media and Communication 1(1):131-139, 2013.

[4] Hirose, K., and Kawanami, H., “Temporal rate change of dialogue speech in prosodic units as compared to read speech”. Speech Communication, 36(1):97-111, 2002.

[5] Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J., “Articulation rate across dialect, age and gender”, Language Variation and Change, 21:233-256, 2009.

[6] Amir, O. and Grinfeld, D., “Articulation rate in childhood and adolescence: Hebrew speakers”, Language and Speech, 54(2):225-240, 2011.

[7] Silber-Varod, V., The SpeeCHain Perspective: Form and Function of Prosodic Boundary Tones in Spontaneous Spoken Hebrew, LAP LAMBERT Academic Publishing, 2013.

[8] Izre'el S. and Mettouchi, A., “Representation of Speech in CorpAfroAs: Transcriptional Strategies and Prosodic Units”, in A. Mettouchi, M. Vanhove, and D. Caubet (Eds), Corpus-based Studies of Lesser-described Languages: the CorpAfroAs Corpus od Spoken AfroAsiatic Languages. Amsterdam-Philadelphia: John Benjamins, to appear.

[9] COSIH: The Corpus of Spoken Israeli Hebrew <http://humanities.tau.ac.il/~cosih/english/>

[10] Segal, O., Nir-Sagiv, B., Kishon-Rabin, L., and Ravid, D., “Prosodic patterns in Hebrew child directed speech”, Journal of Child Language, 36(3):629-656, 2009.

[11] Shriberg, E. “To errrr' is human: ecology and acoustics of speech disfluencies”, Journal of the International Phonetic Association 31(1):153-169, 2001.

Acoustic Cues to Tone and Register in Bai: Adult Baseline Data

Allison Benner, John H. Esling

Department of Linguistics, University of Victoria, Canada

abenner@uvic.ca, esling@uvic.ca

Abstract

This paper presents the results of a study of the acoustic cues associated with the tense/lax distinction in Bai, a Tibeto-Burman register tone language spoken in Yunnan, China. The purpose of the paper is to provide baseline adult data for comparison with infant speech in an acoustic study of infants' acquisition of Bai register tones in the second and third years of life. The results show that among adults, F0, F1, and spectral tilt combine to create the tense/lax contrast in Bai. While these three cues tend to be correlated, individual speakers differ in their use, particularly spectral tilt. The patterns in this study suggest that as Bai infants acquire tones in the second and third years of life, their utterances are likely to become structured around these three acoustic cues in previously unattested ways that exemplify the complex interaction between universal physiological and developmental tendencies and the ambient phonological tone system of Bai.

Index Terms: Bai, tone, acoustic cues, laryngeal constriction, first language acquisition

1. Introduction

Bai is a Tibeto-Burman language spoken by approximately 1.6 million speakers in the three dialect regions of Dali, Jianchuan, and Bijiang in Yunnan, China [1]. As shown in Table 1 below, Bai includes eight tones that are phonologically classified as lax or tense, based on pitch and/or phonation type. All these tones have contrastive nasal variants, with the exception of the rising tone. Tense tones are produced with more laryngeal constriction than lax tones, though tones in both registers may be produced with differing degrees of laryngeal constriction. For example, the tense register includes differing degrees of harshness. Tense tones 55+ and 33+ are often produced with tight or 'pressed' phonation, 31+ with harsh voice, and 21 with aryepiglottic trilling. Lax tones 55 and 33 are produced with modal voice, and lax tone 31 with breathy voice. Tone 35 varies between harsh and modal phases.

Table 1. *Bai register tone system.*

	Lax	Tense
high level	55	55+
mid level	33	33+
mid falling	31	31+
low falling		21
Rising		35

While the literature includes a description and classification of the Bai tone system [2] [3] [4], and while detailed studies of the laryngeal articulatory phonetic realization of these tones are available [1] [5], no acoustic studies on Bai tones exist. This study aims to find general trends in the acoustic cues to these tones. As such, the study contributes to the growing recent literature on acoustic cues to register tones [6] [7] [8] [9] and will be used for baseline comparison in our ongoing study of tone acquisition among infants acquiring Bai [10] [11] [12] [13].

2. Methodology

2.1. Data

To examine the acoustic cues to the tense/lax distinction, data were selected from two pre-existing sources: (1) the audio files that accompany the Phonetic Database article for Bai [14], which comprise recordings of a male native speaker of Bai; and (2) field recordings of six native speakers of Bai (5 female, 1 male) made in Yunnan in 2008. The latter recordings were made to find examples of vowels illustrating all eight tones in the paradigm on a single monosyllabic word, and to verify whether speakers made these contrasts in the same manner observed in the speaker in the Phonetic Database. There are some gaps in the data, because some speakers did not recognize words in the paradigm.

The 291 tokens included in this analysis are monosyllabic words containing the vowels /æ/, /a/, /i/, /e/, /u/, and /o/ spanning, where possible, all the non-nasalized tones in the register tone paradigm. In all, the data include 111 tokens of tones in the lax register, and 180 in the tense register. Table 2 shows the distribution of the data among the eight tones. Within the available data, it was not possible to balance the number of tokens by speaker or tone. Cells in Table 2 with at least 7 tokens include recordings of all 7 speakers in the study. In cells with large numbers of tokens, most utterances represent the single male speaker in the Phonetic Database, whose utterances comprise approximately one-third of the total data (102 tokens).

Table 2. *Data included in the analysis, by vowel and tone.*

Tone	/æ/	/a/	/i/	/e/	/u/	/o/	Total
55	7	8	11	3	7	7	43
55+	7	5	9	7	7	1	36
33	7	7	9	1	6	2	32
33+	7	23	10	9	7	8	64
31	7	7	9	1	7	5	36
31+	7	9	3	7	1	5	32
35	1	--	7	1	3	1	13
21	7	10	3	--	7	8	35
Total	50	69	61	29	45	37	291

2.2. Acoustic Analysis

Three acoustic cues were measured for each of the tokens: F0, F1, and spectral tilt. F0 was measured at the beginning, middle, and end of each token. F1 was measured in the middle of each token. Finally, spectral tilt measures, including H1-H2,

H1-A1*, H1-A2*, and H1-A3*, were taken from the middle of each token. All measurements were taken using Praat [15]. Average values for F0, F1, H1-H2, H1-A1*, H1-A2*, and H1-A3* were calculated for individual speakers and for the group as a whole for each of the eight tones, and for the lax and tense registers. Because of the uneven distribution of the data samples between individuals and tones, no statistical tests were performed.

3. Results

3.1. Fundamental Frequency

Tables 3 and 4 below show the average F0 at the beginning and end of each tone for male and female speakers, respectively. Figure 1 illustrates the average F0 values and direction of F0 movement for each tone across speakers. As shown, the eight contrasting tones in the paradigm are distinguished by pitch. Tense tones tend to be higher in pitch than their lax counterparts. For example, 55+, 33+, and 31+ are higher in pitch than 55, 33, and 31, respectively. The pitch range of the lax 31 tone and the tense 21 tone closely overlap (though they differ maximally in phonation type). All speakers follow this pattern. While lax level tones 33 and 55 are produced in different ways by individual speakers (as slightly rising, slightly falling, or small rise-falls), they are consistently produced with more level pitch than corresponding tense tones 33+ and 55+, which all speakers produce with noticeable falls.

Table 3. *Average F0 of tones (Hz), male speakers.*

	55	55+	33	33+	31	31+	35	21
Start	230	268	179	208	166	224	138	160
Mid	236	269	177	196	138	174	191	135
End	237	248	177	193	107	116	222	110

Table 4. *Average F0 of tones (Hz), female speakers.*

	55	55+	33	33+	31	31+	35	21
Start	270	291	249	272	237	288	245	228
Mid	275	276	248	260	205	247	260	206
End	267	227	231	249	160	179	296	169

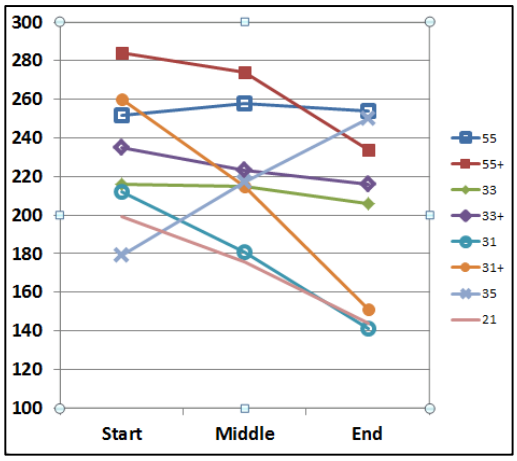


Figure 1: *Average F0 of tones, male and female speakers.*

3.2. F1 Frequencies

The F1 frequency is, on average, higher for tones produced in the tense register (55+, 33+, 31+, 35, and 21) than for tones produced in the lax register (55, 33, and 31) for all vowels studied, reflecting the greater laryngeal constriction in tense tones. Average F1 frequencies for vowels in the tense and lax registers are shown in Figure 2 below.

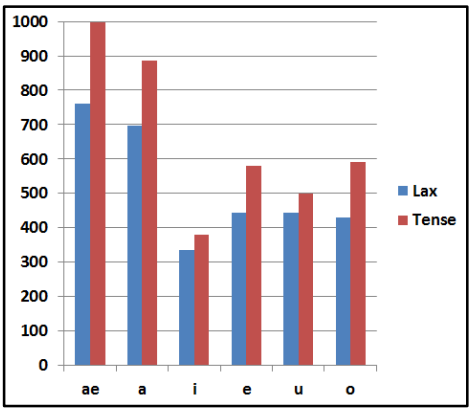


Figure 2: *Average F1 frequencies, all speakers.*

Within this pattern, F1 frequencies differ considerably across tones. Table 5 shows the average F1 for each vowel in the study for each tone. As shown, 55+, 33+, and 31+ have a higher average F1 than 55, 33, and 31, respectively, with the exception of the 33+/33 contrast for /i/. These contrasts in F1 help to preserve distinctions between pairs of tones that are phonologically contrastive and that are potentially confusable on the basis of pitch. Among the tones, across all vowels except /u/, the average F1 for 21 is the highest, and the average F1 for 31 is the lowest, in keeping with the fact that these tones differ maximally in laryngeal constriction: as noted earlier, 21 is produced with harsh voice and aryepiglottic trilling, while 31 is produced with breathy voice.

Table 5. *Average F1, by vowel and tone, all speakers.*

	/æ/	/a/	/i/	/e/	/u/	/o/
55	807	733	357	--	437	457
55+	958	824	405	616	569	593
33	782	681	337	399	400	402
33+	985	867	326	558	442	505
31	694	669	308	392	408	403
31+	1000	868	361	582	493	579
35	875	--	415	490	676	473
21	1056	964	422	--	474	700

F1 frequencies also differ between speakers. Figure 2 depicts the average F1 of /æ/ for speakers 1 to 7 for all tones, except 35 (there was only one recording of this tone in the data). This vowel is chosen as an illustration because the data are balanced across speakers, because tense/lax F1 values occur within a wider range for low vowels, and because the pattern depicted reflects the general trend found for most other vowels in the study. As shown, with the exception of speaker 6, all

speakers produce higher F1 values for tense tones than for their lax counterparts (55+, 33+ and 31+ have higher F1 frequencies than 55, 33, and 31, respectively). For 4 of the 7 speakers, 31 has the lowest F1, and 21 the highest. For some speakers, lax tones 33 or 55 have a lower F1 than lax 31, and/or tense 31+ has a higher F1 than tense 21. Across all speakers, with the exception of speaker 6, the F1 frequencies of tense tones 21 and 31+ are both higher than the F1 of lax (breathy) 31.

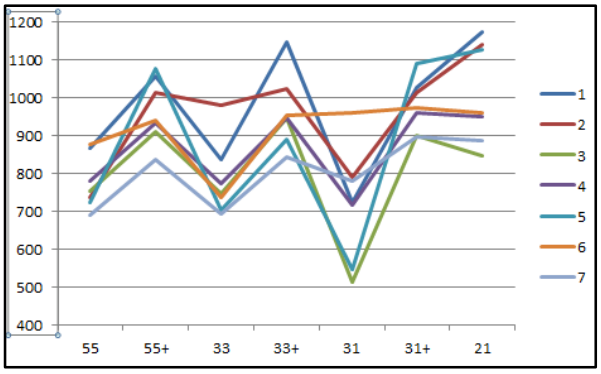


Figure 2: *Average F1 for /æ/, speakers 1-7.*

3.3. Spectral Tilt

As shown in Table 6, average spectral tilt values for the lax register are lower than for the tense register across all six vowels studied and across all four measures of spectral tilt employed (H1-H2, H1-A1*, H1-A2*, and H1-A3*), with the exception of H1-H2 for the vowels /a/ and /i/. These differences reflect the greater laryngeal constriction in the tense tones compared to the lax tones.

Table 6. *Spectral tilt by vowel, tense and lax registers.*

	H1-H2		H1-A1*		H1-A2*		H1-A3*	
	T	L	T	L	T	L	T	L
/æ/	2	7	-10	3	-8	5	4	19
/a/	4	3	-9	-2	-8	3	9	21
/i/	8	7	-1	0	12	18	12	18
/e/	-6	3	-8	3	1	13	6	18
/u/	-2	-1	-3	-2	8	9	27	29
/o/	-3	0	-7	-2	1	15	24	31

As with F1 values, spectral tilt measures varied among tones and among individual speakers. As an illustration of the general trend, Table 7 below shows the average H1-A3* values for each vowel and tone for the seven speakers. Across vowels, H1-A3* is higher for 55, 33, and 31 than for 55+, 33+, and 31+, respectively, with the exception of the 33/33+ contrast for /u/. Among the eight tones, H1-A3* is highest for lax tone 31 for most vowels (/æ/, /i/, /e/, and /u/) and lowest for tense tone 31+ for most vowels (/æ/, /i/, /e/, and /o/). The contrasts in H1-A3* values are most distinct between tense/lax pairs across the tonal paradigm for low vowels /a/ and /æ/, and less so for high vowels /i/ and /u/. These differences likely reflect the degrees of laryngeal constriction to which the production of each vowel is inherently susceptible. More specifically, there may be less latitude to produce tense/lax contrasts with high vowels, which are inherently produced with a more raised tongue and, consequently, a greater degree of pharyngeal expansion than low vowels, which are inherently more susceptible to tongue retraction, larynx raising, and laryngeal constriction [16] [5] [17].

Table 7. *H1-A3*, by vowel and tone, all speakers.*

	/æ/	/a/	/i/	/e/	/u/	/o/
55	16	17	16	17	29	28
55+	5	9	11	8	24	24
33	20	26	16	16	26	29
33+	6	11	15	5	27	28
31	22	22	22	22	33	36
31+	2	8	5	5	27	21
35	11	--	14	11	21	26
21	4	7	8	--	30	20

For all measures of spectral tilt, values range widely across individuals, likely reflecting the different envelope within which speakers produce phonatory contrasts. As an illustration of this point, Figure 3 below shows average H1-A3* values for speakers 1-7 for the vowel /æ/. As can be seen, with the exception of speaker 1, all speakers make a consistent contrast between tense and lax pairs of tones, though some speakers appear to make those contrasts within a more or less constricted setting, and within a narrower or wider range of constriction. Speaker 3, for example, consistently distinguishes between the tense/lax pairs, but does so within a vocal setting that is relatively unconstricted, compared to speakers 4 and 7, who produce these same contrasts within a relatively constricted setting. Of the seven speakers, speaker 5 produces the contrasts within a very wide range, while speaker 1 (with the exception of the 55/55+ contrast) produces most of the contrasts within a very narrow range.

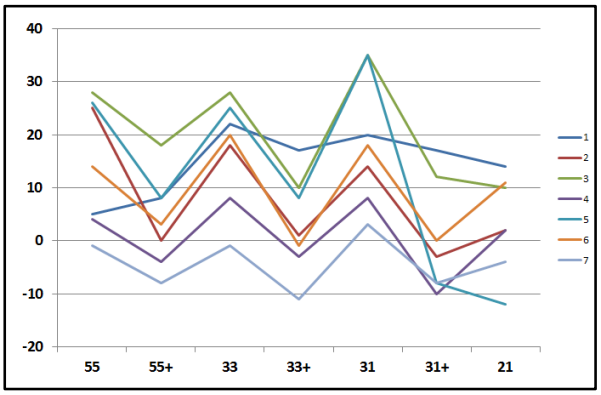


Figure 3: *H1-A3* for /æ/, speakers 1-7.*

On the whole, spectral tilt values are inversely correlated with F1 values, reflecting the relationship of these measures to laryngeal constriction (high spectral tilt values occur in less constricted sounds, which tend to have low F1 values; low spectral tilt values occur in more constricted sounds, which tend to have higher F1 values). In line with this tendency, for all seven speakers, the highest spectral tilt values correspond

to lax tones, and the lowest to tense. Similarly, for all speakers, the highest F1 values correspond to tense tones, and the lowest to lax tones. However, there are individual differences in the ways that spectral tilt and F1 values pattern with specific tones. For example, for speaker 5, the highest and lowest F1 values for /æ/ are for tones 21 and 31, respectively, and the lowest and highest spectral tilt values are for those same tones. However, for speaker 1, whose formant values pattern with speaker 5's (highest for 21, lowest for 31), the lowest and highest spectral tilt values are for tones 55+ and 33, respectively. Thus, speakers differ in the ways that they combine the different cues to the tense/lax contrast, though the relationships between the cues are far from arbitrary.

4. Discussion and Conclusion

In summary, this exploratory study found that Bai speakers distinguish between tense/lax tones on the basis of three acoustic cues: F0, F1, and spectral tilt. Speakers show the greatest homogeneity in their use of F0 as a cue to individual tones, and to tense/lax pairs of tones: across all seven speakers, tense tones have higher F0 than their lax counterparts. Speakers also systematically employ F1 to distinguish between tense and lax tones. In general, across speakers, tense tones have a higher F1 than lax tones. While there are individual differences, the most constricted tones (31+, 21) usually have the highest F1, and the least constricted (31) usually have the lowest. Similarly, for all spectral tilt measures employed (H1-H2, H1-A1*, H1-A2*, and H1-A3*), tense tones tend to have lower spectral tilt values than lax tones, and the most constricted tones (31+, 21) usually have the lowest values, and the least constricted (31) the highest. However, while all speakers use these three cues, individuals differ in the way they distribute these cues among the tones, and in the degree and range of laryngeal constriction they use in producing the tones. These findings suggest that while the phonological system strongly influences the use of acoustic cues, individual portrayals of that phonology differ.

It is possible that given the wide range of articulatory choices at a speaker's disposal in producing laryngeal constriction (i.e. the differing combinations of ventricular fold incursion, larynx raising, tongue and epiglottis retraction, aryepiglottic fold compression, and pharyngeal narrowing, and the differing qualities produced by such combinations, as described in [5]), there is greater scope for individual variation in the production of contrasts based on this feature compared to contrasts that are based exclusively on pitch or vowel formants. A Bai speaker could opt, for example, to produce tense tones with differing degrees of larynx raising, a choice that would tend to result in higher pitch for tense tones relative to their lax counterparts, as well as a more open jaw setting, which would raise F1. This strategy might be sufficient to produce the tense/lax contrast, while seldom resulting in audible harshness or, as a result, low spectral tilt measurements. Another speaker could adopt a more constricted setting across all tones, but employ a greater degree of laryngeal constriction on tense tones, resulting in audible harsh voice and/or aryepiglottic trilling on some tones. Such differences in articulatory strategies may account for the variability in individuals' spectral tilt measurements, while also accounting for the tendency for the F0 and F1 results to pattern similarly across speakers.

The primary intention of this study is to generate baseline adult data for comparison with infant utterances in a study of Bai infants' acquisition of the register tone system in the second and third years of life. This study has several implications for infants' acquisition of the relevant acoustic cues (F0, F1 and spectral tilt). Current research on first language acquisition suggests that in the first year of life (especially the first six months), infants can distinguish between all speech sounds that are employed in languages of the world. Towards the end of the first year, infants lose this sensitivity in favour of developing the ability to distinguish between the sounds used in their ambient language, a result that has been found for non-tonal [18] [19] and tonal [20] [21] [22] contrasts. Thus, like infants learning other tone languages, Bai-learning infants are likely to remain sensitive to pitch differences that are used to distinguish tones. However, in learning Bai, a register tone language, infants also need to remain sensitive to the relationship of these pitch differences to the inter-related cues of vowel quality and spectral tilt, which are used in varied ways by adult speakers of the language. Currently, no research exists on the acquisition of register tone languages, so how Bai infants achieve this perceptual organization remains an open question. Our auditory and acoustic studies of Bai infants' utterances in the first year of life [10] [11] [12] show that Bai-learning infants' babbling includes a higher incidence of, and greater variability in, laryngeal constriction than the babbling of English-learning infants. It remains to be seen, however, how this feature develops in the second and third years of life, and how it interacts with the infants' use of pitch and vowel quality.

Research on infants' and young children's tone production demonstrates that tone acquisition is a protracted developmental process. While infants begin to produce tones in the second and third years of life [23][24][25][26][27][28], their production is not fully adult-like [29][30][31][32], and shows little improvement in accuracy up to the age of five [33]. The latter studies are all based on infants' production of pitch. Our own study will focus on the previously unexplored relationships between pitch, vowel quality, and spectral tilt in the acquisition of register tones. It is likely that Bai infants, like learners of other tone languages, will not fully acquire the Bai tone system until at least the age of five. However, it will be of considerable interest to see which of the eight contrasting tones Bai infants begin to produce in the second and third years of life, and which of the relevant acoustic cues (F0, F1, and spectral tilt) they begin to correlate in their acquisition of the tense/lax contrast.

5. Acknowledgements

The authors would like to thank the Social Sciences and Humanities Research Council of Canada for their support of this research. We would also like to thank Dr. Jerold Edmondson for the use of field recordings of Bai made in Yunnan, China in 2008.

6. References

[1] Esling, J. H. and J. A. Edmondson, "The laryngeal sphincter as an articulator: Tenseness, tongue root and phonation in Yi and Bai", in A. Braun and H. Masthoff [Eds.], Phonetics and its Applications, 38-51, Franz Steiner Verlag, 2002.

[2] Xu, L. and Zhao, Y., "Baiyu gaikuang" (in Chinese), in Zhongguo Yuwen, 321-325, 1964.

[3] Xu, L. and Zhao, Y., "Baiyu jianzhi" (in Chinese), Beijing, 1984.

[4] Edmondson, J. and Li, S., “Voice quality and voice quality change in the Bai language of Yunnan Province”, in Q. Dai [Ed.], Linguistics of the Tibeto-Burman Area, 17(2):49-68, 1994.
[5] Edmondson, J. A., and Esling, J. H., “The valves of the throat and their functioning in tone, vocal register and stress: Laryngoscopic case studies”, Phonology 23:157–191, 2006.
[6] Kuang, J., “Production and perception of the phonation contrast in Yi”, Unpublished manuscript, University of California, Los Angeles, 2011.
[7] Brunelle, M., “Dialect experience and perceptual integrality in phonological registers: Fundamental frequency, voice quality and the first formant in Cham,” J. Acoust. Soc. Am. 131:3088–3102, 2012.
[8] Esposito, C., “An acoustic and electroglottographic study of White Hmong tone and phonation”, J. Phonetics, 40(3):466-476, 2012.
[9] Garellek, M., Keating, P., Esposito, C., and J. Kreiman, “Voice quality and tone identification in White Hmong, J. Acoust. Soc. of Am., 133(2):1078-1089, 2013.
[10] Benner, A., Grenon, I., & Esling, J.H., “Infants’ phonetic acquisition of voice quality parameters in the first year of life”, Proc. of the 16th Intl. C. of the Phonetic Sciences, Saarbrücken, Germany, www.icphs2007.de, 2007, accessed on 28 Dec. 2013.
[11] Benner, A., “Production and perception of laryngeal constriction in the early vocalizations of Bai and English infants”, Proc. Cdn. Acoust. Assn. Annual Conf., 2010.
[12] Benner, A., & Grenon, I., “The relationship between laryngeal constriction and vowel quality in infants learning English and Bai,” Proc. of the 17th Intl. C. Phonetic Sciences, Hong Kong, http://www.icphs2011.hk, 2011, accessed 29 Dec. 2013.
[13] Esling, J.H. & Benner, A., “Laryngeal-pharyngeal ontogeny: Speech in the first several months”, Paper presented at the Intl. Child Phonology Conf., Minneapolis, MI, 2012.
[14] Esling, J.H., “University of Victoria phonetic database (version 4.0)”, Victoria, BC & Lincoln Park, NJ, 1999.
[15] Boersma, P. & Weenink, D., “Praat: Doing phonetics by computer [computer program], (version 5.3.51)”, 2013, http://www.praat.org/, retrieved 2 June 2013.
[16] Esling, J.H., “There are no back vowels: The laryngeal articulator model”, Cdn. J. of Ling., 50:13-44, 2005.
[17] Moisik, S.R. & Esling, J.H., “Evaluating the vowel space effects of larynx height using laryngeal ultrasound”, Cdn. Acoust., 39:180-181, 2011.
[18] Werker, J.F. & Curtin, S. (2005). “PRIMIR: A developmental framework of infant speech processing”, Lang. Learning & Dev., 1(2):197-234, 2005.
[19] Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P., “Infants show a facilitation effect for native language phonetic perception between 6 and 12 months”, Developmental Science, 9(2):F13-F21, 2006.
[20] Mattock, K. & Burnham, D., “Chinese and English infants’ tone perception: Evidence for perceptual reorganization”, Infancy, 10(3):241-265.
[21] Mattock, K., Molnar, M., Polka, L., & Burnham, D., “The developmental course of lexical tone perception in the first year of life”, Cognition, 106(3):1367-1381, 2008.
[22] Yeung, H., Kenny, H., & Werker, J.F., “When does native language input affect phonetic perception? The precocious case of lexical tone”, J. of Memory and Lang., 68(2):123-139, 2013.
[23] Clumeck, H., “Studies in the acquisition of Mandarin phonology”, Unpublished doctoral diss., U. of California, Berkeley, 1977.
[24] Clumeck, H., “The acquisition of tone”, in G.H. Yeni-Komshian, J.F. Kavanaugh, & C.A. Ferguson [Eds.], Child phonology: Vol. 1, Production, 257-275, NY, Academic Press, 1980.
[25] Li, C.N. & Thompson, S.A., “The acquisition of tone in Mandarin-speaking children”, J. of Child Lang., 4:185-199, 1977.
[26] Zhu, H. & Dodd, B, “The phonological acquisition of Putonghua (Modern Standard Chinese)”, J. Child Lang., 27:3-42, 2000.
[27] Ota, M., “The development of lexical pitch accent systems: An autosegmental analysis”, Cdn. J. Ling., 48(3/4):357-383, 2003.
[28] To, C.K.S., Cheung, P.S.P., & McLeod, S., “A population study of children’s acquisition of Hong Kong Cantonese consonants, vowels, and tones”, J. Speech, Lang, & Hearing Research, 56(1):103-122, 2013.
[29] Wong, P., Schwartz, R., & Jenkins, J., “Perception and production of lexical tones by 3-year-old Mandarin-speaking children”, J. Speech, Lang., & Hearing Research, 48(5):1065-1079, 2005.
[30] Yang, J., & Lee, H.-T., “Lexical variation and rime-tone correlation in early tonal acquisition: A longitudinal study of Mandarin Chinese”, Int. Symp. Tonal Aspects Lang., 126-131, 2006.
[31] Wong, P., “Acoustic characteristics of three-year-olds’ correct and incorrect monosyllabic Mandarin lexical tone productions”, J. Phonetics, 40(1):141-151, 2012a.
[32] Wong, P., “Monosyllabic Mandarin tone productions by 3-year-olds growing up in Taiwan and the United States: Interjudge reliability and perceptual results”, J. Speech, Lang. & Hearing Research, 55(5):1423-1437, 2012b.
[33] Wong, P., “Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan”, J. Acoust. Soc. Am., 133(1):434-443, 2013.

Word accent and intonation in Baltic

José I. Hualde [1], *Tomas Riad* [2]

[1] Department of Spanish, Italian and Portuguese and Department of Linguistics, University of Illinois at Urbana-Champaign, U.S.A.
[2] Department of Swedish Language and Multilingualism, Stockholm University, Sweden
jihualde@illinois.edu, tomas.riad@su.se

Abstract

We examine the realization of word accent contrasts in Standard Latvian and East Aukštaitian Lithuanian across intonational contexts. In our Latvian data the contrast is manifested as level vs. falling pitch in most contexts, in addition to a durational difference. In Aukštaitian Lithuanian, instead, differences in vowel quality and duration cue the lexical contrast in the nuclei that we examine. While Latvian retains a tonal contrast, in Aukštaitian Lithuanian it has been replaced with a combined segmental/quantitative contrast, where the so-called circumflex tone corresponds to relatively shorter duration and, in the case of diphthongs, centralized quality in the first half. We discuss the implications of these findings for further typological work.

Index Terms: Lithuanian, Latvian, word accent, prosody

1. Introduction

The Baltic languages, Latvian and Lithuanian, are traditionally described as possessing lexical tone in some syllables, although the nature of the lexical contrast, as being based on tone or other features remains controversial. We will use the more phonetically neutral term 'word accent'. Here we investigate the realization of word accent in Latvian and Lithuanian in different intonational contexts, going beyond previous work in this respect. To the extent that existing phonological analyses [1] [7] may be based on an incomplete understanding of the phonetic facts, they may be in need of revision.

Our data are based on utterances elicited from educated native speakers by means of a questionnaire and containing minimal or near minimal pairs. Given the controversies regarding the nature or even the existence of the contrasts that we are investigating, we decided that using a maximally explicit methodology for data elicitation was appropriate. Naturally, this methodology will tend to produce somewhat hyperarticulated realizations of any existing phonological contrasts. All participants were either faculty or students at Stockholm University at the time of the experiment.

2. Latvian

2.1. Status Quaestionis

In Latvian, stress generally falls on the initial syllable of the word. If the stressed syllable is heavy (that is, if it contains a long rhyme: long vowel, diphthong or a postvocalic consonant), in some dialects, a three-way contrast in accent type in syllables is said to obtain. The three accents have been described as level (also known as "even" or "circumflex"), falling (or "acute") and glottalized (or "broken"). In linguistic work the level accent is represented with a tilde, e.g. [ã], the falling accent with a grave accent mark, [à] and the glottalized accent with a circumflex accent mark, [â]. An acoustic characterization of these three accents in words obtained in medial position in a carrier phrase is found in [5] [6]. In this context, in terms of tonal contour, the level accent has a high level, slightly rising F0 throughout the stress syllable, whereas the other two accents, falling and broken, are characterized by an F0 peak towards the middle of the syllable, followed by a sharp drop in F0 in the second mora. The broken accent is distinguished from the falling accent in showing glottalization in the second mora. There are no accentual contrasts in light syllables.

In most dialects, however, there is only a two-way word-accent distinction. In particular, contemporary Standard Latvian, the variety that we examine here, distinguishes only "even" from "non-even" accent (the latter having resulted from the historical merger of the falling and broken accents, [7]).

2.2. Methods

We constructed a list of 15 mini-dialogues in order to elicit the same target word in different intonational contexts (e.g. Contrastive focus: Did Peter say apple? No, Peter say TARGET; Postfocus: So, Peter wrote TARGET? No Peter said TARGET; etc.). Target words were members of minimal pairs. Here we report on one of these pairs: *zāle* [zãle] (even, E) 'hall' vs. *zāle* [zâle] (non-even, N) 'grass'. Participants were asked to read the dialogues (both the context and the sentence containing the target word) naturally and at a comfortable speed. To avoid confusion between members of each minimal pair, all 15 mini-dialogues involving the same target word were presented together. Participants were three speakers of Standard Latvian, from the central dialectal area: one male from Riga and two females, from Seldus and Valmiera.

2.3. Results

Consistent with previous description, the even accent was realized as relatively flat high pitch over the heavy stressed syllable in contexts where the test word is under focus (Fig. 1, left). In postfocal and prefocal position, however, the even accent was realized as flat *low* pitch (Fig. 2). It thus appears that what characterizes the even accent across contexts is the absence of tonal differences between the beginning and the end of the syllable, with a relatively flat pitch throughout. The crucial aspect appears to be that both morae in the heavy stressed syllable receive the same tone, be it high or low.

In contrast, the non-even accent is produced in most contexts with a pitch fall from the first to the second mora (Fig. 1, right). This is except in final position in questions, where there may be a final rise.

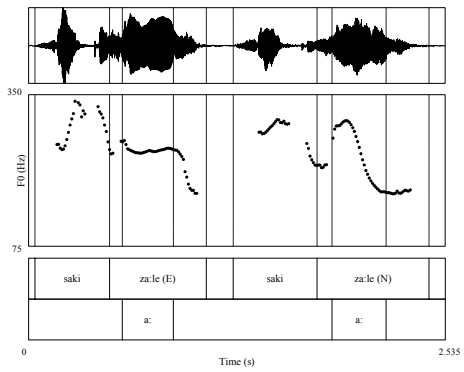


Fig. 1. *saki zāle* [zāle] (E) 'say hall' vs. *saki zâle* [zâle] (N) 'say grass'. Female speaker from Saldus.

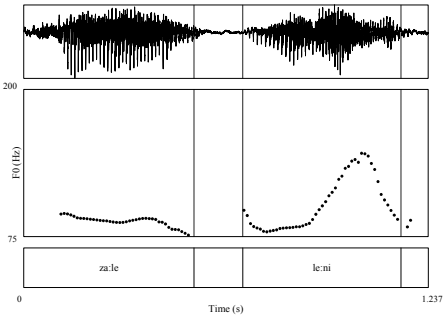


Fig. 2 (*Vai Peteris teica*) *zāle LENI*? '(Did Peter say) hall SLOWLY?' Even accent realized as low flat low pitch. Male speaker from Riga.

The contrast between the F0 contours can be captured numerically as the difference between F0 maximum and minimum during the stressed nucleus. For our two female speakers, average values including all contexts were 83 Hz under E accent and 123 Hz under N accent (t-test: t = –2.6895, df = 63.269, p = 0.009138). For our male speaker this value could not be accurately calculated for a large number of tokens because of the presence of glottalization during the second half of the stressed nucleus under non-even accent. Glottalization was realized either as a glottal stop (Fig. 3) or as irregular pulsation.

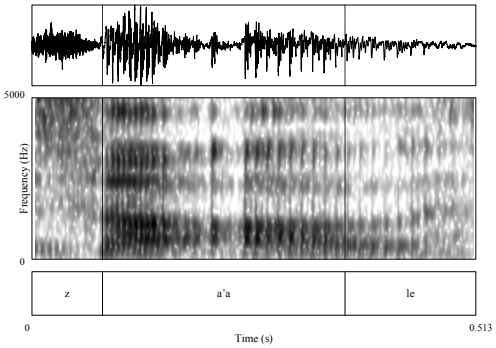


Fig. 3. *zâle* (N) 'grass'. Example of non-even accent accompanied by glottal stop. Male speaker from Riga.

For this speaker we have analyzed a second minimal pair, *loks* [luõks] 'spring onion' vs. *logs* [luôks] 'window'. A measure of the presence of glottalization can be obtained by calculating the percentage of the stressed nucleus with regular pulsation. This is shown in Fig. 4, with the results returned by the function VoicedFrames in PRAAT [2] for the stressed nucleus of E and N words. For both N words (*logs*-N, *zale*-N) glottalization is common, but not systematic. That is, for this speaker the neutralization of the falling and broken or glottalized accent in Standard Latvian (see [7]) appears to have resulted in an allophonic range that includes different degrees of glottal constriction. Our female speakers did not produce glottalization.

A significant difference in duration was also found for all speakers, with E words being realized with a longer vowel (means, *zãle* E = 335 ms, *zâle* N = 288 ms, t = 5.7486, df = 97.287, p = 1.035e-07)

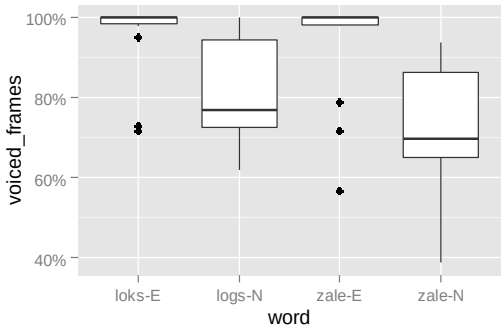


Fig. 4. 'Voiced Frames' in stressed nucleus for E and N words, male speaker from Riga.

2.4. Discussion

Kariņš [6] proposes a phonological lexical tone analysis of the Latvian accentual contrast, based on the F0 contours of words within a carrier phrase. In his analysis, the E accent is phonetically unspecified and receives a phrasal H tone that attaches to the second mora of the stressed syllable. In words with a falling, N, accent, on the other hand, there is a lexical L on the second mora and the phrasal H attaches to the first mora. (Kariņš assigns a more complex tonal specification, LH to words with a broken accent.) The contours that we have obtained in declarative sentences where the target word is under focus are similar to those described in [6], except that our speakers do not have a lexical contrast between non-even accents with and without glottalization ("falling" and "broken"). By expanding the elicitation to other phrasal and intonational contexts, we notice that there is no consistent tonal contour across all contexts. The E accent may be H or L. In final position in questions, this accent may show up without the fall, perhaps because a final H% boundary tone pushes the lexical L from the second mora to the first, under crowding. No phrasal H is then assigned to the accented syllable. Glottalization in the speech of our male speaker (from Riga), appears to be a secondary correlate of N accent.

3. Lithuanian

3.1. Status quaestionis

Lithuanian has a complex stress system, reminiscent of that of Russian, where some nouns have stress on the stem in some declensional cases and stress on the suffix in other cases, other nouns have stress on the suffix in all cases, etc.

A further complication with respect to Russian is that if the stressed syllable is heavy it may bear one of two lexical accents, traditionally known as "acute" (A) and "circumflex" (C). Although traditionally described as a contrast in lexical tone, the nature of this contrast in present-day Lithuanian has been disputed and may differ according to geographical variety. In the standard variety, the contrast is manifested not by pitch, but by differences in spectral structure and duration [3] [4] [9] [11]. Dogil & Möhler [3] report that, regarding pitch, in words pronounced in carrier phrases, A has a clear rise-fall pattern within the stressed syllable, whereas C has a very variable realization and its F0 shape cannot be defined.

3.2. Methods

We selected three near-minimal pairs, stressed on the same syllable and contrasting in word accent: *láužą* A 'bonfire' vs. *laũmę* C 'pixie', *láimę* A 'happiness' vs. *laĩvą* C 'ship' and *výrą* A 'man' vs. *vỹną* C 'wine'). Two of our minimal pairs illustrate the falling diphtongs /ai/ and /au/ and the third one is used to test the contrast with a long vowel, /i:/ (orthographical <y>). To elicit the data we used a questionnaire that included the same 15 pragmatic contexts as for Latvian: broad focus, all new, contrastive, prefocal, narrow focus final, narrow focus non-final, obviousness, postfocus, yes-no question, confirmation question, surprise, suggestion, command, vocative and continuation. Our two subjects are both college-educated female speakers of Lithuanian, from the Highland Lithuanian or Aukštaitian area: one of them is from Utena and the other one is from Panevėžys. They were asked to read both the context and the target sentence with natural intonation.

3.3. Results

Regarding pitch contours, we find no consistent difference between words with A and C accent in the same sentential and intonational context. Both speakers produced very similar F0 contours in each of the contexts tested. Under focus, words of both accent classes show a rise in pitch throughout the stressed syllable with a peak towards the middle or end of this syllable, see Fig. 5 (although, the alignment within the syllable nucleus is different given the longer duration of the first element of the diphthong in A words, see below).

For both classes of words, quite different contours were obtained under other contextual conditions, including a later rise with a peak after the stressed syllable in prenuclear position (Fig. 6) and a low flat contour in phrase-final position under broad focus (Fig. 7).

For words with the diphthongs orthographically represented as <au> and <ai>, we find, however, considerable spectral and durational differences according to etymological class. In particular, in C words, the first element in the diphthong is considerably reduced in duration and centralized in quality, cf. [9]. For all three pairs of words, the stressed nucleus is significantly longer in A than in C words: *láimę* A 'happiness' vs. *laĩvą* C 'ship', t=2.2, df=57.8, p=0.028 ; *láužą* A 'bonfire' vs. *laũmę* C 'pixie', t=4.2 , df =57.57, p <0.0001; *výrą* A 'man' vs. *vỹną* C 'wine', t= 2.9153, df = 57.668, p=0.005.

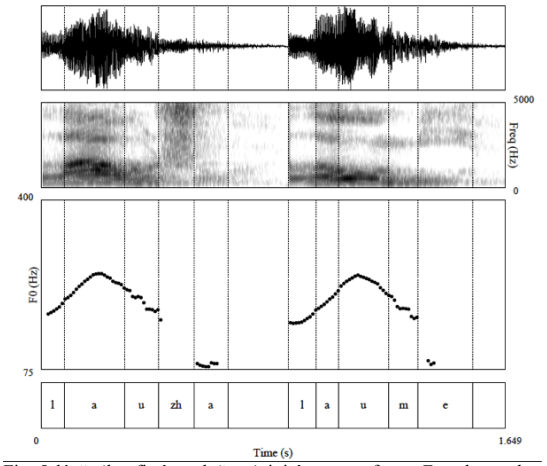


Fig. 5 *láužą* 'bonfire' vs. *laũmę* 'pixie', narrow focus. Female speaker from Utena.

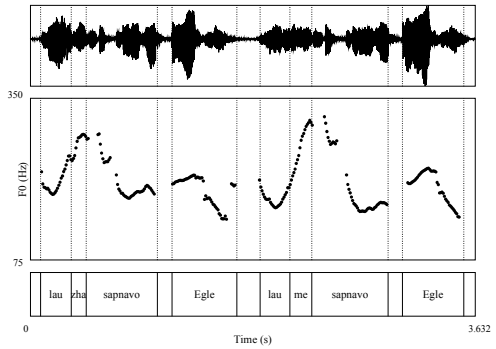


Fig. 6. Late rise on *láužą* vs. *laũmę* in prefocal position: *láužą*/*laũmę sapnãvo Ẽglė*, lit. 'about bonfire/about a pixie dreamt Eglė'. Female speaker from Panevėžys.

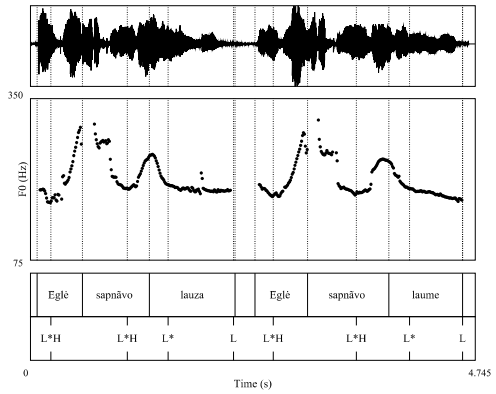


Fig. 7. Flat low pitch on *láužą* and *laũmę* in phrase-final position, broad focus: *Eglė sapnãvo láužą*/*laũmę* 'Eglė dreamt about a bonfire/a pixie'. Female speaker from Panevėžys.

For words with the diphthongs orthographically represented as <au> and <ai>, we find, however, considerable spectral and durational differences according to etymological class. In particular, in C words, the first element in the diphthong is considerably reduced in duration and centralized in quality, cf. [9]. For all three pairs of words, the stressed nucleus is significantly longer in A than in C words: *láimę* A 'happiness' vs. *laĩvą* C'ship', t=2.2, df=57.8, p=0.02856; *láužą* A 'bonfire' vs. *laũmę* C 'pixie', t =4.2 , df=57.5 , p<0.0001; *výrą* A 'man' vs. *vỹną* C 'wine', t =2.9, df=57.6, p=0.005.

In Fig. 8, F1 maxima in stressed nuclei are compared. It is apparent that the A diphthong words *láimę* and *láužą* have a much higher F1 maximum than their C counterparts *laĩvą* and *laũmę*, indicating a considerably lower first element. (t-tests: *láimę* vs. *laĩvą,* t=21.6, df = 40.5, p < 0.0001; *láužą* vs. *laũmę,* t =28.5, df =50.9, p<0.0001). There is no significant difference between the two items containing a long monophthong /i:/.

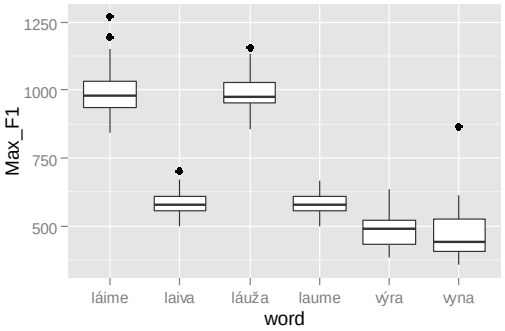


Fig 8. F1 maximum in Hz. The C words, *laĩvą*, *laũmę* have a significantly lower maximum F1 than their A counterparts.

In Fig. 9 we plot the means of formant values at F1 maximum and minimum for <ai> and <au> diphthongs. The graph shows that the A words *láimę* and *láužą* have a low vowel as first element of the diphthong, whereas in the circumflex words, *laĩvą* and *laũmę* this element is a mid vowel. On the other hand, the glide has more extreme values for the C words: higher for <ai> and higher and more posterior for <au>.

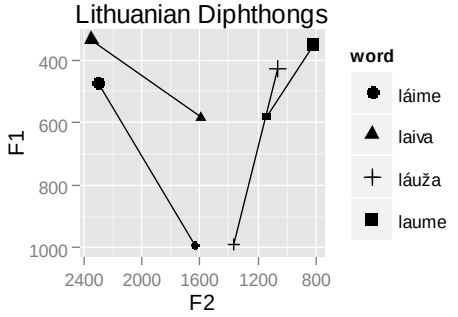


Fig. 9. Plot of F1 maxima and minima and F2 values at those two points for each of the four diphthong items: *láimę*, *laĩvą*, *láužą* and *laũmę*. (Diacritics are missing in the figure for typographical reasons.)

3.4. Discussion

In our East Aukštaitian Lithuanian data, vowel quality and relative duration cue the A vs. C contrast. In particular, the diphthongs <ai>, <au> have a shorter, raised nucleus and a more extreme glide in circumflex words. The long vowel <y> is also slightly shorter in circumflex words. Other heavy syllables, including those closed by a sonorant consonant remain to be investigated. As in other recent acoustic work, we have not found a difference in tonal contours between the two lexical classes.

4. General discussion

The two systems that we have examined are at some typological distance from each other. The stress systems are different, Latvian having developed fixed initial stress, and Lithuanian retaining mobile stress. Whereas for Latvian we find a consistent tonal contrasts in stressed syllables (enhanced by duration), in Lithuanian the etymological accentual contrast is cued by durational and vowel quality differences. For both languages an analysis of the word accent contrast in terms of domain of prominence (syllable vs. mora, [3]) seems possible. However, such an analysis has implications for phonology (e.g. metrical theory), as well as for typology, where intermediate varieties (e.g. mobile stress combined with a tonal contrast) would need to be examined, before an adequate phonological analysis could be determined.

From a diachronic perspective an important question that arises is how a contrast in lexical tonal accent develops into a vowel-quality/durational contrast.

We end with a cautionary note. Given the small number of speakers and test items, the results of this paper must be taken as provisional and awaiting further confirmation.

5. Acknowledgements

We are thankful to our Latvian and Lithuanian speakers for their participation, to Peteris Vanags and Kristina Bukelskytė-Čepele for their help with the experimental materials and to Daniel Scarpace for help with the PRAAT and R scripts. We also gratefully acknowledge the support of the Illinois-Sweden Program for Education and Research (INSPIRE).

6. References

[1] Blevins, J. 1993. "A tonal analysis of Lithuanian nominal accent". *Language* 69(5), 237–273.

[2] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341–345.

[3] Dogil, G. 1999. "Baltic languages". In: H. van der Hulst, ed., *Word prosodic systems in the languages of Europe*, 877–896. Berlin: Mouton de Gruyter.

[4] Dogil, G. & Möhler, G. 1998. "Phonetic invariance and phonological stability: Lithuanian pitch accents. *Proceedings of the International Conference on Spoken Language Processing*, Sydney, vol. 7, 2891–2894.

[5] Ekblom, R. 1933. *Die Lettischen Akzentarten*. Uppsala: Almqvist & Wiksells boktryckeri.

[6] Kariņš, A. K. 1997. "Lexical tone and stress in Latvian". *Berkeley Linguistics Society* 23, 186–197.

[7] Kenstowicz, M. 1972. "Lithuanian phonology". *Studies in the Linguistic Sciences* 2, 1–85.

[8] Prauliņš, D. 2012. *Latvian: An essential grammar*. London: Routledge.

[9] Robinson, D. 1968. “Some acoustic correlates of tone in Standard Lithuanian”. *The Slavic and East European Journal* 12, 206–212.
[10] Stundžia, B. 1996. *Lietuvių kalbos kirčiavimas: mokytojo knyga*, Vilnius: Baltos lankos.
[11] Vaitkevičiūtė, V. 2004. *Bendrinės lietuvių kalbos kirčiavimas. Antrasis leidimas*. Kaunas: Šviesa.

Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information

Neville Ryant[1], Malcolm Slaney[2], Mark Liberman[1], Elizabeth Shriberg[3], and Jiahong Yuan[1]

[1]Linguistic Data Consortium, Philadelphia, PA, USA
[2]Microsoft Research, Mountain View, CA, USA
[3]SRI International, Menlo Park, CA, USA

nryant@gmail.com, malcolm@ieee.org, markyliberman@gmail.com
ees@icsi.berkeley.edu, jiahong.yuan@gmail.com

Abstract

A deep neural network (DNN) classifier based only on 40 mel-frequency cepstral coefficients (MFCCs) achieved 29.99% frame error rate (FER) and 16.86% segment error rate (SER) in recognizing five tonal categories in Mandarin Chinese broadcast news. With the addition of subband autocorrelation change detection (SACD) pitch-class features [1], the classifier scored 27.58% FER and 15.56% SER. These results are substantially better than the best previously reported results on broadcast news tone classification [2] and are also better than a human listener achieved in categorizing test stimuli created by amplitude- and frequency-modulating complex tones to match the extracted F_0 and amplitude parameters [3]. The same DNN architecture scored substantially worse when trained and tested with SACD pitch-class parameters alone: 39.22% FER and 24.89% SER. RAPT F_0 estimates are worse yet: 44.37% FER and 27.28% SER. The 40 MFCC parameters do not encode F_0 in any obvious way and attempts to predict SACD or other pitch features from them work badly. These surprising results raise difficult questions for theories of Chinese tone.

Index Terms: speech recognition, Mandarin, tone modeling, deep neural networks

1. Introduction

Typically, Chinese speech-recognition systems have included tonal features in order to improve performance in the integrated task of recognizing tonally-specified segments [4, 5, 6, 7]. More recently, there has been increased interest in the more specific problem of automated recognition of tonal categories alone in continuous speech [8, 2, 9, 10]. For instance, Pui-Fung [8] uses decision trees and a segmental representation based on the fitting of polynomials to the F_0 contour to achieve 27.8% segment error rate (SER). Lei [2] achieves 23.8% SER using MLPs and contextual information. Most recently, Kalinli [10] achieved 21% SER, albeit for command-and-control utterances, with the incorporation of biologically inspired auditory features.

Of the above papers, all save Kalinli perform explicit pitch tracking (though even Kalinli includes parameters that are probably an excellent proxy for F_0 slope). However, pitch is notoriously hard to accurately estimate even in cases where it is not inherently ambiguous [11]. Moreover, for the task of interest, tone classification, absolute pitch is not itself even particularly relevant but, rather, changes in pitch over an interval of time. Such being the case, it has been suggested that it is more appropriate to estimate pitch changes directly [1]. Using subband autocorrelation change detection (SACD) features, Slaney achieves superior performance for 4-way tone classification on a corpus of Mandarin phone speech with the SACD features providing relative reductions in error ranging from 10% for clean materials to 17% for speech corrupted by white noise.

Yet more recent work demonstrated successful Mandarin tone classification for broadcast news materials in the absence of any explicit pitch-related information whatsoever [3]. Using a deep neural network (DNN) based classifier and an input representation consisting of 21 consecutive frames of 40 mel frequency cepstral coefficients, our previous work achieves an SER of 16.62%, a 7.04% absolute reduction relative to a baseline system incorporating explicit, but perhaps errorful, F_0 information.

Jointly the findings of Slaney and Ryant suggest that at least for some tone languages, highly accurate tone classification is possible in the absence of explicit pitch tracking; indeed, that tone is not "just" about F_0. In this paper we extend this work and directly compare the efficacy of these features for Mandarin tone classification using the same training/test sets and machine learning infrastructure. We also consider possible explanations for why the MFCC frontend is so successful.

2. Data and evaluation

Testing and training sets were constructed using the 1997 Mandarin Broadcast News Speech corpus [12][1]. We extracted all "utterances" (the between-pause units that are time-stamped in the transcripts) from the corpus and manually excluded those containing background noise or music. Utterances from speakers whose names were not tagged in the corpus or from speakers with accented speech were also excluded. In total 7,849 utterances from 20 speakers were selected. From these we randomly selected 50 utterances from each of six speakers to compose a test set, with the remaining 7,549 utterances reserved for training. The 300 test utterances were manually labeled and segmented into initials and finals by a native Mandarin speaker. Tones were marked on the finals, including Tone1 through Tone4, and Tone0 for the neutral tone. The total number of utterances, segments, and hours of speech are detailed in Table 1.

	Hours	Utterances	Segments	TBUs
Train	6.05	7,549	196,330	96,697
Test	0.22	300	7,189	3,464

Table 1: Train/test set composition. TBU = tone-bearing unit, defined as the syllable final.

System performance is measured in two ways. As an initial evaluation of the quality of the representation learned by the network, we consider its frame error rate (FER), defined as the percentage of frames incorrectly classified by the DNN. Our primary metric, however, is segment error rate (SER), defined as the percentage of TBUs incorrectly classified.

3. System description

We propose attacking the problem of explicit tone classification as follows:

1) Train a DNN to classify each frame of speech into one of six tone classes: Tone0, Tone1, Tone2, Tone3, Tone4, No-tone.

2) Compute "tonal features" for each segment, defined as the mean of the outputs of the DNN over all frames contained within that segment. These are similar to Chao's articulatory features [9].

3) Use these "tonal features", along with segment duration and contextual features, to classify the tone-bearing units (TBUs).

3.1. Features

We train four separate tone-classification systems using different feature frontends:

1. *RAPT F_0 estimate*
   Our first feature consists of F_0 as estimated by peaks in the normalized cross-correlation function using RAPT [13] as implemented in ESPS's *get_f0* with the following parameters: wind_dur=0.01, min_f0=60, max_f0=650.

2. *SAaC F_0 estimate*
   A second F_0 estimate is computed using the SAaC system [14]. A correlogram is constructed by running the signal through an auditory filterbank and calculating the autocorrelation for each channel. The size of this representation is reduced using PCA and the retained principal components serve as input to an MLP that classifies frames into one of 70 pitch classes (67 classes spanning 60-400 Hz on a logarithmic axis plus additional classes corresponding to unvoiced, out of range low, and out of range high). Viterbi decoding of the MLP outputs produces a smoothed pitch track.

3. *SACD*
   We also consider SACD features [1]. As with SAaC, an MLP is trained to classify frames into one of 70 pitch classes on the basis of the principal components of a correlogram. The MLP-derived pitch class probabilities are smoothed across frames using a 5-frame moving average window and cross-correlation between adjacent frames calculated for a range of lags. The final SACD features consist of the cross-correlation values corresponding to lags from -2 to 2.

4. *MFCC*
   Forty mel frequency cepstral coefficients (MFCCs) were extracted using the following analysis parameters: i) 0.97 pre-emphasis factor; ii) 25 ms Hamming window; iii) 1024-point DFT; iv) 40 filter mel-scale filterbank[2].

In addition to systems trained using the RAPT, SAaC, SACD, and MFCC features individually, we also consider each combination of the MFCC features and the pitch-related featuress. All features, including the F_0 estimates, were computed every 10 ms and normalized to have 0 mean and unit variance on a per-utterance basis[3].

3.2. Network training

For each feature combination a DNN was trained [16] to classify frames of the signal as one of the six targets.

[1] The specific dataset used in these experiments will be published by the LDC and meanwhile is available from the authors by request.

[2] Our MFCCs may be reproduced using *melfcc* [15] with the following parameter values: wintime=0.025, hoptime=0.010, nbands=40, numcep=40, lifterexp=-22, sumpower=0, minfreq=0, maxfreq=8000, dcttype=3.

[3] We examined other normalization schemes, including one in which F_0 normalization was restricted to voiced segments, but this choice had negligible impact on the final accuracy.

Input to the DNN consisted of a high-dimensional feature vector derived by concatenating the extracted features for all frames in a 21-frame context window (10-1-10). Training targets were derived by forced alignment of the HUB-4 training utterances using an HMM-based forced aligner built on the training utterances with the CALLHOME Mandarin Chinese Lexicon [17] and HTK. The aligner employed explicit phone boundary models [18] and achieved 93.1% agreement within 20 ms compared to manual segmentation on the test set. Additionally, we checked 100 training utterances on the tone labels automatically generated by the aligner. Among the 1,252 syllables in the 100 utterances, 15 syllables had a wrong tone, an error rate of 1.2%[4].

The full network topology consisted of: i) the input layer; ii) 4 hidden layers, each consisting of 2000 rectified linear units (ReLUs) [19]; iii) an output layer consisting of 6 softmax units. The network was trained for 60 epochs (each epoch consisting of 250,000 examples) using stochastic gradient descent with a minibatch size of 128, 20% dropout [20] in the input layer, 30% dropout in the hidden layers, and a cross-entropy objective. Learning rate was kept constant within epochs and followed the schedule $\eta(n) = \eta(0)\frac{500}{n+500}$, where $\eta(0) = 0.5$, while momentum was kept constant at 0.5 throughout training. No L_2 weight decay was used, but the incoming weight vector at each hidden unit was constrained to have a maximum L_2-norm of 3.

3.3. Segment-level classification

Segment-level classification decisions were made using a single-layer neural network trained to assign tone classes to the TBUs. Input features consisted of the tonal features of the segment, duration (in seconds) of the segment (as determined by the forced alignment boundaries), and tonal features and durations of the two immediately preceding and two immediately following segments. The neural network contained a single hidden layer of 128 ReLUs and was trained for 1,000 epochs (epoch=100,000 instances) using stochastic gradient descent with minibatch size of 512, 30% hidden layer dropout, a decaying learning rate beginning at 1, and a constant momentum of 0.9. The incoming weight vector at each hidden unit was constrained to have a maximum L_2-norm of 1.

4. Results

FERs and SERs for the trained systems are shown in Table 2. Because silences and other unvoiced regions are relatively easy to recognize in material of this kind and, therefore, a FER that includes such regions will depend on the amount of silence that is included in the test set, we depict not only overall FER, but also FER excluding frames that do not correspond to a tone bearing unit in the gold standard segmentation. Three results are immediately apparent. One, in accord with earlier findings [1], the SACD features are more informative than either RAPT or SAaC-derived F_0 estimates. Indeed, FER on TBUs for the the system trained on SACD features is 39.22% and SER 24.89%, which represents relative error reductions of 11.61% and 8.76% respectively from the figures achieved by the system using RAPT F_0 estimates. Two, replicating our earlier findings [3], the system trained only using the MFCC frontend trounces the systems trained using only pitch related features, reducing TBU FER by 23.53% and SER by 32.36% relative to the system trained using SACD. Three, while inclusion of F_0 alongside MFCCs fails to improve (hurts actually) performance, adding SACD features does appear to help, resulting in relative reductions of 8.04% for FER on TBUs and 8.35% for SER. This result suggests that, whatever information is contained in the MFCCs, it is complementary to that contained in the SACD features.

	Frame Error Rate (FER)			
	Overall	TBUs	Tones 1−4	SER
RAPT	29.09	44.37	42.05	27.28
SAaC	32.39	49.64	47.55	28.67
SACD	25.25	39.22	37.05	24.89
MFCC	18.88	29.99	29.35	16.86
MFCC+RAPT	18.70	29.98	29.43	17.47
MFCC+SAaC	18.79	29.38	28.75	17.52
MFCC+SACD	17.57	27.58	27.00	15.56

Table 2: Frame error rates and segment error rates (%) on test set for DNNs trained using various combinations of the feature frontends.

5. General discussion

The success of MFCCs with a context window of many frames of MFCCs is, at first glance, perplexing: how does a representation in which information about pitch should be eradicated, or at least substantially blurred, do so well at predicting tones on segments, a task that is supposedly entirely about pitch? One possible explanation for our performance is that the DNN system is actually somehow implicitly learning to do overall phone recognition with the tone recognition merely a byproduct. While perfect (toneless) phone recognition is implausible[5], we do put the idea to the test by comparing FER and SER of the MFCC-trained system with the FER and SER of an oracle

[4]These errors are primarily due to application of third tone sandhi across word boundaries.

[5]When we trained a DNN with the same topology and hyperparameters used for the tone classification experiments to predict the (toneless) phone categories using the MFCC features as input, final frame error rate on the test set came to 21.2%, suggesting that, in the absence of a language-model or other higher-level information, we would be unlikely to do better than 80% accuracy at predicting phones, much less 100%.

with perfect knowledge of the pinyin of each initial/final, but no other information about tone at all (Table 3). An oracle making maximum-likelihood guesses given perfect phone knowledge produces 51.96% SER compared to 16.86% for the MFCC only system.

Perhaps more context helps? Just in case, we also consider the performance of a second oracle, which predicts the tone class of each segment using perfect (toneless) knowledge of the phone class of the preceding, current, and following segments and a neural network with a single hidden layer of 128 rectified linear units (depicted as Oracle (tri) in the table). Inclusion of this additional context does improve performance markedly, bringing TBU FER down to 21.27% and SER to 20.76%, suggesting that in the unlikely event of perfect recognition of a span of three pinyin initials or finals, reasonably good tone recognition is possible, even in the absence of a language model. However, these error rates remain substantially higher than what the DNN is achieving, suggesting some other mechanism is at work.

	Frame Error Rate (FER)			
	Overall	TBUs	Tones 1−4	SER
Oracle (mono)	28.28	52.14	53.79	51.96
Oracle (tri)	11.54	21.27	21.84	20.76

Table 3: Frame error rates and segment error rates (%) on test set for two oracle systems.

Alternately, it may be the case that the DNN is making use of a multitude of other, non-pitch, phonetic dimensions, which jointly are predictive of tone class. Acoustic analysis of Mandarin syllables suggests that duration [21, 22], temporal envelope [22], and formant structure [23] differ for different lexical tones of the same syllable. Moreover, it is well established that, though impaired relative to clean speech, native speakers are able to identify tone in both real [24, 25] and synthetic [26, 22] whispered speech at well above chance levels. In light of these findings the thesis tying the DNN's performance to efficient use of non-pitch information represented in the MFCCs is plausible.

Finally, it should be considered that the DNN may be recovering F_0 information from the MFCC parameters, either in terms of the actual pitch track or some other form. Conventional wisdom suggests that MFCC and its ilk are good for speech recognition because they represent the rough shape of the spectrum, but without the pitch information[6]. Nevertheless, as a test of the hypothesis that F_0 is being extracted, we trained a DNN to predict the SAaC/SACD pitch classes using the MFCC features as input. While this network was able to achieve a frame error rate of 29.48% on the test set for pitch-class prediction, error analysis reveals that this is principally because the network is very good at making voicing distinctions as opposed to actually successfully determining pitch in voiced segments (unvoiced frame error: 1.83%; voiced frame error: 45.3%).

However, this does not rule out the other possibility: that the network is pulling out some other, pitch-related information from the MFCC representation that has predictive power. To explore this idea, we performed a simple comparison experiment by synthesizing a large number of static vowels with random pitches centered at 120 Hz (male) and 200 Hz (female) using Praat [29]. Figure 1 shows the relative contrast between three different vowels (/a/, /ɪ/, /u/), and the same vowel at two pitches separated by 2 semitones. As we vary the number of cepstral coefficients between 1 and 40, the MFCC representation does a better job of capturing their differences, as reflected in the Euclidean distance. For these highly stylized vowels (fixed pitch, no noise, no coarticulation) the female pitch change leads to longer distances, suggesting that the pitch change is reflected in the MFCC coefficients at least for widely spaced harmonics. Interestingly, this difference only shows up when the number of cepstral coefficients is more than 20. This difference might allow a classifier to more easily notice the two classes. Yet, we were not able to see any significant difference in the performance of the DNN network when we looked at male vs. female speakers.

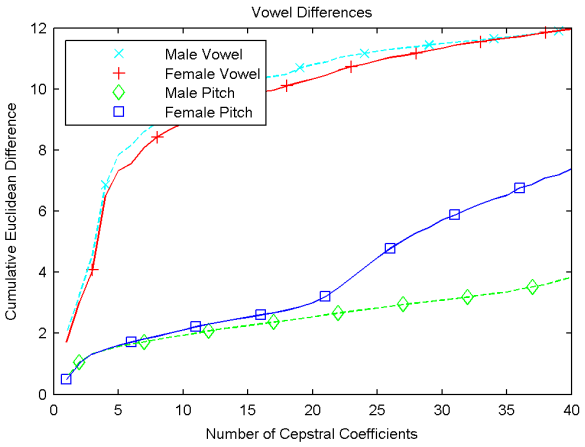


Figure 1: The average distance between two vowels, as a function of the size of the cepstral vector. The top two curves are for two different vowels at the same pitch. The bottom two curves are for the *same* vowel at two pitches that differ by 2 semitones. The synthetic vowels have an average pitch of 120 Hz for the male examples, and 200 Hz for the female examples, in line with the Mandarin database used in this paper.

Most probably, all three hypotheses are true to an extent and the DNN is using all three sources of information jointly to make its final predictions.

[6]Though see also [27, 28], who report success in predicting F0 in English read speech using a standard 23 channel, 13 cepstral coefficient MFCC representation.

6. References

[1] M. Slaney, E. Shriberg, and J.-T. Huang, "Pitch-gesture modeling using subband autocorrelation change detection," in *INTERSPEECH*, 2013, pp. 1911–1915.

[2] X. Lei, M.-H. Siu, M.-Y. Hwang, M. Ostendorf, and T. Lee, "Improved tone modeling for Mandarin broadcast news speech recognition," in *INTERSPEECH*, 2006.

[3] N. Ryant, J. Yuan, and M. Liberman, "Mandarin tone classification without pitch tracking," in *Proceedings of ICASSP*, 2014.

[4] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous Mandarin speech recognition," in *Eurospeech*, 1997.

[5] E. Chang, J.-L. Zhou, S. Di, C. Huang, and K.-F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *INTERSPEECH*, 2000, pp. 983–986.

[6] H. C.-H. Huang and F. Seide, "Pitch tracking and tone features for Mandarin speech recognition," in *Proceedings of ICASSP*, 2000, pp. 1523–1526.

[7] R. Sinha, M. Gales, D. Kim, X. Liu, K. Sim, and P. Woodland, "The CU-HTK Mandarin broadcast news transcription system," in *Proceedings of ICASSP*, 2006.

[8] W. Pui-Fung and S. Man-Hung, "Decision tree based tone modeling for Chinese speech recognition," in *Proceedings of ICASSP*, 2004, pp. 905–908.

[9] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *Proceedings of ICASSP*, 2012, pp. 4741–4744.

[10] O. Kalinli, "Tone and pitch accent classification using auditory attention cues," in *Proceedings of ICASSP*, 2011, pp. 5208–5211.

[11] R. N. Shepard, "Circularity in judgments of relative pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, 1964.

[12] S. Huang, J. Liu, X. Wu, L. Wu, Y. Yan, and Z. Qin, *1997 Mandarin Broadcast News Speech (HUB4-NE)*. Linguistic Data Consortium, 1998.

[13] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis*. New York: Elsevier, 1995, pp. 495–518.

[14] B. S. Lee and D. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *INTERSPEECH*, 2012.

[15] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/∼dpwe/resources/matlab/rastamat/

[16] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] S. Huang, X. Bian, G. Wu, and C. McLemore, *CALLHOME Mandarin Chinese Lexicon*. Linguistic Data Consortium, 1997.

[18] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *INTERSPEECH*, 2013, pp. 2306–2310.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Procedings of ICML*, 2010, pp. 807–814.

[20] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[21] C.-y. Tseng, *An acoustic phonetic study on tones in Mandarin Chinese*. Brown University, 1981.

[22] Q.-J. Fu and F.-G. Zeng, "Identification of temporal envelope cues in Chinese tone recognition," *Asia Pacific Journal of Speech Language and Hearing*, vol. 5, no. 1, pp. 45–58, 2000.

[23] Y.-Y. Kong and F.-G. Zeng, "Temporal and spectral cues in Mandarin tone recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2830–2840, 2006.

[24] M. K. Jensen, "Recognition of word tones in whispered speech," *Word*, vol. 14, no. 2-3, pp. 187–96, 1958.

[25] M. Gao, "Tones in whispered chinese: articulatory features and perceptual cues," Ph.D. dissertation, University of Victoria, 2002.

[26] D. H. Whalen and Y. Xu, "Information for Mandarin tones in the amplitude contour and in brief segments," *Phonetica*, vol. 49, no. 1, pp. 25–47, 1992.

[27] B. Milner and X. Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 24–33, 2007.

[28] J. Darch, B. Milner, and S. Vaseghi, "Analysis and prediction of acoustic speech features from mel-frequency cepstral coefficients in distributed speech recognition architectures," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3989–4000, 2008.

[29] P. Boersma and D. Weenink, "Praat speech processing software," *Institute of Phonetics Sciences of the University of Amsterdam*. [Online]. Available: http://www. praat. org

Intended intonation of statements and polar questions in Polish in whispered, semi-whispered and normal speech modes.

Marzena Żygis[1], *Daniel Pape*[2], *Luis M.T. Jesus*[2,3], *Marek Jaskuła*[4]

[1] Centre for General Linguistics (ZAS) & Humboldt University, Berlin, Germany
[2] IEETA, University of Aveiro, Portugal
[3] ESSUA, University of Aveiro, Portugal
[4] West Pomeranian University of Technology, Szczecin, Poland
zygis@zas.gwz-berlin.de,danielpape@ua.pt,lmtj@ua.pt, Marek.Jaskula@zut.edu.pl

Abstract

This paper examines acoustic correlates of intonation in Polish whispered, semi-whispered and normal speech modes. In particular, it investigates correlates of utterance-final rising intonation in polar questions and falling intonation in statements. The paper examines not only properties of vowels but also properties of the following voiceless consonant clusters.

The study is based on measurements of 4608 sibilants (fricatives and affricates) produced by 16 native speakers of Polish. The results point to differences in spectral properties of both utterance-final vowels and consonants where falling intonation in statements contrasts with rising intonation in polar questions. Regarding the consonants, both fricatives and affricates are produced with higher spectral peaks, higher intensity and higher COG and STD values in questions than in statements. Skewness and kurtosis values are lower in questions than in statements. Some spectral differences of sibilants, including spectral slopes, are more distinguishable for questions versus statements in the whispered speech mode than in other speech modes. The more pronounced role of these cues in whispered speech suggests their compensatory function for the fundamental frequency, which is the main correlate of intonation in phonated speech but is completely absent in whispered speech.

In summary, the study shows that speakers produce intended intonation patterns by varying the choice of cues as well as their magnitude in dependence on both (i) speech modes and (ii) intonation patterns.

Index Terms: whispered speech, intonation, voiceless clusters, Polish

1. Introduction

Interaction between segments and prosody, in general, and between segments and intonation in particular, still belongs to an understudied area of phonetic and phonological research. Relatively little is known about this interaction in whispered speech, and less so in semi-whispered speech; cf. [1] for studies of this interaction in the normal speech mode.

The overall goal of this paper is two-fold. First, it is aimed at providing an insight into the realisation of intended intonation in whispered speech where F0, the main correlate of intonation is completely absent and semi-whispered speech, where F0 is partially absent. Second, the study is intended to contribute to a better understanding of the role of voiceless segments in conveying different intonation patterns in various speech modes.

In particular, the paper addresses two questions:

(1) How are different patterns of intonation produced in whispered speech in comparison to semi-whispered and normal speech modes?

(2) How do voiceless consonant clusters contribute to intended intonation patterns in all three speech modes?

In order to answer these questions we conducted an acoustic speech production experiment on Polish, a language which provides suitable test material due to its abundance of complex consonant clusters.

2. Methods

We recorded eight different items ending in clusters consisting of voiceless retroflex fricatives followed by retroflex affricates. Each item was presented in a polar question ('Widzi ten blu[ʂt͡ʂ]?' '*Does he see the ivy*?') and a statement ('Widzi ten blu[ʂt͡ʂ]' '*He sees the ivy*'). All words were monosyllabic. The polar question was expected to be produced with a rising intonation and the statement with a falling intonation. In order to compare the realisation of intonation in whispered speech with other speech modes we recorded the questions and statements described above in three different speech modes: normal, whispered and semi-whispered.

Sixteen native speakers of Polish (eigth male), aged 20 to 52 years, took part in the experiment. All speakers were monolingual and spoke Standard Polish. They were asked to read the sentences in all three speech modes, as described above. All recordings were conducted in the sound-proof room at the Electrical Engineering Department of the West Pomeranian University of Technology in Szczecin using a TLM103 Neumann microphone (20cm distance from lips) connected to a ProTools system with a Digi 003 interface (sample rate 44100 Hz). The items were analysed with Praat (version 5.3.57 [2]) and MATLAB (version R2007b [3]). In total, measurements of 4608 sibilants were taken (8 items × 2 sentence types (question, statement) × 3 speech modes (normal, semi-whispered, whispered) × 2 sibilant types (fricative, affricate) × 3 repetitions × 16 speakers).

We investigated the following acoustic parameters, see [4] for all parameter details. Note that the parameters displayed in (f), (g) and (h) were calculated at the onset, midpoint and offset of the frication of both sibilants.

(a) duration of the vowel, fricative and affricate;

(b) maximum and mean of F0 of the preceding vowel;

(c) F0 difference between the vowel offset and onset;

(d) formants F1, F2, F3 (using the formant extraction of PRAAT [2]) at the (i) vowel onset, (ii) midpoint and (iii) vowel offset;

(e) mean intensity over the complete duration of the vowel, the fricative and the affricate;

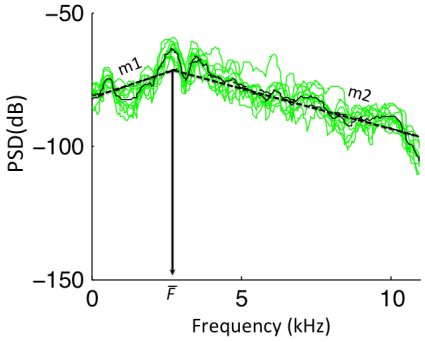
(f) frequency of the highest spectral peak of the frication noise in the range from 20Hz to11kHz;

(g) spectral Centre of Gravity (COG), its standard deviation (STD), skewness, and kurtosis;

(h) the spectral regression slopes, as described in [5]: m1 is the slope of the spectral regression line for the frequency range between 500 Hz and 3000 Hz, and m2 is the slope of the spectral regression line for the range between 3000 Hz and 11025 kHz (see Figure 1).

We computed multitaper spectra with a 12 ms window for the frication noise midpoint (512 point Hamming window). The power spectral density (PSD) was estimated via the Thomson multitaper method (linear combination with unity weights of individual spectral estimates and the default FFT length) available in the MathWorks Signal Processing Toolbox Version 6 [3].

Figure 1 shows an example of 10 different multitaper spectra, from one individual speaker, the overlaid mean spectrum and the computation of the regression lines m1 and m2, with the endpoint/startpoint $\bar{F}$.



(i)

Figure 1: *Multitaper spectra (light colour) with mean spectrum (black solid) and regression lines (dashed black) used to calculate the low-frequency slope (m1) and high-frequency slope (m2), with the end/starting point at the mean frequency* $\bar{F}$.

Regarding statistical analysis, linear mixed effect models were employed for the investigated variables, which were studied as effects of INTONATION (rising vs. falling) and SPEECH MODE (normal, semi-whispered and whispered), as well as their interaction (INTONATION*SPEECHMODE). GENDER (male, female) was included as fixed effect as well. In addition, speaker-specific random slopes for INTONATION TYPE and SPEECH MODE were included into the model. ITEM and SPEAKER were taken as random effects.

All analyses were conducted in the R environment software (version 3.0.2) [6].

3. Results

Our results show that different acoustic cues are used to a different extent when producing questions versus statements in various speech modes. In the following we report only those results which – due to their significance – are important for answering the main questions of this study, cf. (1) and (2).

Regarding duration of the vowel, whispered vowels were generally longer than vowels produced in semi-whispered (t=7.304, p<.0001) and normal speech mode (t=5.799, p<.0001). No significant differences were found between statements and questions in all three speech modes.

As was expected, the F0 maximum and the F0 mean of the vowel preceding the sibilant cluster were significantly higher in questions than in statements (F0 max: t=11.21, p <.0001, F0 mean: t=13.41, p <.0001). In addition, we calculated the F0 difference between the vowel offset and onset which in fact points to raising F0 in questions (female:142 Hz, male:81 Hz) and falling F0 in statements (female:-15 Hz, male:-17 Hz) confirming our initial assumption about F0 differences in intonation patterns (t=10.726, p<.0001); cf. also [7], [8]. Due to the complete absence of the F0 in whispered speech and the partial absence of the F0 in semi-whispered speech, (where the F0 cannot be reliably extracted) we did not analyse this parameter in those two modes.

The absence of F0 in whispered speech leads us to a key question for the present study, namely: How can a distinction in intonation be realized in questions and statements in whispered speech if the F0, the most important correlate of intonation, does not play a distinguishing role?

First, our results point to the importance of formants. The first formant frequency is significantly higher in questions than in statements in whispered (t= 6.23, p<.0001) and normal speech (t=5.174 p<.0001) but lower in semi-whispered speech (t=-3.726 p<.0001). Furthermore, the second formant frequency of the vowel is significantly higher in questions than in statements only in whispered speech (t=2.812, p<.01) but lower in semi-whispered speech (t=-6.485 p<.0001). In normal speech mode the formants do not show any significant difference when questions and statements are performed. Finally, the third formant frequency is higher in questions than in statements in whispered (t=4.15, p<.0001) and normal speech (t=1.76, p<.05).

Regarding the mean intensity of the vowel, the results show that it is significantly higher in questions than in statements across all speech modes: whispered (t=16.76, p<.0001), normal (t=20.67, p<.0001) and semi-whispered (t=14.91, p<.0001).

Besides clear differences in the realisation of whispered vowels in questions and statements, our results point to significant differences in sibilant clusters depending on the intonation type.

Regarding duration, fricatives were longer in whispered speech than normal speech mode (t=2.941, p.<01). Similarly, duration of affricates was longer in whispered speech than semi-whispered (t=1.774, p<.05) and normal speech modes (t=2.493 p<.001). No differences in duration of both fricatives and affricates were found regarding question vs. statement distinction across all three speech modes.

Furthermore, considerable differences were found in spectra of the consonants.

Figure 2 presents multitaper spectra of all recorded items of 8 male speakers obtained at the acoustic midpoint of frication for all three speech modes, where the black lines show statement conditions and the grey lines question conditions.

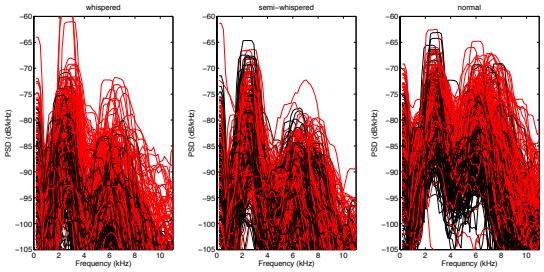


Figure 2: Multitaper spectra of 1145 affricate spectra from 8 male speakers obtained at the acoustic frication midpoint for all three speech modes in separate panels.

With regard to the spectral regression lines slopes of the fricative (m1, m2, cf. Figure 1), m1 exhibits significantly lower values in whispered than in normal speech (t=-2.288, p<.05) and semi-whispered speech (t=-1.997, p<.05). The m2 value is significantly higher in whispered than in normal speech (t=4.01, p<.0001) and semi-whispered speech (t=4.35, p<.0001). Only in whispered fricatives, a distinction in m1 is significant when comparing questions vs. answers, i.e. whispered fricatives display higher m1 values in questions than in statements (t=3.349, p<.001). Regarding m2, questions are produced with a lower m2 in comparison to statements in whispered (t=-8.581, p<.0001) and semi-whispered fricatives (t=-2.19, p<.05).

The spectral slope m1 at the midpoint of the following affricate shows significantly lower values in whispered than in normal speech mode (t=-3.757, p<.0001). Similar to fricatives, a distinction between questions and statements was only found in whispered speech mode (t=3.290, p<.001). Furthermore, with respect to m2, the results show significantly higher m2 values in whispered than in semi-whispered (t=3.915, p<.0001) and normal speech mode (t=-4.67, p<.0001). Again, questions are produced with lower m2 values than statements across all three speech modes differing in t-values: whispered speech mode (t=-12.60, p<.0001), normal speech mode (t=-4.67, p<.0001) and semi-whispered speech mode (t=-4.194, p<.0001). Both slopes are shown in Figure 3; cf. also Figure 1.

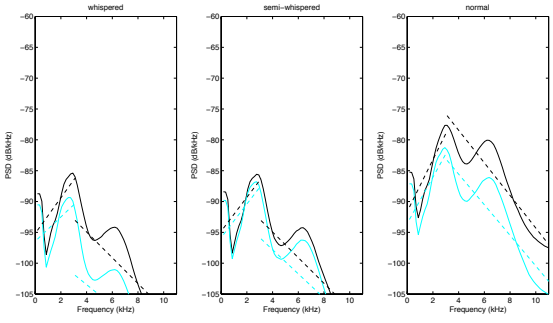


Figure 3: *Multitaper spectra (mean plots over all items and speakers) for the frication midpoint of affricates in whispered, semi-whispered and normal speech mode. Black solid lines correspond to the question and lighter colour to the statement condition. Dotted lines are the spectral regression lines m1 and m2.*

The mean intensity of both fricatives and affricates differs significantly for questions compared to statements: the fricative displays a higher intensity mean for questions than for statements (whispered: t=17.01, p<.0001, normal: t=15.76, p<.0001, semi-whispered: t=5.29, p<.0001). A similar pattern applies for affricates (whispered: t=12.99, p<.0001, normal: t=16.63, p<.0001, semi-whispered: t=4.11, p<.0001).

The frequency of the highest peak of the fricative at its midpoint is higher in questions as opposed to statements for all three speech modes: whispered (t=2.737, p<.01), semi-whispered (t=3.259 p<.001) and normal (t=4.886, p<.0001). Similar observations apply to the affricate midpoint, but exclusively to whispered (t=3.149, p<.001) and normal speech (t=8.726, p<.0001).

Significant differences are also found for the four spectral moments (COG, STD, skewness and kurtosis) at the sibilants' midpoint. The COG values are significantly higher for questions than for statements across all speech modes. This conclusion holds true for both fricatives and affricates whereby the t-values are higher for whispered and normal speech in comparison to the semi-whispered speech mode. The results are presented in Table 1 together with the COG's standard deviation (SD) at the midpoints of both sibilants.

Table 1: Comparison of COG and SD values for the difference between questions and statements across three different speech modes.

	whispered	semi-whisp.	normal
Fricative (COG)	t=5.316 p<.0001	t=2.977 p<.01	t=4.99 p<.0001
Affricate (COG)	t=6.101 p<.0001	t=1.574 n.s.	t=7.88 p<.0001
Fricative (STD)	t=5.923 p<.0001	t=1.786 p<.05	t=0.82 n.s.
Affricate (STD)	t=3.375 p<.001	t=1.153 n.s	t=2.43 p<.01

Regarding skewness, the third spectral moment, our results indicate significant differences between all three speech modes. For fricatives, skewness is significantly higher in whispered speech compared to normal (t=4.418, p<.0001) and semi-whispered speech (t=3.297, p<.0001). In the same vein, frication in affricates displays higher skewness values in whispered than in normal (t=7.096, p<.0001) or semi-whispered speech mode (t=4.343, p<.0001). If we compare the production of questions in comparison statements, the results indicate lower skewness values for questions as compared to statements for fricatives in whispered speech mode only (t=-3.949, p<.0001). In affricates, the skewness is lower in questions than in statements for both whispered (t=-6.636, p<.0001) and normal speech (t= -1.975, p<.05).

The fourth spectral moment, kurtosis, is significantly higher for fricatives in whispered than in semi-whispered (t=3.290, p<.001) and normal speech mode (t=3.842, p<.0001). However, when comparing questions to statements, kurtosis values are significantly smaller in whispered speech (t=-5.656 p<.0001). Regarding affricates, the results show that kurtosis was significantly higher in whispered than in semi-whispered (t=3.968, p<.0001) and normal speech mode (t= 6.066, p<.0001). However, a significant difference in the production of statements vs. questions is found only in whispered speech, where kurtosis is lower in questions (t=-8.163, p<.0001).

4. Discussion

The results show that, in Polish, different intonation patterns between questions and statements are produced by means of various acoustic cues which are, in turn, dependent on the speech mode. In whispered speech, where the F0 is entirely absent, intended rising intonation in questions is produced by both vocalic and consonantal cues. Regarding the former, the results point to a higher F1 and F2 in the utterance-final vowel, a higher amplitude in questions only and no difference in vowel duration between questions and statements.

These results are in accordance with findings reported for the Dutch language where it was shown that in whispered speech F1 and F2 are higher in questions than statements for /ə/ [9]. However, a higher F1 was not found when other Dutch vowels were investigated [10]. Similarly to the results of the present study, the amplitude of the vowel was higher in questions as opposed to statements [9], [10]. In addition, no difference in vowel duration was found between whispered questions and statements [9].

Whereas the majority of previous studies have focused on the properties of vowels when investigating intonation/pitch [11], [12], [13], the results of the present study also point to spectral differences of utterance-final consonants [14], [15]. Significant differences in virtually all spectral parameters are found when comparing questions to statements. The former are produced with higher spectral COG values. This finding is in line with [14], where higher COG values of voiceless fricatives were reported for questions in German phonated speech mode. In the present study, higher COG values are also found for both fricatives and affricates in whispered and semi-whispered questions as compared to statements. The higher COG values are accompanied by higher SD values (with the exception of affricates in semi-whispered speech mode).

The third and the fourth spectral moments are of special importance as they differ considerably between questions and statements across all speech modes. The significantly higher skewness in whispered speech indicates that the mass of the spectral distribution is towards lower frequency values in whispered speech in contrast to the other speech modes. However, for whispered questions, the mass of the spectral distribution moves towards higher frequencies as compared to whispered statements. In affricates, the latter difference applies to normal speech as well, but it is considerably less pronounced in normal speech than in whispered speech.

The higher kurtosis (peakedness; width of peak) in whispered speech indicates a narrower spectral peak for fricatives and affricates in comparison to their semi-whispered and normal speech counterparts. The width of peak in frication of both fricatives and affricates is wider in questions than in statements in whispered speech. In affricates, the width of peak is wider in questions than in statements in both speech modes. Thus, the question vs. statement differences with regard to the third and fourth spectral moment are found exclusively in whispered speech.

Furthermore, intensity is higher in questions than in statements in both sibilants across all three speech modes. Similarly, the highest spectral peak is found at higher frequencies in questions than statements (with the exceptions of affricates in semi-whispered speech).

Finally, regarding the spectral regression line slope m1, the spectra of semi-whispered and normal speech rise more steeply in the frequency range of 500-3000Hz than the spectra of whispered speech. However, a significant m1 difference between questions and statements is found in whispered fricatives and affricates where questions are produced with steeper m1 slopes in comparison to statements. Regarding the spectral regression line slope m2 (from 3kHz to 11 kHz), the spectra for questions in whispered and semi-whispered fricatives fall more steeply above 3kHz compared to the spectra of statements. For affricates, the spectra of questions also fall more steeply above 3kHz than statements but the difference in steepness was largest in whispered followed by normal and then semi-whispered speech modes.

5. Conclusions

In summary, our results point to differences in spectral properties of not only vowels but also consonants when statements and polar questions are produced. These differences, especially with regard to consonants, are intensified when speakers switch from normal speech to whispered speech. In particular, questions are not only realized with a higher F1 and F2 of the vowel but also with a higher intensity, higher spectral peak frequency, higher COG, and SD as well as lower kurtosis and skewness of the consonant. Further differences are found in the spectral regression line slopes.

Some spectral differences between the production of questions and statements are found exclusively, or to the greatest extent, in whispered speech, emphasising a relevance of these cues for this particular speech mode. Moreover, the more pronounced role of these cues in whispered speech suggests a compensation for the distinguishing role of F0 in phonated speech mode. The perceptual relevance of this finding requires further investigation.

This study also sheds further light on the interaction of segments and intonation. It shows, in fact, that taking intonation patterns into consideration seems to be indispensable if spectral properties of segments - in this case sibilants - are to be investigated. Differences in intonation patterns significantly change consonantal properties. This conclusion applies to all speech modes: whispered, semi-whispered and normal.

Finally, due to the investigation of different speech modes, the study provides new aspects for a discussion of the phonological concept of intonation. It shows that speakers produce the intended intonation patterns, encoded at the underlying level, by varying a choice of acoustic cues, i.e., cues-trading as well as their magnitude in dependence on both (i) speech modes and (ii) intonation patterns.

6. Acknowledgements

This research has been supported by Bundesministerium für Bildung und Forschung (BMBF, Germany) Grant Nr. 01UG1411 to Marzena Żygis. The study was also funded by FEDER through the Operational Program Competitiveness Factors – COMPETE and by Portuguese National Funds through FCT – Foundation for Science and Technology in the context of the project FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011) to IEETA to Luis Jesus and Daniel Pape, and the post-doctoral fellowship from FCT (Portugal) SFRH/BPD/48002/2008 to Daniel Pape.

7. References

[1] K. Kohler (ed.), "Bridging the Segment-Prosody Divide in Speech Production and Perception", *Phonetica,* vol. 69, 2012.

[2] Boersma, P. & D. Weenink, David (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.57, retrieved 27 October 2013 from http://www.praat.org/.

[3] MathWorks, "Signal Processing Toolbox 6 User's Guide", N. MathWorks, Ed., 2007.

[4] M. Żygis, D. Pape, and L. M. T. Jesus, "(Non)retroflex Slavic affricates and their motivation. Evidence from Czech and Polish", *Journal of the International Phonetic Association,* 42: 281-329, 2012.

[5] L. M. T. Jesus and C. Shadle, "A parametric study of the spectral characteristics of fricatives," *Journal of Phonetics,* 30: 437-464, 2002.

[6] R Development Core Team "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, http://www.R-project.org/, version 3.0.2.

[7] G. Demenko & A. Wagner, "Prosody annotation for unit selection text-to-speech". Archives of acoustics 32(1):25-40.

[8] A. Wagner "A comprehensive model of intonation for application in speech synthesis", Ph.D. thesis, Poznań. 2008.

[9] W. Heeren and V.J. Van Heuven, "Perception and Production of Boundary Tones in whispered Dutch", Proceedings of Interspeech, 2011-2414, 2009.

[10] W. Heeren and V.J. Van Heuven, "Acoustics of whispered boundary tones: Effects of vowel type and tonal crowding", *Proceedings of the ICPhS XVII,* 851-854, 2011.

[11] X.-L., Li and B-L. Xu, "Formant comparison between whispered and voiced vowels in Mandarin", *Acta Acustica* 91: 1079-1085, 2005.

[12] W. Meyer-Eppler, "Realization of prosodic features in whispered speech", *Journal of the Acoustical Society of America,* 29: 180-182, 1957.

[13] I.B. Thomas, "Perceived pitch of whispered vowels" Journal of the Acoustical Society of America (46): 468-470, 1969.

[14] O. Niebuhr, "At the edge of intonation: The interplay of utterance-final F0 movements and voiceless fricative sounds", *Phonetica* (69), 7-27, 2012.

[15] O. Niebuhr, C.Lill and J.Neuschulz, "At the segment-prosody divide: The interplay of intonation, sibilant pitch and sibilant assimilation," *Proceedings of the ICPhS XVII*, 1478-1481, 2011.

Intonational Aspects of Imperatives in Mexican Spanish

Alina Lausecker[1], *Annika Brehm*[1], *Ingo Feldhausen*[1,2]

[1] Goethe-Universität Frankfurt am Main, Germany
[2] UMR 7018-LPP, Université Paris-Sorbonne Nouvelle, France
alinalausecker@hotmail.de, a.brehm@stud.uni-frankfurt.de, ingo.feldhausen@gmx.de

Abstract

This paper sheds new light on the intonation of imperatives in Mexican Spanish. Results from a production experiment based on scripted speech show that imperative sentences have two different nuclear configurations depending on the position of the imperative verb (V_I): (i) (L+)H* L% with V_I in sentence-final position, and (ii) L* L% with V_I in non-final position. The pitch accent on V_I in non-final position is characterized by a late peak (L+>H*). However, if the sentence is uttered with some sort of emphasis, the nuclear configuration in the non-final context can also be rising. While these results partly confirm claims made concerning the nuclear configuration in [1], they contradict the findings in [2], who attested strong pitch accent variation on V_I.

Index Terms: Intonation, Imperatives, Mexican Spanish, Spanish ToBI, Pitch Accent, Nuclear Configuration

1. Introduction

While intonational aspects of imperatives in Peninsular Spanish have attracted considerable attention (e.g. [3], [4], [5], [6]), there are only few studies on Mexican Spanish which also follow different goals. Work [2], for example, examines the pitch accent realization on V_I in sentence-initial position (as in *Abre* in *¡Abre la puerta!* 'Open the door!'). He does not consider the nuclear configuration (i.e. the nuclear pitch accent plus the following boundary tone), which is located on the sentence final DP *la puerta* 'the door'. His results show a three-way variation in the pitch accent on V_I (a predominant *late H peak*, but also an *early H peak* and *a peak with no distinct rise*; [2], p. 355). [1], in turn, concentrates solely on the nuclear configuration, thereby neglecting the pitch accent on V_I, since V_I is non-final. They transcribe the nuclear configuration as L+H* L%. Despite the thorough study, the results in [1] are typically based on imperatives which include final elements such as *ahorita mismo* in *¡Ven aquí, ahorita mismo!* 'Come here, right now' (see [1], p. 340). In this example, a prosodic boundary (H-) occurs between *ven aquí* and *ahorita mismo* (with another between the V_I *ven* and the adverb *aquí*). For this reason, it is not clear whether the elements after V_I are actually part of the imperative or whether they should be understood as added, extra-sentential elements. Thus, the question arises as to whether [1] really transcribe the imperative intonation or merely the intonation of some added, facultative material. Both studies ([1] and [2]) share the common characteristic of failing to consider V_I in sentence-final position, and as such it is unclear whether their claims also hold for V_I in final position. In a pilot study based on semi-spontaneous speech (and not on scripted speech, as is done here), [7] show that the nuclear configuration of imperatives actually differs depending on the position of V_I (final V_I: L+H* L% vs. non-final V_I: L* L%). However, their study is based on a very small set of data in which only thirteen instances of final V_I occur (such as *¡Cállate!* 'Be quiet!'). While the utterances with non-final, here sentence-initial V_I, have the form of the sentences used in [2], all utterances with final V_I contain only one metrically strong position. The pilot study thus seems to confirm the assumption that long declaratives (consisting of at least two metrically strong positions) in Spanish and Catalan are realized by a rising prenuclear accent and a low or falling nuclear configuration, while short declaratives (with only one metrically strong position) are realized by a rising nuclear accent and a low boundary tone (see [8]). The question arises as to whether V_I in sentence-final position also shows a rising accent when preceded by other material. We are not aware of any previous study addressing this issue in Mexican Spanish.

This paper unifies different perspectives and examines both (a) the nuclear configuration of imperative utterances in Mexican Spanish with V_I in sentence-final and sentence-initial position and (b) the pitch accent on the imperative verb. Three hypotheses H1-3 are established:

- *Hypothesis 1:* short imperatives, i.e. declaratives consisting of only an imperative verb (such as *¡Dámelo!* 'Give it to me!'), are realized with a rising nuclear accent (in accordance with [8]) and as such they don't show any pitch accent variation (as opposed to [2]).
- *Hypothesis 2:* Long imperatives (such as *¡Dame la mermelada!* 'Give me the marmalade!') should be realized with a low or falling nuclear configuration and a rising accent on V_I (in accordance with [8], but contradicting [1]); consequently, we do not expect any pitch accent variation either (in contrast to [2]).
- *Hypothesis 3:* Utterances with sentence-final V_I preceded by additional material are expected to be realized by a low or falling nuclear contour as in the case of long imperatives. There are several studies which question whether an imperative intonation really exists in Spanish ([3], [9], [2], [6]). If these authors are on the right track, long utterances with V_I in final position should not behave differently than long declaratives.

2. Methodology

A production experiment based on scripted speech was conducted in which three monolingual, native speakers of Mexican Spanish (one female from Torreón aged 23 [TF], one female and one male from México DF aged 25 and 33 [MF, MM]) uttered a total of 180 sentences (3 subjects x 30 target sentences x 2 repetitions). The material consisted of three conditions, C1-3 (which correspond to hypotheses 1-3):

- *Condition 1:* short imperatives consisting of one prosodic word, as shown in (1);
- *Condition 2:* long imperatives, with V_I in sentence-initial position, as shown in (2); and
- *Condition 3:* long imperatives with V_I in sentence-final position, as shown in (3). Since imperatives are typically

characterized by omitting the preverbal subject, the preceding material in (C3) is a conditional clause.

The imperative verb V_I is marked by bold letters. The metrically strong position of V_I is underlined, while the position of the nuclear accent is indicated by capitals.

¡**MÍrala**! (1)
'Look at her!'

¡**Mira** a BÁRbara! (2)
'Look at Barbara!'

Si estás interesada, ¡**MÍrala**! (3)
'If you are interested, look at her!'

As can be seen in (1) and (3), the metrically strong position of V_I overlaps with the nuclear position when V_I is in final position. Only in (2), in which V_I is sentence-initial, can the two positions be distinguished from one another. By using sentences like that in (2), in which V_I is followed by an argument, we make sure that the prompts do not evoke additional, facultative material.

There were a total of 30 target sentences (i.e. 10 sentences per condition). These target sentences were accompanied by 15 filler clauses (consisting of simple DPs such as *Las manzanas* 'the apples' or constructions with clitic left-dislocations such as *El águila, la vendió mi hermano* 'The eagle, my brother sold it.'; (for details on the intonation of left-dislocations in Spanish see [10])). The stimuli were presented in a pseudo-randomized order on sheets of paper with approximately 6 sentences per sheet. The subjects were recorded in a quiet room and were told to read the stimuli out loud at a normal rate of speech only after having silently read and understood the sentence. The subjects read the entire set of stimuli two times. The recording sessions started with a small practice session. The data were recorded in a quiet room in Frankfurt (Germany) and Paris (France) using an audio interface (44 KHz sample frequency, 24-bits precision) and a condenser headset microphone. All recordings were stored as digitized as wav-audio files. The F0 tracks were analyzed using *praat* [11]; the pitch tracks and spectrograms were used to guide the segmentation and the text-to-tune alignment. The tonal analysis is based on the ToBI annotation system for Spanish (Sp_ToBI, [12], [13], [1]).

3. Results

3.1. Results between speakers

The results for the nuclear configuration are given in Table 1. In Condition 1 (C1, short imperatives), the metrically strong syllable of V_I is tonally realized by either a high (H*) or a rising (L+H*) nuclear pitch accent (55% and 45%, respectively). A chi-square test shows that the difference in frequency between H* and L+H* is not significant ($\chi^2 = 0.6$, df = 1, p = 0.44; [14] was used for the calculations). The IP edge tone is always low (L%). In Condition 2 (C2, long imperatives with initial V_I), the predominant nuclear accent is a low tone (L*; 36 instances, 60%), followed by rising or high accents (25% and 10%, respectively) and a few instances of falling accents (H+L*, 3 times, 5%). The difference between L* and L+H* / H* is significant ($\chi^2 = 3.947$, df = 1, p = 0.047). Again, there were only low IP edge tones (100%). In Condition 3 (C3, long imperatives with final V_I), the metrically strong position of V_I is predominantly realized by a high (H*) or a rising nuclear accent (L+H*), 46 times, 77%. There are some instances of low nuclear accents (9 times, 15%) and few cases of a falling accent (5 times, 8%). A chi-square test shows that the difference in frequency between H* / L+H* and L* / H+L* is significant ($\chi^2 = 17.067$, df = 1, $p < 0.01$), i.e. the number of high or rising accents is significantly higher than the number of low or falling accents. All intonational phrase boundaries were realized by a low edge tone (L%).

While the metrically strong syllable of V_I in C1 and C3 is simultaneously the nuclear accent in the utterance, V_I in C2 is not part of the nuclear configuration, but belongs to the prenuclear area. This is not shown in Table 1. V_I in C2 is predominantly realized by a delayed peak L+>H* (92%, 55 times) and sometimes can also be realized by a rising accent L+H* (5%, 3 times) and a high tone H* (3%, 2 times).

Table 1, *Nuclear configurations for each condition for all subjects (given in % and absolute numbers)*

C1	H* L% (55%, 33)	L+H* L% (45%, 27)		
C2	L* L% (60%, 36)	L+H* L% (25%, 15)	H* L% (10%, 6)	H+L* L% (5%, 3)
C3	H* L% (47%, 28)	L+H* L% (30%, 18)	L* L% (15%, 9)	H+L* L% (8%, 5)

Typical pitch contours for the three conditions are given in Figure 1 to 4. A typical contour for the short imperatives (C1) is given in Figure 1. The nuclear pitch accent on V_I is realized with a high plateau (H*), followed by a low IP edge tone (L%). Pitch contours for long imperatives (C2) are given in Figure 2 and Figure 3. Both contours share a pre-nuclear accent located on V_I, which is realized by a rising tone with a delayed peak (L+>H*). Additionally, the IP edge is realized by a low tone (L%). The two contours differ in the shape of the nuclear accent. While it is low in Figure 2, it is rising (L+H*) in Figure 3. As a consequence, the contours differ with respect to their nuclear configuration (L* L% vs. L+H* L%).

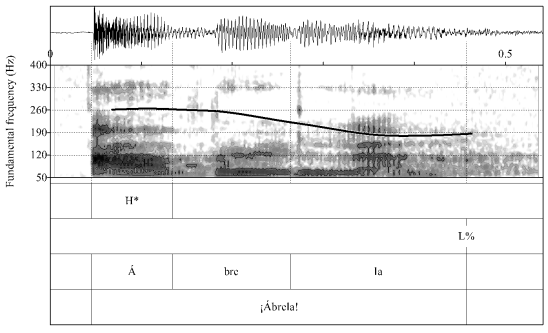


Figure 1: *Waveform, spectrogram, and F0 trace for the short imperative (C1)* ¡Ábrela! *'Open it!' of speaker TF (sentence 2_17), produced with a high nuclear pitch accent (H*) and a low edge tone (L%).*

A pitch contour for long imperatives with final V_I (C3) is given in Figure 4. The conditional clause is composed of rising accents (L+>H*, L+H*) and a high edge tone (H-). A pause (of 132ms) separates the following V_I from the conditional clause. The nuclear configuration on V_I is composed of a high

tone (H*) and a low IP edge tone (L%). As can be seen in Table 1, the typical nuclear configurations are no different from those in conditions 1 and 2 (cf. the contours on V_I in Figure 1 and in Figure 4). C3 differs from C1 and C2, however, through the presence of a sentence-internal break: While all sentences within conditions C1 and C2 show no sentence internal break, there is always a break in the sentences in C3. This break is located between the conditional clause and the imperative main clause, and is characterized phonetically by a rising contour at the end of the conditional clause, followed by a pause. The mean duration of the pause for all speakers is 236ms (not normalized).

Finally, we comment on some additional differences observed between the three conditions. The difference in frequency between H* / L+H* in C1 and H* / L+H* in C2 is significant ($\chi^2 = 57.778$, df = 1, p = 0), as is the difference between H* / L+H* in C1 and H* / L+H* in C3 ($\chi^2 = 15.849$, df = 1, p < 0.01). This means that the frequency of high and rising accents is significantly higher in C1 than in C2 or C3.

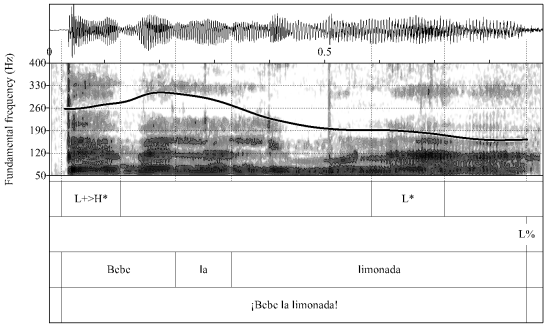


Figure 2: *Waveform, spectrogram, and F0 trace for the long imperative (C2)* ¡Bebe la limonada! *'Drink the lemonade!' of speaker TF (sentence 2_25), produced with a delayed peak (L+>H*) on V_I and a low nuclear configuration (L* L%).*

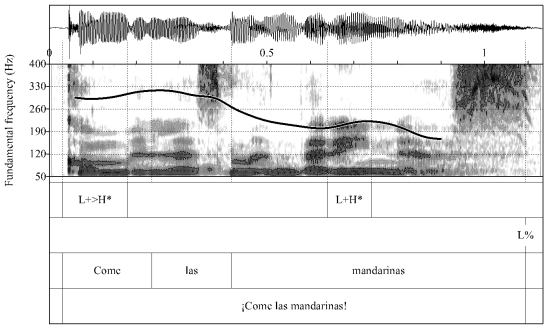


Figure 3: *Waveform, spectrogram, and F0 trace for the long imperative (C2)* ¡Come las mandarinas! *'Eat the tangerines!' of speaker TF (sentence 2_21), produced with a delayed peak (L+>H*) on V_I and a rising/falling nuclear configuration (L+H* L%).*

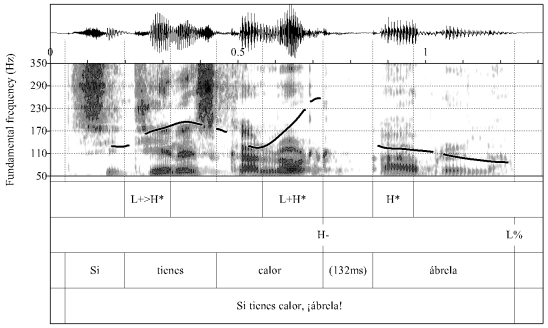


Figure 4: *Waveform, spectrogram, and F0 trace for the long imperatives with final V_I (C3)* Si tienes calor, ¡ábrela! *'If you are warm, open it!' of speaker MM (sentence 2_7), produced with a high nuclear pitch accent (H*) and a low IP edge tone (L%), preceded by a pause (of 132ms).*

3.2. Individual results

As for the nuclear configurations attested in C1, each speaker realizes the two configurations H* L% and L+H* L%, see Table 2. There are some differences, however, with respect to the frequency of the realizations. While the two speakers from México DF prefer H* L% over L+H* L% (MF: 70% vs. 30%; MM: 55% vs. 45%), the speaker from Torreón realizes H* L% less often than L+H* L% (40% vs. 60%). For the nuclear configurations attested in C2, each speaker has a repertoire of three configurations, see Table 2. Only two configurations, L* L% and L+H* L%, were realized by each speaker. The speakers from México DF additionally use H* L%, while TF uses H+L* L%. The low nuclear configuration L* L% is the prevailing choice for each speaker (MF: 60%; MM: 45%; TF: 75%). The rising-falling or circumflex contour L+H* L% is the second most common choice for the speakers from México DF, while H+L* L% is the second most common for TF.

Table 2, *Nuclear configurations for each condition and each individual speaker (given in % and absolute numbers)*

C1	MF	H* L% (70%, 14)		L+H* L% (30%, 6)	
	MM	H* L% (55%, 11)		L+H* L% (45%, 9)	
	TF	L+H* L% (60%, 12)		H* L% (40%, 8)	
C2	MF	L* L% (60%, 12)	L+H* L% (25%, 5)	H* L% (15%, 3)	
	MM	L* L% (45%, 9)	L+H* L% (40%, 8)	H* L% (15%, 3)	
	TF	L* L% (75%, 15)	H+L* L% (15%, 3)	L+H* L% (10%, 2)	
C3	MF	L+H* L% (55%, 11)	H* L% (45%, 9)		
	MM	L* L% (45%, 9)	H* L% (25%, 5)	H+L* L% (15%, 3)	L+H* L% (15%, 3)
	TF	H* L% (70%, 14)	L+H* L% (20%, 4)	H+L* L% (10%, 2)	

As for C3, there is not only clear variation between the speakers with respect to the number of different contours used (MF: 2, MM: 4; TF: 3), but also with respect to the most popular choice of the speakers. While MF and TF prefer high or rising nuclear accents, MM prefers a low nuclear accent. Furthermore, he is the only speaker to use a low nuclear configuration in C3.

Finally, in terms of the realization of the prenuclear accent located on V_I in C2, all three speakers strongly prefer the delayed peak L+>H* (MF: 90%; MM: 95%; TF 90%). As a consequence, the frequency of the other realizations (which are H* and L+H*) is very low. The delayed peak is realized significantly more often than the other tones across the speakers ($\chi^2 = 41.667$, df = 1, p = 0).

4. Discussion

Hypothesis 1 (short imperatives are realized with a rising nuclear accent and do not show variation) can be taken as fulfilled, since the nuclear accent is either high or rising. This supports the claims by [8] for short declaratives and contradicts [2], in which considerable tonal variation is attested on V_I. In defense of [2], however, it is important to note that short imperatives were not investigated in this work. We return to [2] in greater detail when discussing the results in the light of the second hypothesis.

Hypothesis 2 (long imperatives have a low or falling nuclear configuration and a rising accent on V_I in sentence-initial position) cannot be taken to be completely fulfilled, as the data give a rather mixed picture. Even though 65% of the data is realized by L* L% (60%) and H+L* L% (5%), 35% of the nuclear accents are either high or rising. Thus, the low or falling contours support the claims made in [8], while the high and circumflex contours show that the authors of [1] appear to be on the right track. In contrast to [1], though, no optional emphatic upstep of H was attested in our data. Additionally, our data show that long imperatives with V_I in sentence-initial position can indeed be realized with a low or falling nuclear configuration. This is in line with studies on neutral, broad focus declaratives in Mexican Spanish, which show that both contours, L* L% and L+H* L%, are typical nuclear configurations ([15], [16], and also [1]). As such, our work confirms the studies on Mexican declaratives, leading us to conclude that hypothesis 2 is simply too strict for Mexican Spanish. In terms of pitch accent variation, variation of pitch accents on V_I is possible in C2, but in contrast to the results of [2], the delayed peak is chosen by nearly all of the speakers (92%), with instances of non-delayed peaks being very rare (totaling 8%). A considerably larger number of instances of non-delayed peaks were attested in [2] (see p.359). Furthermore, [2] reports on a variation between three different pitch accents in 3 out of his 4 speakers, while in our data two speakers (MM and TF) use only two different pitch accents. In addition, the third pitch accent used by speaker MF occurs only once. Thus, as for C2, there is considerably less pitch accent variation in our data than in [2].

Hypothesis 3 (long imperatives with final V_I are realized by a low or falling nuclear configuration) was shown not to be fulfilled. The nuclear contour is typically either high or rising (77%), while the rest is low or falling (23%). In addition, the low or falling instances were mainly uttered by a single speaker (MM), while the other two speakers either did not realize any falling nuclear accent (MF) or did so to only a small extent (TF).

Due to the fact that the material preceding V_I is a conditional clause whose edge is obligatorily marked by a high edge tone accompanied by a pause, we wonder whether the entire sentence can really be considered to be a typical declarative utterance. The main clause, consisting of only V_I, behaves as the short imperative of C1. [17], [18], [19] have already noted that adjunct clauses have a considerable impact on the intonation of a sentence. This might explain the great variation in pitch accent observed between our speakers. While speaker MM varies between the low contour of typical declaratives and rising accents of short imperatives, MF and TF (almost) always choose the short imperative pattern. Interestingly, it was the male speaker who showed the greatest variation. The male speakers in [2] also use a greater number of different pitch movements.

Finally, our data confirm the results of [7], in which final V_I is shown to have a rising intonation for semi-spontaneous speech. The comparison of our study and that in [7] indicates that some differences exist with respect to the tonal realization of non-final V_I in semi-spontaneous and scripted speech with scripted speech showing less variation (see [20], [21], and [22] for studies on spontaneous speech in Spanish and a discussion of differences between laboratory and spontaneous speech).

5. Conclusions

The contribution of this paper consists in presenting a unified perspective on (a) the pitch accent located on the imperative verb and (b) the nuclear configuration of imperative sentences - a view that has not taken been before. Coming from this perspective, we show that imperative sentences in Mexican Spanish have two different nuclear configurations depending on the position of the imperative verb: (i) (L+)H* L% with V_I in sentence-final position (as in short imperatives and imperatives preceded by a conditional clause), and (ii) L* L%, with V_I in non-final position (as in long imperatives). Long imperatives can also have a rising accent, which is in line with [1]. However, in contrast to [1], the circumflex nuclear configuration is not the prevailing choice of the speakers for (long) imperatives. In contrast to [2], we could not attest strong variation in pitch accent on V_I across the conditions. Our speakers strongly preferred the delayed peak L+>H* in long imperatives and hardly realized the other pitch accents. High/rising accents such as H* or L+H* are either the most popular or sole choice in short imperatives (C1) and in long imperatives with final V_I (C3).

In this study, we hoped to answer the question of whether V_I in sentence-final position also shows a rising accent when preceded by other material. Our data were able to confirm this. Nevertheless, it would be interesting to see whether high or rising accents also occur when V_I is preceded by material other than a conditional clause. If so, this might shed further light on the ongoing discussion on whether differences exist on the tonal level between declarative and imperative intonation.

6. Acknowledgements

We would like to thank Pilar Prieto and the two anonymous reviewers for supportive comments. We also thank Fabián Santiago Vargas, Ellenit Hernández Mendoza, and our subjects for their help during the recording session. Moreover, our gratitude goes to Audrey MacDougall for her assistance with editing.

7. References

[1] De-la Mota, C., Butragueño, P. M. and Prieto, P., "Mexican Spanish Intonation", in P. Prieto and P. Roseano [Ed], Transcription of Intonation of the Spanish Language, 319-350, München, 2010.

[2] Willis, E. W., "Is there a Spanish imperative intonation revisited: local considerations", Linguistics 40(2): 347-374, 2002.

[3] Navarro Tomás, T., "Manual de entonación española", New York, 1944.

[4] De-la-Mota, C., "La representación gramatical de la información nueva en el discurso", [Unpublished Ph.D. dissertation], Barcelona, 1995.

[5] Estebas-Vilaplana E. and Prieto, P. "Castilian Spanish Intonation", in P. Prieto and P. Roseano [Ed], Transcription of Intonation of the Spanish Language, 17-48, München, 2010.

[6] Robles-Puente, S., "Looking for the Spanish Imperative Intonation", in S. Alvord [Ed], Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology, 153–164, Somerville, 2011.

[7] Brehm, A., Lausecker, A. and Feldhausen, I., "The Intonation of Imperatives in Mexican Spanish", Proceedings of the 10th International Seminar on Speech Production, Köln (Germany), 5-8 May, 2014.

[8] Prieto, P., "Tune-text association patterns in Catalan: An argument for a hierarchical structure of tunes", Probus 14: 173-204, 2002.

[9] Kvavik, K. "Is there a Spanish imperative intonation?", in H. and M. C. Resnick [Ed], Studies in Caribbean Spanish Dialectology, 35-49, Washington, D.C., 1988.

[10] Feldhausen, I., "The Relation between Prosody and Syntax: The case of different types of Left-Dislocations in Spanish", in M. Armstrong, N. Henriksen and M. Vanrell [Ed], Interdisciplinary approaches to intonational grammar in Ibero-Romance intonation [Issues in Hispanic and Lusophone Linguistics], Amsterdam, to appear.

[11] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", [Computer program]. Online: http://www.praat.org, accessed on 22 Dec 2013.

[12] Aguilar, L., De-la-Mota, C. and Prieto, P. [Ed], "Sp_ToBI Training Materials". Online: http://prosodia.upf.edu/sp_tobi/, accessed on 22 Dec 2013.

[13] Prieto, P. and Roseano, P. „Transcription of Intonation of the Spanish Language", München, 2010.

[14] Preacher, K., "Calculation for the chi-square test", [Computer software]. Online: http://quantpsy.org, accessed on 23 Dec 2013.

[15] Sosa, J. M., "La entonación del español. Su estructura fónica, variabilidad y dialectología", Madrid, 1999.

[16] Willis, E. W., "Tonal Levels in Puebla Mexico Spanish Declaratives and Absolute Interrogatives", in R. Gess and E. Rubin [Ed], Theoretical and Experimental Approaches to Romance Linguistics, 351–363, Amsterdam, 2005.

[17] Bolinger, D., "Intonational Signals of Subordination", Proceedings of the Tenth Annual Meeting of the Berkeley Linguistics Society, 401-413, 1984.

[18] Chafe, W., "How people use adverbial clauses", Berkeley Linguistics Society 10: 437-449, 1984.

[19] Chafe, W., "Linking intonation units in spoken English", in J. Haiman and S.A. Thompson [Ed], Clause combining in grammar and discourse, 1-27, Amsterdam, 1988.

[20] Face, T., "Intonation in Spanish declaratives: differences between lab speech and spontaneous speech", Catalan Journal of Linguistics 2: 115–131, 2003.

[21] Feldhausen, I., Benet, A. and Pešková, A., "Prosodische Grenzen in der Spontansprache: Eine Untersuchung zum Zentral-katalanischen und porteño-Spanischen", Working Papers in Multilingualism 94: Series B, 2011.

[22] Pešková, A., Feldhausen, I., Kireva, E. and Gabriel, C., "Diachronic prosody of a contact variety: Analyzing Porteño Spanish spontaneous speech", in K. Braunmüller and C. Gabriel [Ed], Multilingual Individuals and Multilingual Societies [HSM 13], Amsterdam, 365-390, 2013.

The Acquisition of English Lexical Stress by Cantonese-English Bilingual Children at 2;06 and 3;0

Jingwen Li, Peggy Pik Ki Mok

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
joanneljw@gmail.com, peggymok@cuhk.edu.hk

Abstract

This study investigates the acquisition of English lexical stress by simultaneous Cantonese-English bilingual children at the age of 2;06 and 3;0 respectively, comparing them with the English monolingual peers. Research on early bilingual phonological acquisition often focuses on segmental level. Few studies are available when it concerns prosodic features, especially in children speaking non-Indo-European languages. This study examines an important prosodic feature, lexical stress, in Cantonese-English bilingual children. The results showed that there is delayed acquisition of English lexical stress among the bilingual children, as reflected in less contrastive syllable duration and peak F0, possibly due to a lack of lexical stress in Cantonese, a typical syllable-timed language. This study helps to understand the bilingual interaction of two distinctive prosodic systems, and broaden our knowledge about early bilingual prosodic development.

Index Terms: lexical stress, Cantonese, English, bilingual acquisition

1. Introduction

Lexical stress (word stress) is the stress placed on a given syllable, and the assignment of lexical stress in words is language-specific. For example, in languages like Czech and Finnish, stress is always assigned on the first syllable of a word; while in some other languages like English and Russian, the position of stress in a word is less predictable. In English, there are two main types of stress patterns in English: trochaic and iambic [1]. Trochaic word refers to a disyllabic word with stressed-unstressed pattern (e.g. baby); iambic word refers to an unstressed-stressed pattern (e.g. behind). The acoustic correlates of lexical stress in English have been examined extensively ([4][7][11][12]). Most of these studies focused on minimal pairs, in which the placement of stress determines whether the word is a noun or a verb. The results consistently indicate that within a lexical word, the stressed syllable has higher fundamental frequency (F0), longer syllable duration, and greater intensity than the unstressed syllable. Besides, vowel quality has also been considered as an important acoustic correlate of English lexical stress, and failure in reducing vowels in unstressed syllable contributes to non-native accent ([10][14][24]).

Previous studies suggested that English-speaking children's early acquisition of lexical stress involves enlarging the stress-unstressed ratios in all acoustic correlates (F0, syllable duration and intensity), especially by reducing the unstressed syllable duration but maintaining the stressed syllable duration ([1][8][19][21]). An equally important question is how bilingual children at a young age acquire lexical stress. However, there is a dearth of empirical data in this line of research. Most previous studies focused on preference of word truncation by bilingual children ([18][20][22]). For example, [18] conducted a nonsense-word repetition task in French-English bilingual children. English and French contrast in that the majority of English words have a trochaic rhythm while French words have an iambic rhythm. [18] found that the English-dominant bilinguals tend to preserve trochaic pattern and the French-dominant bilinguals tend to preserve iambic pattern, indicating cross-linguistic effects and the prominent influence of language dominance on the directionality of the effects. In contrast, based on an English-French bilingual child's speech, [20] found no evidence for trochaic bias in the data. By far, few studies have used an acoustic approach to investigate prosodic development in young bilingual children, especially before the age of three. Mok ([16] [17]) investigated the acquisition of speech rhythm in Cantonese-English bilingual children at the age of 2;06 and 3;0. She found that the bilinguals displayed less variable syllable duration and less vowel reduction in unstressed syllables, compared with their English monolingual peers. Contrastive syllable duration and vowel reduction are important indicators for lexical stress, too. Therefore, Mok's studies ([16][17]) provide insights into the present study that cross linguistic effects in the acquisition of lexical stress in Cantonese-English bilingual children can also be found.

Unlike English, Cantonese does not make use of lexical stress, and neither has it phonological vowel reduction. As a tone language, F0 in Cantonese is primarily used to differentiate lexical meaning. Additionally, Cantonese is a typical syllable-timed language ([15]). It is thus worth investigating whether the different experience with F0, syllable duration and vowel quality in Cantonese would influence the acquisition of English lexical stress in simultaneous Cantonese-English bilingual children. Given that the period between age 2;0 and 3;0 is an important stage for children's prosodic and lexical development, the present study examines the acoustic correlates (F0 and syllable duration) of lexical stress in Cantonese-English bilingual children and in English monolingual children at 2;06 and 3;0 respectively.

2. Method

2.1. Subjects

Based on the age of the children, there are two groups of data: 2;06 and 3;0. In the 2;06 group, there are seven simultaneous Cantonese-English bilingual children and five English monolingual children; in the 3;0 group, there are eight bilingual children and six monolingual children. The bilingual children all came from the YipMathews corpus in CHILDES (http://childes.psy.cmu.edu/media/Biling/YipMatthews/). Yip and Matthews [23] gave detailed introduction to the background of the children. The data of monolingual children came from various sources, which will be introduced in the following sections.

Table 1. *Background information of the bilingual children.*

Child	Input languages	Sex	Language dominance	Age range for data used	
				2;06	3;0
B1	BrE/Cantonese	M	Cantonese	2;5.12-2;7.07	2;11.12-3;1.13
B2	BrE/Cantonese	F	Cantonese	2;5.16-2;7.01	2;11.05-3;0.09
B3	BrE/Cantonese	F	Cantonese	2;6.02-2;7.28	2;11.19-3;0.24
B4	BrE/Cantonese	F	Cantonese	---------	2;11.27-3;0.18
B5	BrE/Cantonese	F	English	2;5.19-2;6.16	2;10.29-3;0.03
B6	BrE/Cantonese	M	Cantonese	2;6.20-2;7.04	2;11.29-3;0.27
B7	HKE/Cantonese	M	Cantonese	2;5.05-2;7.00	2;11.05-3;0.03
B8	HKE/Cantonese	M	Cantonese	2;4.29-2;7.24	2;10.03-3;2.03

2.1.1. Bilingual children

Table 1 shows the background information of the bilingual children. Six of them, B1, B2, B3, B4, B5 and B6, were exposed to Cantonese and English from birth and grew up in a 'one parent one language' environment, with one parent being a native speaker of British English and the other a native speaker of Cantonese. The other two children, B7 and B8, grew up in Hong Kong families and were exposed to Hong Kong English and Cantonese from birth. Except for B5, all the other children were Cantonese dominant.

2.1.2. Monolingual children

Table 2 shows the background information of the monolingual children. Their parents are all British English native speaker.

Table 2. *Background information of the monolingual children.*

Child	Input Languages	Sex	Age of recordings	
			2;06	3;0
M1	British English	F	2;5-2;7	2;11
M2	British English	F	---	3;1
M3	British English	F	2;7.01	3;1
M4	British English	M	2;7.01	3;0
M5	British English	M	---	3;0
M6	British English	F	2;5-2;6	---
M7	British English	M	2;6.00-2;6.22	3;0

Data of M1 and M7 came from the Forrester corpus ([6]) and the Thomas corpus ([13]) in CHILDES respectively. M1's natural conversations were recorded and all participants involved in the dialogues were British, white, and middle class. M7 was born in a middle class family and he was primarily cared for by his mother. The frequency of data between 2;00,12 and 3;00,12 is very intensive, and during this period, M7 was recorded for one hour each time, five times a week, every week for the entire period. Two of the English speaking children M2 and M5 were recruited in an English-medium kindergarten in Hong Kong for children from expatriate families. The other two English children M3 and M4 were twins from an expatriate family living in Hong Kong.

2.2. Materials

Disyllabic words of both trochaic (strong-weak) and iambic (weak-strong) patterns were used. We have gone through every video/audio recording, extracted the target words, and saved them as .wav files for acoustic analyses. The quality of some recordings was not very good, so we only used interpretable utterances that have clear formant structure in the spectrogram. To exclude excessive initial F0 raising and duration shortening, as well as final lengthening and lowering, we only chose disyllabic words in sentence-medial position, which means the target word together with the preceding word and the following word all come from the same intonational phrase. We also excluded words with excessive stress on one of the syllables. For example, in many cases, the child was shouting, singing, crying or arguing with his/her siblings. Utterances produced under such occasions were not used. Since corpus data were used, there may be segmental and sentential context effects that affect the acoustic properties of the targets words, but this is unavoidable.

The number of the extracted utterances varies from child to child, but it does not affect the overall results because the F0 ratios and duration ratios were averaged across all the tokens from the same child. We use stressed/unstressed ratios for cross group comparison because ratios can demonstrate how contrastive the stress pattern is, and more importantly, they normalize the data for individual variation. We obtained many more trochaic words than iambic words. The imbalance of the two types of disyllabic words can be explained by the fact that there are many more trochaic words than iambic words in English ([9]).

2.3. Measurements

Each disyllabic word was labeled, and the syllable duration and peak F0 of the vowel of each syllable were measured with Praat ([3]).

When measuring syllable duration, if the word begins or ends with a stop, then the closure phase of the stop would be excluded because there is no reliable cue for marking it ([14][16]). Each disyllabic word was segmented into two syllables, and word-medial consonants were segmented based on the maximal onset principle. For example, the /p/ in 'paper' was treated as the onset of the second syllable. Peak F0, which is the highest point on the F0 contour of the vowel, was automatically tracked by Praat. After the values of syllable duration and syllable peak F0 of all the disyllabic words had been measured, stressed-/unstressed-syllable ratios were calculated, and then they were averaged across tokens for each child.

3. Results

3.1. Children at 2;06

The stressed/unstressed ratios of syllable duration in trochaic words for the 2;06 years old bilingual and monolingual children are shown in Table 3. Calculating ratio means the value of the stressed syllable duration is divided by the value of the unstressed syllable duration. Therefore, a value >1 means that the stressed syllable duration is longer than the unstressed syllable duration (as expected), and vice versa.

It can be seen that for trochaic words, the stressed and unstressed syllable duration ratios are comparable across the bilingual children, hovering around 1. In contrary, the stressed syllable durations are consistently higher than the unstressed syllable durations within the same disyllabic words for monolingual children. An independent-samples t-test was conducted to compare the ratios in the bilinguals and the monolinguals. There was a significant difference in stressed and unstressed syllable duration ratios between the bilinguals and the monolinguals [$t(10) = -5.50$, $p<0.01$]. The results showed that the difference between stressed and unstressed syllable duration is larger in monolingual than bilingual children.

Table 3. *Syllable duration ratios (s.d.) for bilingual children: 2;06.*

Bilinguals	Duration ratio	Monolinguals	Duration ratio
B1	1.12(0.42)	M1	1.35(0.24)
B2	1.19(0.36)	M3	1.64(0.35)
B3	0.94(0.32)	M4	1.57(0.48)
B5	0.93(0.21)	M6	1.65(0.57)
B6	1.08(0.28)	M7	1.25(0.31)
B7	1.01(0.30)	---	---
B8	1.09(0.23)	---	---
Mean	1.05(0.10)	Mean	1.49(0.18)

Table 4. *Peak F0 ratios (s.d.) for bilingual children: 2;06.*

Bilinguals	Peak F0 ratio	Monolinguals	Peak F0 ratio
B1	1.03(0.10)	M1	1.75(0.18)
B2	1.00(0.04)	M3	1.04(0.05)
B3	1.24(0.69)	M4	1.17(0.42)
B5	1.15(0.19)	M6	1.15(0.25)
B6	1.04(0.06)	M7	1.04(0.19)
B7	1.08(0.44)	---	---
B8	1.02(0.03)	---	---
Mean	1.08(0.09)	Mean	1.23(0.30)

The stressed/unstressed ratios of peak F0 in trochaic words for all the children are listed in Table 4 and the data show more individual variation. Among the bilingual children, B3 (1.24) had a larger difference in peak F0 between stressed and unstressed syllables, while the others had comparable stressed and unstressed peak F0, hovering around 1. The monolinguals did not show as distinct a stress pattern in peak F0 as they did in syllable duration, e.g., M1 having very contrastive peak F0 (1.75) while M7 showing similar F0 peaks (1.04). Overall, in terms of peak F0, the monolingual children did not show more distinct pattern than the bilingual children. Independent t-test confirmed that the difference between the bilinguals and the monolinguals was not significant [$t(10) = -1.28$, $p>0.05$].

In the 2;06 age group, iambic words were not analysed because of the lack of data. For example, in all the monolingual speech, only two iambic words were found in M6's utterances.

3.2. Children at 3;0

3.2.1. Trochaic words

The stressed/unstressed ratios of syllable duration of trochaic words for the bilingual and monolingual children at the age of 3;0 are shown in Table 5.

Table 5. *Syllable duration ratios (s.d.) for bilingual children: 3;0.*

Bilinguals	Duration ratio	Monolinguals	Duration ratio
B1	1.02(0.26)	M1	1.71(0.30)
B2	1.15(0.28)	M2	1.60(0.52)
B3	1.15(0.31)	M3	1.98(0.71)
B4	1.02(0.23)	M4	1.67(0.52)
B5	1.06(0.40)	M5	1.63(0.38)
B6	0.99(0.46)	M7	1.69(0.63)
B7	1.32(0.51)	---	---
B8	0.99(0.30)	---	---
Mean	1.09(0.11)	Mean	1.71(0.14)

Table 6. *Peak F0 ratios (s.d.) for bilingual children: 3;0.*

Bilinguals	Peak F0 ratio	Monolinguals	Peak F0 ratio
B1	0.99(0.12)	M1	1.17(0.24)
B2	0.99(0.15)	M2	1.26(0.66)
B3	1.14(0.40)	M3	0.97(0.19)
B4	1.08(0.13)	M4	1.10(0.24)
B5	1.09(0.23)	M5	1.23(0.37)
B6	1.01(0.12)	M7	1.15(0.40)
B7	1.07(0.36)	---	---
B8	1.13(0.24)	---	---
Mean	1.06(0.06)	Mean	1.15(0.10)

It can be seen that the stressed-unstressed syllable duration ratios are comparable among bilingual children. Although B7 (1.32) appears to have more distinct stress pattern than the other bilingual children, his duration ratio is still lower than the lowest value for monolingual children (M2: 1.60). On the other hand, the stressed/unstressed ratios of syllable duration in monolingual children are consistently higher than those in the bilingual children, suggesting that the monolinguals have more contrastive stressed/unstressed syllable durations than the bilinguals. An independent-samples t-test confirms the significant difference in stressed and unstressed syllable duration ratios between the bilinguals and

the monolinguals [t(12) = -9.48, p<0.01]. The results showed that the difference between stressed and unstressed syllable duration is larger in monolingual than bilingual children at the age of 3;0.

The stressed/unstressed ratios of peak F0 in trochaic words for all the bilingual and monolingual children are listed in Table 6. The values are comparable across the bilingual children, and the condition is similar for the monolingual children, except that M2 (1.26) and M5 (1.23) seem to have more contrastive stressed/unstressed peak F0. Independent t-test indicated no significant difference between the bilingual group and the monolingual group in terms of ratios of peak F0 [t(12) = -1.93, p>0.05].

3.2.2. Iambic words

The number of iambic words in the 3;0 group is also very limited. All the iambic words are listed in Table 7, and the value in bracket is the number of tokens obtained from all the utterances of all the children.

Table 7. *Iambic words: 3;0.*

Bilingual		Monolingual
about (3)	*behind* (1)	*about* (6)
alright (1)	*cannot* (1)	*again* (1)
around (1)	*forgot* (2)	*around* (2)
away (2)	*Michelle* (3)	*because* (1)
because (6)	*upstairs* (1)	*behind* (1)

It can be seen in Table 7 that the majority of the iambic words are function words. Given the small number of them, no statistics can be used to compare the stressed/unstressed ratios in terms of peak F0 and syllable duration between the two groups of children. Nevertheless, the average stressed/unstressed syllable duration ratio is much larger in the monolinguals (2.72) than that in the bilinguals (1.55); while the average ratios in peak F0 are comparable across the two groups (monolingual: 1.18; bilingual: 1.13). The observed patterns of iambic words confirm the findings in trochaic words very well that monolingual children displayed more distinct stress pattern than the bilingual children did, demonstrated by more contrastive syllable duration. It is interesting to note that, the bilinguals seem to perform better in contrasting stressed syllable duration from unstressed syllable duration in iambic words than in trochaic words. It is possible that because the iambic function words are more frequently heard in the input language, and so acquired better by the bilingual children.

3.3. Cross-age comparison

Besides cross language comparisons, cross age comparisons were also carried out within each group of children. For instance, paired sample t-test were conducted to compare the stressed and unstressed syllable duration ratios in the monolingual children at 2;06 and that in the same children at 3;0. The results show that the stressed and unstressed syllable durations are more contrastive in the monolinguals at 3;0 (M = 1.76, SD = 0.14) than at 2;06 (M = 1.45, SD = 0.18), with a significant difference [t(3) = -4.2, p<0.01]. Interestingly, this was the only significant result in all the cross age comparisons. It suggested that English monolingual children were enlarging the contrast between the stressed and unstressed syllable durations between 2;06 and 3;0, but the development of syllable durational contrast is much slower in Cantonese-English bilingual children.

4. Discussion

This study aims to investigate whether the linguistic experience of Cantonese would affect the acquisition of English lexical stress in simultaneous Cantonese-English bilingual children. Data of eight Cantonese-English bilingual children and six English monolingual children were used, including two ages: 2;06 and 3;0.

In both age groups, monolingual children displayed more contrastive syllable duration than the bilingual children did in disyllabic words, and the difference is significant. Both monolingual and bilingual children have comparable stressed and unstressed syllable peak F0. It was expected that Cantonese-English bilingual children would use F0 to contrast stressed/unstressed syllable better than the monolinguals, since they have more experience with F0 variation in Cantonese. However, the finding that even English monolingual children did not contrast stressed/unstressed peak F0 suggests that, syllable duration, rather than peak F0, is the primary cue to distinguish stressed syllable from unstressed syllable by children, at least before the age of 3;0. Unlike the monolingual children, the Cantonese-English bilingual children do not show clear stress pattern in either of the acoustic correlates, and they have less reduction in unstressed syllable duration. Besides, cross age group comparison indicated the development of English lexical stress pattern is much slower in the Cantonese-English bilingual children between the age of 2;06 and 3;0, when they are compared with the English monolingual children.

The results confirm our prediction, and the delayed development of lexical stress is possibly due to the fact that Cantonese lacks lexical stress and phonological reduction. Lack of durational variation in Cantonese has affected the acquisition of durational contrast in English. But the finding that both bilinguals and monolinguals have comparable stressed and unstressed peak F0 suggests that, first, F0 is not the primary cue for lexical stress contrast in children by the age of 3;0; second, though the bilingual children have more intricate use of F0 in Cantonese tones, they had not applied it in contrasting English lexical stress.

Further investigation is still required to examine whether and when the bilingual children can catch up with the English monolingual children and acquire adult-like lexical stress pattern.

5. Conclusions

The results of the current study have broadened our knowledge about early bilingual acquisition of lexical stress from several aspects. Firstly, for English monolingual children before the age of three, syllable duration, rather than F0, is the primary cue for lexical stress contrast. Secondly, Cantonese-English bilingual children have a delay in developing syllable durational contrast for clear lexical stress pattern. The delayed development is possibly due to the fact that Cantonese, as a typical syllable-timed language, lacks lexical stress, and the experience with less variable syllable duration causes negative cross-linguistic effects.

6. References

[1] Allen, G. and Hawkins, S., ‘‘Phonological rhythm: Definition and development,’’ in Child Phonology, edited by G. Yeni-Komshian, J. Kavanagh, and C. Ferguson (Academic, New York), Vol. 1, 1980

[2] Bauer, R. S. and Benedict, P. K., Modern Cantonese phonology. Berlin: Mouton de Gruyter, 1997.

[3] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer (Version 5.1.12) [computer program]. http://www.praat.org/Boula de, 2009.

[4] Bolinger, D. L., “A theory of pitch accent in English,” Word 14, 109–119, 1958.

[5] Fletcher, P., Leung, S. C. S., Stokes, S. F. and Weizman, Z. O., Cantonese preschool language development: A guide. Hong Kong: Department of Speech and Hearing Sciences, 2000.

[6] Forrester, M., Appropriating cultural conceptions of childhood: Participation in conversation. Childhood, 9, 255–278, 2002.

[7] Fry, D. B., “Duration and intensity as physical correlates of linguistic stress,” J. Acoust. Soc. Am. 27, 765–768.10.1121/1.1908022 [Cross Ref], 1955.

[8] Goodell, E. and Studdert-Kennedy, M., ‘‘Acoustic evidence for the development of gestural coordination in the speech of 2-year-olds: A longitudinal study,’’ J. Speech Hear. Res. 36, 707–726, 1993.

[9] Giegerich, H., English phonology: an introduction. Cambridge, UK: CUP, 1992.

[10] Lee, B., Guion, S. G., and Harada, T., “Acoustic analysis of the production of unstressed English vowels by early and late Korean and Japanese bilinguals,” Stud. Second Lang. Acquis. 28, 487–513, 2006.

[11] Lieberman, P., “Some acoustic correlates of word stress in American English,” J. Acoust. Soc. Am. 32, 451–454.10.1121/1.1908095, 1960.

[12] Lieberman, P., Intonation, Perception and Language (M.I.T. Press, Cambridge, Massachusetts), 1975.

[13] Lieven, E., Salomo, D. and Tomasello, M., Two-year-old children’s production of multiword utterances: A usage-based analysis. Cognitive Linguistics, 20(3), 481-508, 2009.

[14] Low, E. L. and Grabe, E., A contrastive study of prosody and lexical stress placement in Singapore English and British English. Language and Speech, 42, 39-56, 1999.

[15] Mok, P. On the syllable-timing of Cantonese and Beijing Mandarin. Chinese Journal of Phonetics, 2: 148-154, 2009.

[16] Mok, P. P. K., The acquisition of speech rhythm by three- year-old bilingual and monolingual children: Cantonese and English. Bilingualism: Language and Cognition, 14, 458–472, 2011.

[17] Mok, P., Speech rhythm of monolingual and bilingual children at 2;06: Cantonese and English. Bilingualism: Language and Cognition, 16: 693-703, 2013.

[18] Paradis, J., Do bilingual two-year-olds have separate phonological systems? International Journal of Bilingualism, 5(1), 19-38, 2001.

[19] Pollock, K., Brammer, D. and Hageman, C., “An acoustic analysis of young children’s production of word stress,’’ J. Phon. 21, 183–199, 1993.

[20] Rose, Y. and Champdoizeau, C., There is no Innate Trochaic Bias: Acoustic Evidence in Favour of the Neutral Start Hypothesis. Language Acquisition and Development: Proceedings of GALA 2007. Anna Gavarró & Maria João Freitas (eds.). Newcastle, UK: Cambridge Scholars Publishing. 359-369, 2008.

[21] Schwartz, R. G., Petinou, K., Goffman, L., Lazowski, G. and Cartusciello, C., Young children’s production of syllable stress: An acoustic analysis.The Journal of the Acoustical Society of America, 99, 3192–3200, 1996.

[22] Vihman, M. M., DePaolis, R. A., and Davis, B. L., Is there a “Trochaic Bias” in Early Word Learning? Evidence from Infant Production in English and French. Child Development 69: 935-949, 1998.

[23] Yip, V. and Matthews, S., The bilingual child: Early development and language contact. Cambridge: Cambridge University Press, 2007.

[24] Zhang, Y. H., Nissen, S. L. and Francis, A. L., Acoustic characteristics of English lexical stress produced by native Mandarin speakers. The Journal of the Acoustical Society of America. 2008 June; 123(6)4498, 2008.

Disentangling sources of rhythmic variability between dialects

Adrian Leemann[1], Volker Dellwo[1], Marie-José Kolly[1], Stephan Schmid[1]

[1]Phonetics Laboratory, Department of General Linguistics, University of Zurich
{adrian.leemann, marie-jose.kolly, schmidst}@pholab.uzh.ch, volker.dellwo@uzh.ch

Abstract

Speech rhythm is highly variable. Previous studies reported variability between languages, dialects, speakers, and labelers. Research further revealed an effect of *sentence* in the rhythmic characteristics of speakers of the same language. In the present study we tested whether the effect of sentence material is constant across varieties of the same language. We addressed this question by an example of analyzing rhythmic variability between eight dialects of Swiss German in three different sentences. Results showed a significant interaction for *dialect*sentence* for most of the tested rhythm metrics. We take this as evidence that differences between dialects are contingent upon the sentences used in the experiment. We further investigated which sources in the sentence material caused between-dialect differences in rhythm scores to vary. We found exemplary evidence that dialect-specific phonological and morphological phenomena contained in the individual sentences are the prime suspects. Implications for future speech rhythm research are discussed.

Index Terms: Speech rhythm; dialectology; Swiss German, rhythm metrics

1. Introduction

Acoustic measures of speech rhythm that are based on temporal features of speech have been reported to vary significantly between and within languages. Yet, relatively little is known about the actual sources behind this variability. [1] suggested that metrics such as the percentage over which speech is vocalic (%V) reflect the degree of vowel reduction, and metrics such as the standard deviation of consonantal intervals (ΔC) capture syllable complexity. [2], however, provided evidence that differences in rhythm metrics of typologically different languages emerge even when syllable structure complexity is controlled for. They reported that the durational marking of prosodic heads or pre-final heads accounted for more rhythmic variability for the language set investigated. [3, 4] examined the degree to which sentence material affected rhythm scores. Both studies reported effects of *sentence*, implying that rhythm scores strongly differ depending on the sentence material being analyzed.

While these studies have shown an effect of *sentence* in the rhythm scores of speakers of the same language, it is unclear whether such effects are constant across varieties of the same language. In this preliminary study we address this question by an example of analyzing rhythmic variability between dialects of Swiss German (SwG) in different sentences. Our paper addresses the following research questions:

(1) Does sentence material have an influence on rhythmic differences between dialects?
(2) Which parameters exactly in the sentence material cause between-dialect differences in rhythm scores to vary?

To test these hypotheses, we applied rhythm metrics on the following 8 SwG dialects, see Table 1.

	West	East
Midland	**BS:** Basel	**TG:** Thurgau
	BE: Bern	**ZH:** Zurich
Alpine	**SB:** Sensebezirk	**SZ:** Schwyz
	VS: Valais	**GR:** Grisons

Table 1: Selected dialects and their abbreviations.

The sentences used in the present study are a subset of those used in [5], where the declarative intonation patterns of the dialects mentioned were examined based on 7 read sentences per speaker. Since the current study is preliminary in nature, we analyzed only 3 of [5]'s 7 sentences per speaker. The other 4 out of 7 sentences used in [5] were the basis of between-dialect rhythmic analyses in [6]. [6] reported significant differences for these 8 SwG dialects particularly in the variability of vocalic intervals, while consonantal variability was less discriminant. The current study is thus a follow-up of [6].

We hypothesize that dialect-specific phonological processes present in some but not all sentences are possible reasons as to why rhythm scores vary according to *sentence*. Consider, for instance, the following example: a typical feature of Basel SwG is extensive vowel lengthening before [r]; often the [r] is elided completely. In *stark* 'strong', for example, Basel speakers articulate [ˈʃtɑːx], while most other SwG dialect spekaers realize [r] before the [x] without vowel lengthening, resulting in [ˈʃtɑrx]. If some sentences in the data set contain words that allow for this phonological process to take place, it is conceivable that this has an effect particularly on %V and vocalic variability measures such as VarcoV, ΔV, and nPVI_V. To test this hypothesis, we selected three sentences that vary in the proportions of vowels and consonants as well as in syllabic make-up. Given these examples, we expect to find significant interaction of *dialect*sentence*.

2. Data and methods

2.1. Subjects

6 speakers were recorded for each of the 8 dialects, adding up to a total of 48 speakers (33 females, 15 males). The subjects aged between 17 and 69 confirmed to be using the dialects in question on a daily basis. None of the speakers reported speaking or hearing problems.

2.2. Material

Each speaker was asked to read in their respective dialects three sentences that were written in Standard German:

(1) *Warum verfolgt der Hund die Katze?*
'Why does the dog chase the cat?'

(2) *Der Bildschirm leuchtet stark*
'The monitor glows brightly'

(3) *Die Union von den Nonnen hat einen neuen Namen*
'The union of the nuns has a new name'

Sentence 1 on average consisted of 12 consonants and 6 vowels, sentence 2 of 14 consonants and 6 vowels, and sentence 3 of 12 consonants and 13 vowels. With respect to the syllabic make-up of the sentence material, the three sentences demonstrated the following distributions of syllable structures averaged over all dialects, see Figure 1.

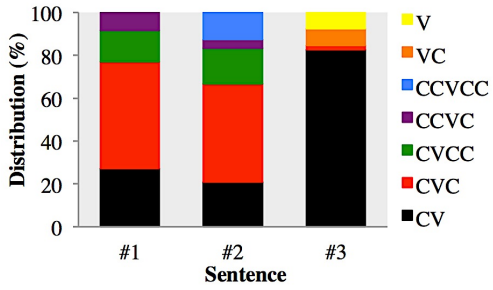


Figure 1: Syllabic make-up of sentences 1, 2, and 3.

Figure 1 reveals that sentences 1 and 2 were very different from sentence 3. Averaged over all dialects, the latter contained 83% CV syllables (black) while sentences 1 and 2 featured only 27% and 21% CV syllables. Sentences 1 and 2 contain much more CVC syllables (red) (sentence 1: 50%, sentence 2: 49%, sentence 3: 2%), however. Syllabic make-up of the sentences was further different between the dialects examined. Figure 2 shows the syllabic make-up of sentence 1 by dialect, for example.

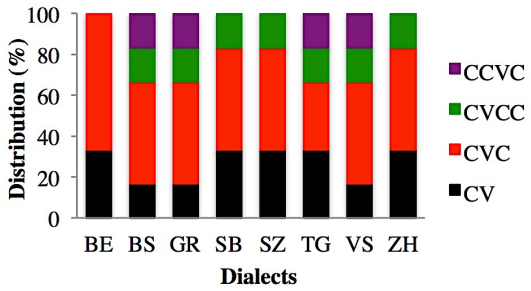


Figure 2: Syllabic make-up of sentence 1 by dialect.

Figure 2 reveals that dialects differed with regard to the relative proportions of syllable types in sentence 1. The BE speakers only exhibited CV (33%) and CVC (67%) syllables, for instance.

2.3. Procedure

Subjects were recorded in their respective locations with a Marantz PMD-671 solid-state recorder (sampling rate of 44.1 kHz and 16 bit quantization) and a Sennheiser clip-on ME2 omni lavalier microphone. Sentences were transcribed in SAMPA and labeled on the segmental level in *Praat* [7]. Consecutive vowels or consonants were merged into vocalic and consonantal intervals respectively, which provided the preferred labeling format for a subsequent application of rhythm metrics and statistical analyses. The following metrics were calculated using the *Praat* plugin 'Duration analyzer' (http://www.pholab.uzh.ch/leute/dellwo/software.html).

C:V ratio measure
The percentage over which speech is vocalic: %V [1]

Vocalic variability measures
a). The rate-normalized standard deviation of vocalic intervals: VarcoV [8]
b). The rate-non-normalized average differences between consecutive vocalic intervals: rPVI_V [9]
c). The rate-normalized average differences between consecutive vocalic intervals: nPVI_V [9]
d). The rate-non-normalized standard deviation of vocalic intervals: ΔV [1]

Consonantal variability measures:
a). The rate-normalized standard deviation of consonantal intervals: VarcoC [10]
b). The rate-non-normalized average differences between consecutive consonantal intervals: rPVI_C [9]
c). The rate-normalized average differences between consecutive consonantal intervals: nPVI_C [9]
d). The rate-non-normalized standard deviation of consonantal intervals: ΔC [1]

3. Results

3.1. Statistical analyses

All data were analyzed using R [11] and the R packages *lme4* [12] and *languageR* [13, 14]. If not indicated otherwise, we analyzed data using linear mixed effect models (LMEs). Normality was checked by visual inspection of quantile plots. *Dialect* was treated as a fixed effect, *speaker* and *sentence* as random effects. Effects were tested by model comparison between a full model, in which the factor in question was entered as either a fixed or a random effect, and a reduced model without this effect. p-Values were obtained by comparing the results from the two models using ANOVAs. For the assessment of the relative goodness of fit, we report *AIC* (Akaike Information Criterion) values that decrease with goodness of fit. Only p-values that are considered significant at the $\alpha = 0.05$ level are reported.

3.2. Interaction of *sentence* and *dialect*

Table 2 summarizes the statistics for the effects of *dialect*, *sentence*, and *dialect*sentence* by rhythm measure.

Rhythm measure	Factor	Result	Rhythm measure	Factor	Result
ΔV	dialect	AIC=-748, p<.0004*	ΔC	dialect	AIC=-502, p<.007*
	sentence	AIC=-748, p<.0001*		sentence	AIC=-502, p<.0001*
	dialect*sentence	AIC=-519, p<.003*		dialect*sentence	AIC=-519, p<.003*
VarcoV	dialect	AIC=-180, p<.0001*	VarcoC	dialect	ns.
	sentence	AIC=-180, p<.0001*		sentence	AIC=-110, p<.0001*
	dialect*sentence	AIC=-194, p<.006*		dialect*sentence	AIC=-124, p<.004*
rPVI_V	dialect	AIC=603, p<.0005*	rPVI_C	dialect	AIC=719, p<.0008*
	sentence	AIC=603, p<.0001*		sentence	AIC=719, p<.0001*
	dialect*sentence	AIC=572, p<.0001*		dialect*sentence	AIC=705, p<.01*
nPVI_V	dialect	AIC=1195, p<.0001*	nPVI_C	dialect	ns.
	sentence	AIC=1195, p<.0001*		sentence	AIC=1147, p<.0001*
	dialect*sentence	AIC=1173, p<.0002*		dialect*sentence	AIC=1132, p<.008*
%V	dialect	AIC=846, p=.024*			
	sentence	AIC=846, p<.0001*			
	dialect*sentence	ns.			

Table 2: Summary of mixed model comparisons for *dialect* and *sentence* by rhythm measure.

The comparison between the full and reduced models showed a significant difference in all rhythm metrics for *dialect* and *sentence* (except for VarcoC (cf. [6]) and nPVI_C for *dialect*), with the full model exhibiting an increased goodness of fit. Moreover, for all metrics except for %V we obtained a significant interaction of *dialect*sentence*. We take this as evidence that between-dialect variability strongly depends on the sentence used. To study simple effects of *dialect*, we conducted individual model comparisons for each of the three sentences for those rhythm metrics that showed significant interaction. 16 of the 24 model comparisons (3 sentences*8 rhythm metrics that showed interaction) revealed significant effects of *dialect* (Bonferroni adjusted for sentence, α=.017).

To visually illustrate that rhythmic differences between the dialects are contingent upon *sentence*, Figures 3 and 4 show boxplots of the dialects' VarcoV and nPVI_V by *sentence*.

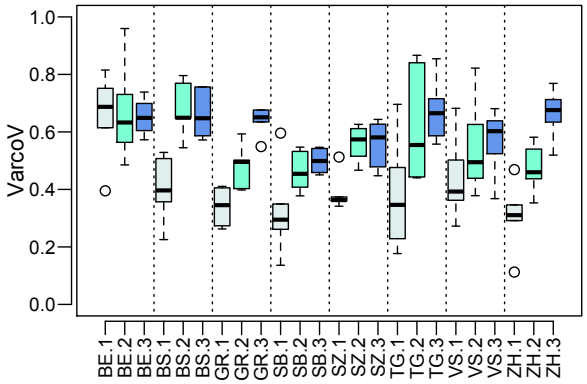


Figure 3: Boxplots of the dialects' VarcoV by sentence.

For VarcoV (see Figure 3), simple effect tests showed significant effects for *dialect* in all three sentences (Bonferroni adjusted for *sentence*, α= .017; sentence 1: *AIC*=-50, *p*<.0001, sentence 2: *AIC*=-51, *p*<.005, sentence 3: *AIC*=-90, *p*<.002). Post-hoc tests (also model comparisons, Bonferroni adjusted for *dialect* and *sentence*, α=.002) revealed, however, that dialectal differences in many cases were contingent upon the sentence being examined. In sentence 1 (grey), for example, BE speakers revealed significantly more vocalic variability (M=.66, SD=.15), than ZH speakers (M=.31, SD=.11). In sentences 2 (turquois) and 3 (blue), however, differences between these two dialects were no longer present.

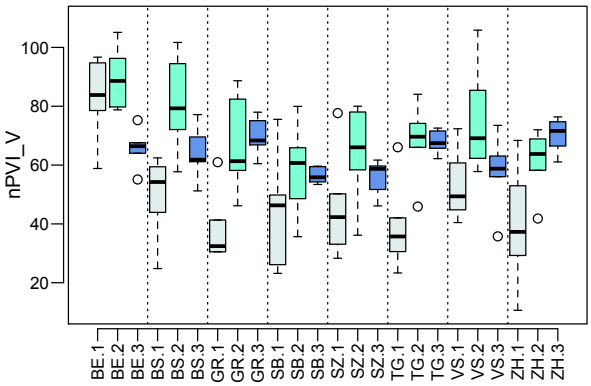


Figure 4: Boxplots of the dialects' nPVI_V by sentence.

For nPVI_V (see Figure 4), simple effect tests revealed significant effects of *dialect* in all three sentences (Bonferroni adjusted for *sentence*, α=.017; sentence 1: *AIC*=413, *p*<.0001, sentence 2: *AIC*=408, *p*<.005, sentence 3: *AIC*=339, *p*<.0005). Post-hoc tests also showed, however, that dialectal differences often depended on the sentence being examined. In sentence 1 (grey), for instance, BE speakers had significantly more vocalic variability (M=.83, SD=14), than BS speakers (M=50, SD=14). In sentences 2 (turquois) and 3 (blue), however, these differences were no longer present.

It is interesting to see that in both Figures, Figure 3 and 4, sentence 1 exhibits the lowest VarcoV and nPVI_V values for nearly all dialects, even though sentence 1 and sentence 2 demonstrate a nearly identical distribution of syllable structures (cf. Figure 1). It is conceivable that this has to do with the rhythm metrics at work in these examples: measures such as VarcoV and nPVI_V capture vowel duration variability and not syllable duration variability.

3.3. Searching for potential triggers in the material

In this section we examine possible sources in the material that cause rhythm scores between dialects to vary. The data provided are exemplary evidence and described impressionistically, largely based on visual inspection. Only VarcoV and nPVI_V are considered because these measures are controlled for speech rate, they revealed significant interaction for *dialect*sentence* (see Table 2), and they have proven to be highly discriminative for the 8 SwG dialects examined [6].

VarcoV: SB and TG

Figure 3 revealed that VarcoV differentiates SB (M=.49, SD=.04) from TG speakers (M=.67, SD=.10) in sentence 3. Post-hoc tests showed that the difference between the dialects was only significant for this sentence. In the other two sentences the dialects do not differ. If we look at a typical SB and TG realization of sentence 3, we detect two phenomena that may contribute to a lower VarcoV in the SB dialect, see Figure 5.

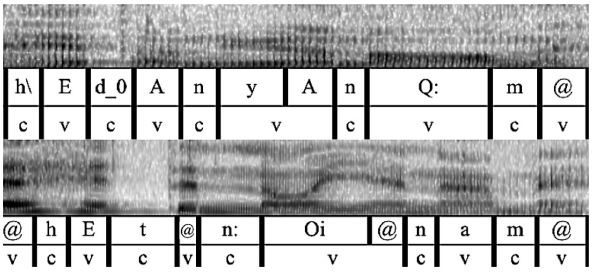


Figure 5: Typical articulation of the constituent *hat einen neuen Namen* (sentence 3) by an SB speaker (top panel) and a TG speaker (bottom panel).

Firstly, the indefinite article in 'a new name' in Figure 5 (top panel) is realized as a full vowel [ɑ] by the SB speaker, with a duration of 59 ms. The TG speaker, on the other hand, articulates a short schwa with a duration of 42 ms. Alpine dialects such as SB SwG have a tendency of realizing unstressed light syllables as full vowels. Secondly, Figure 5 reveals that the SB speaker lengthens the [ɑ] in *Name*, 'name'. SB speakers have a tendency of lengthening Middle High German short vowels [15]. This [ɑ] in the SB SwG is 155 ms long, while the TG speaker's [a] is only 112 ms long. If a sentence contains tokens that allow for full vocalic articulation and lengthening of short vowels, it is plausible that this leads to less vocalic variability and thus smaller VarcoV values for SB SwG.

nPVI_V: BE and TG

Figure 4 revealed that BE (M=83, SD=14) and TG (M=39, SD=15) differed significantly in nPVI-V in sentence 1. Post-hoc tests showed that the two dialects differ in this sentence, not, however, in sentences 2 and 3. If we look at a typical BE and TG realization of sentence 1, we find one phenomenon which may contribute to the higher nPVI_V values in the BE dialect, see Figure 6.

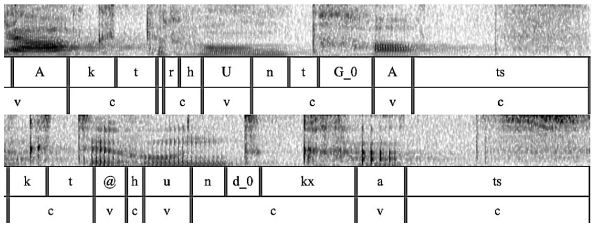


Figure 6: Typical realization of the constituent *jagt der Hund die Katze?* (sentence 1) by a BE speaker (top panel) and a TG speaker (bottom panel).

One of the triggers for a higher nPVI_V value in the BE dialect in this sentence may be the realization of an [r] in the masculine NOM. definite article *der*, as in *der Hund*, for which most other SwG dialects use the form [d̥ə]. Because of this [r], the [ə] is strongly reduced (15 ms). The TG speaker realizes *der* as [d̥ə], where the schwa is articulated more fully (76 ms). This morphological property is particularly typical of BE SwG and may have contributed to the higher proportions of CVC syllables for the BE speaker group in sentence 1 (cf. Figure 2). If a sentence contains such dialect-specific morphological properties, it is possible that this leads to different distributions in syllable structures between the dialects, and possibly more vowel reduction, which in turn may cause an increase in nPVI_V.

4. Discussion and conclusions

Results of the current study support an answer to question (1) *Does sentence material have an influence on rhythmic differences between dialects?* as a *yes*. Effects of *sentence* are not only found within one and the same language, as shown by [3, 4], but also across varieties of the same language: in the majority of the studied rhythm metrics we found significant effect of *sentence* as well as a significant interaction of *dialect*sentence*. That is, cross-dialectal rhythmical differences heavily depend on the sentence material being examined. To complicate matters, the *dialect*sentence* interaction was manifested differently depending on the applied rhythm metric (see Figures 3 and 4). For VarcoV, for example, we found that in VS and ZH SwG sentence 3 has a higher VarcoV than sentence 2. In nPVI_V, however, VS SwG sentence 3 has a lower value than sentence 2, while in ZH SwG, sentence 3 still has a higher value than sentence 2. Rhythmic variability between dialects is more complex than hitherto predicted.

These results provide further support that the selection of sentences in speech rhythm research must be given extra attention, cf. [3, 4, 16]. One can choose to work with a small data set and a special consideration for issues of dialect- or language-specific phonotactic and prosodic representativeness [3, 4], or one can work with a data set large and manifold enough to level out sentence effects. While the applied rhythm metrics have shown an interaction of *dialect*sentence*, it remains unclear to what degree these metrics reflect audible rhythmic aspects of speech [2, 3, 4, 8, 17]. It would be interesting to conduct further experiments to test whether between-sentence variability in different SwG dialects is perceptually salient.

In answer of research question (2), this study reported a number of potential triggers in the sentences material that may cause variation in the rhythm scores. By merely looking at two exemplary instances, the following dialect-specific phonological and morphological phenomena were suspected to contribute to variation in rhythm scores: SB SwG: Full vocalic articulation of unstressed light syllables, cf. [15], lengthening of short vowels [15]. BE SwG: NOM. definite article *der* is realized as [d̥ər], which increases the number of CVC syllables in this dialect (cf. Figure 2) and quite possibly results in reduced realizations of [ə].

If, hypothetically, a study uses only a small set of sentences where some sentences contain words that allow for these processes to apply and other sentences do not feature such words and consequently no such processes, dialects are likely to differ in rhythm scores in one sentence yet behave similarly in another sentence. While our preliminary findings clearly showed sentence-dependencies for between-dialect differences, these dependencies are in fact a result of dialect-specific phonological processes present in the sentence material. It is conceivable that the mentioned dialect-specific phonological processes in fact strongly contribute to between-dialect differences in rhythm scores. Such insights call for further microanalyses of the kind presented in Figures 5 & 6. In a follow-up study, one could control for features such full vocalic articulation of unstressed syllables. This would reveal the significance of such a dialect-specific phonological process for cross-dialectal comparisons in speech rhythm.

5. References

[1] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73:265-292, 1999.

[2] Prieto, P., Vanrell, M., Astruc, L., Payne, E. and Post, B., "Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish", Speech Communication, 54:681-702, 2012.

[3] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O. and Mattys, S.L., "How stable are acoustic metrics of contrastive speech rhythm?", Journal of the Acoustical Society of America, 127:1559-1569, 2010.

[4] Arvaniti, A., "The usefulness of metrics in the quantification of speech rhythm", Journal of Phonetics, 40:351-373, 2012.

[5] Leemann, A. and Zuberbühler, L., "Declarative sentence intonation patterns in 8 Swiss German Dialects", Proceedings of Interspeech, 26.-30.10.2010, Makuhari, Japan:1768-1771.

[6] Leemann, A, Dellwo, V, Kolly, M.-J. and Schmid, S., "Rhythmic variability in Swiss German dialects", Proceedings of Speech Prosody, 21.-25.5.2012, Shanghai, PRC.

[7] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer". www.praat.org, 05.15.2012.

[8] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics, 35:501-522, 2007.

[9] Grabe, E. and Low, E. L., "Durational variability in speech and the Rhythm Class Hypothesis", in C. Gussenhoven and N. Warner Eds], Laboratory Phonology 7, 515-545, Berlin/New York: Mouton de Gruyter, 2002.

[10] Dellwo, V., "Rhythm and speech rate: a variation coefficient for DeltaC", in P. Karnowski and I. Szigeti [Eds], Language and language processing: proceedings of the 38th Linguistics Colloquium, 231-241, Frankfurt: Lang, 2006.

[11] R Core Team, "R: A Language and Environment for Statistical Computing". Version 3.0.0. http://www.R-project.org, 2013.

[12] Bates, D. M. and Maechler, M., "lme4: Linear mixed-effects models using S4 classes", R package version 0.999375-32, 2009.

[13] Baayen, R. H., "Analyzing Linguistic Data: A Practical introduction to statistics using R", CUP: Cambridge, 2008.

[14] Baayen, R. H., "languageR: Data sets and functions with 'Analyzing Linguistic Data: A practical introduction to statistics using R'", R package version 0.955, 2009.

[15] Christen, H., Glaser, E. and Friedli, M., "Kleiner Sprachatlas der deutschen Schweiz", 5th ed, Stuttgart: Huber, 2013.

[16] Knight, R. A., "Assessing the temporal reliability of rhythm metrics", Journal of the International Phonetic Association, 41(3):271-281, 2011.

[17] Dellwo, V., "Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence", PhD-Thesis, Universität Bonn, 2010.

Dialectal variation at the Prosody-Syntax interface: Evidence from Catalan and Spanish interrogatives

Maria del Mar Vanrell[1], *Olga Fernández Soriano*[2]

[1] Institut für Romanische Philologie, Freie Universität Berlin, Berlin, Germany
[2] Departamento de Filología Española, Universidad Autónoma de Madrid, Madrid, Spain
mariadelmar.vanrell@fu-berlin.de, olga.fernandez@uam.es

Abstract

In this study we investigate how prosody interacts with word order in the expression of interrogativity in different varieties of two Ibero-Romance languages, Catalan and Spanish. We analyze a corpus obtained by means of the Discourse Completion Task Methodology. The collected data were prosodically and syntactically annotated and show that the absence of syntactic marking (wh-word, subject-verb inversion or subject dislocation) for questions tends to correspond to a more salient intonational marking. Thus, wh-questions favor general falling intonational patterns. By contrast, yes-no questions can be classified depending on the nuclear tone (with preference for low tones in Catalan and high tones in Spanish) and final tone (low for language varieties with subject inversion or dislocation, but optionally high for those that do not present syntactic marking in a mandatory way).

Index Terms: prosody, word order, interrogativity, Catalan, Spanish, dialectal variation.

1. Introduction

The interface between prosody and word order in Ibero-Romance has not been consistently studied. A considerable amount of research has been devoted to languages like Spanish, but other languages such as Catalan are much less known. In addition, most of the work concentrates on declarative modality, particularly on the expression of focus ([1], [2], [3]), as well as on the location of the main prominence (nuclear accent) and the theoretical implications of the location of this prominence. However, somewhat less attention has been devoted to how focus structure can influence pitch accent choice.

Regarding previous studies on Catalan and Spanish questions, although we find exceptions such as the work in [4] and [5] for Catalan and Spanish respectively, there is often a stark division between those studies that emphasize the syntactic perspective (word order in the marking of interrogativity: [6] for Catalan) and those that focus on the prosodic perspective (pitch accent choice in different question types: [7], [8] for Catalan; [9], [10] for Spanish). Moreover, syntactic studies tend to disregard dialectal variation, dealing instead primarily with standard varieties or specific characteristics of particular varieties (e.g., the absence of subject inversion in Caribbean Spanish). It is also necessary to note that, as pointed out by [2], syntactic works and phonological studies tend to use different methodological approaches. Whereas syntactic works make use of introspection and grammaticality judgments, phonological studies have a tendency to use experimental methods. This makes the results from the two linguistic fields very difficult to compare.

This paper makes an attempt to encompass all those aspects and reconcile the two perspectives (prosodic and syntactic), while dealing with dialectal variation and applying a uniform, controlled methodology. We investigate the interaction between nuclear configuration types and word order in the marking of interrogativity in different varieties of Catalan and Spanish.

2. Methodology

2.1. Participants

The participants in our production experiment were 4 men and 10 women aged between 22 and 45 from the following locales of the two languages under study: a) Central Catalan (CCat): 2 female speakers, province of Barcelona; Balearic Catalan (BCat): 1 female speaker and 1 male speaker, island of Majorca; Valencian Catalan (VCat): 1 female speaker and 1 male speaker, province of Alacant and Valencia respectively; and b) Castilian Spanish (CasSpa): 2 female speakers, Madrid; Spanish as spoken in the Basque Country: two female L1-Spanish speakers, province of Biscay (BCSpaL1Spa) and 2 female L1-Basque speakers (BCSpaL1Bas), province of Biscay too; Canarian Spanish (CanSpa): 2 male speakers, island of Gran Canaria.

2.2. Materials

The corpus analyzed in this paper was obtained by means of the Discourse Completion Task methodology or DCT ([11], [12], [13]). It is an inductive method in which the researcher presents the subject with a series of situations and then asks him or her to respond accordingly. The full survey is made up of 130 situations that allowed us to obtain a wide range of interrogative contours controlling for the type of verb, the type of subject and the degree of certainty about the likelihood that the speaker will get a "yes" answer to his/her utterance. We elicited in this fashion a total of 1820 contours (130 contours x 14 speakers). In this paper we will present the results for 779 contours (224 Catalan y/n questions, 118 Catalan wh-questions; 292 Spanish y/n questions, 145 Spanish wh-questions). Indirect questions and y/n questions other than information-seeking y/n questions were not analyzed in this paper.

2.3. Procedure

The descriptions of the prompt situations provided in the DCT were read aloud to the participants. After each description, the participant was asked to respond appropriately to the situations as spontaneously as possible. Speakers were recorded on a Zoom H4n digital audio recorder using an AKG C520 condenser microphone. The following example illustrates a situation used to elicit a y/n question in Spanish with nominal subject: 'You have no idea whether Juan bought the car or not.

Ask your friend about this.' Intended target response: 'Did Juan buy the car?'.

2.4. Analysis

Data were annotated in Praat ([14]) for the following fields: (i) orthographic transcription, (ii) position of the subject (postverbal or preverbal, right or left dislocated, dropped), (iii) additional lexical markers such as *que* or *oi?* for Catalan and (iv) prosodic transcription using the latest version of the Cat_ToBI and Spa_ToBI systems ([8] for Catalan and [10] for Spanish). The annotations were compiled automatically in .txt format through a Praat script and then transferred to a SPSS file for purposes of subsequent statistical exploration.

3. Results

3.1. Word order results

Table 1 reports the frequency of observation (e.g., number of occurrences) of the variable SUBJECT EXPRESSION AND POSITION (dropped, preverbal, postverbal and dislocated) for nominal and pronominal subjects in y/n questions and different language varieties. A clear-cut division emerges between language varieties that can present postverbal subjects and those that cannot. Thus, VCat and Spanish in general can have a postverbal subject, see (1), whereas Eastern Catalan (BCat and CCat) tend to dislocate the subject, as in (2). Interestingly, there are two varieties, VCat and CanSpa,[1] which—although they can resort to subject inversion as a question marker—prefer to have the verb in preverbal position, as in (3). Our data show that the probability of a subject appearing in preverbal position increase in the case of 1st and 2nd person pronominal subjects. That is, sentences like (4) are possible, whereas sentences like (5) are more marginal though still possible. Our results also show that in Spanish the 2nd person formal form *usted*, which agrees in 3rd person with the verb, appears more frequently in postverbal position, as in (6), than 1st and non-formal 2nd personal pronominal subjects, which tend to be dropped, due to the fact that they usually cannot be distinctive in these contexts (hence the contrast with full NPs). The analogous Catalan form *vostè* tends to be dislocated but can also appear in preverbal position.

(1) *¿Nació el hijo de la vecina en Madrid?* (CasSpa)
be born.PAST.3SG the son of the neighbor in Madrid

(2) *Treballa fins tard, el fill de la veïna?* (CCat)
work.PRES.3SG until late the son of the neighbor

(3) *¿La mujer de Juan es francesa?* (CanSpa)
the wife of Juan is French

(4) *Tu vares comprar un cotxe?* (BCat)
you buy.PAST.3SG a car

(5) *Ets francès, tu?* (CCat)
be.2sg French, you

(6) *¿Nació usted en Madrid?* (BCSpaL1Spa)
be born.PAST2SG2SG.FORMAL in Madrid

Table 1. *Frequency of observation of the variable SUBJECT EXPRESSION AND POSITION for nominal and pronominal subjects in y/n questions across different language varieties. The highest numbers or (in some cases) the second highest numbers after subject drop in pronominal subjects are indicated in bold.*

Language varieties	Nominal subject			Pronominal subject			
	Pr	Po	Di	Dr	Pr	Po	Di
BCat	1	0	**17**	5	7	0	**14**
CCat	0	0	**18**	5	0	0	**14**
VCat	**14**	3	0	12	**7**	1	0
CanSpa	**16**	0	0	7	2	**7**	0
CasSpa	6	**13**	0	11	2	**5**	0
BCSpaL1Spa	**14**	6	0	8	**6**	4	0
BCSpaL1Bas	3	**17**	0	9	3	**5**	0

Table 2 shows the frequency of observation of the variable SUBJECT EXPRESSION AND POSITION (dropped, preverbal, postverbal and dislocated) for nominal and pronominal subjects in wh-questions and different language varieties. Again BCat and CCat tend to dislocate the subject, as in (7), whereas VCat and Spanish clearly do not, as in (8). Like in y/n questions, pronominal subjects tend to be more commonly produced in preverbal position than nominal subjects. Again, it is of interest to single out the behavior of the form *usted*. As was seen for y/n questions, this pronominal form does not behave like other Spanish personal pronouns, nor does it behave like a full DP (see [15]).

(7) *I quan fa feina, n'Aina?*(BCat)
and when make.PRES.3SG work, PERS.ART-Aina

(8) *Què volia el fill de la veïna?* (VCat)
what want.past.3sg the son of the neighbor
¿Dónde nació Ana? (CSpa)
where be.born.PAST.3SG Ana

The results of a two-way ANOVA test with SUBJECT EXPRESSION AND POSITION as the dependent variable and SUBJECT TYPE, LANGUAGE VARIETY and QUESTION TYPE as independent variables revealed statistically significant effects of SUBJECT TYPE ($F(1,751)=103.320$, $p<.001$), LANGUAGE VARIETY ($F(6,751)=18.581$, $p<.001$) and QUESTION TYPE ($F(1,751)=4.492$, $p<.05$). Post-hoc analyses confirmed that Eastern Catalan (BCat and CCat) differed significantly from the other language varieties (Tukey $p<.001$).

Table 2. *Frequency of observation of the variable SUBJECT EXPRESSION AND POSITION for nominal and pronominal subjects in wh-questions across different language varieties. The highest numbers or (in some cases) the second highest numbers after subject drop in pronominal subjects are indicated in bold.*

Language varieties	Nominal subject			Pronominal subject			
	Pr	Po	Di	Dr	Pr	Po	Di
BCat	3	4	**34**	18	10	0	**11**
CCat	0	0	**42**	2	2	0	**25**
VCat	6	**36**	0	18	4	**5**	0
CanSpa	6	**37**	1	13	7	**8**	1
CasSpa	1	**41**	0	10	4	**14**	1
BCSpaL1Spa	0	**43**	1	15	**7**	**7**	1
BCSpaL1Bas	0	**42**	0	9	0	**22**	0

3.2. Intonation results

Table 3 shows the frequency of occurrence (as counts) of the variable NUCLEAR PATTERN for two question types (y/n questions and wh-questions) and the different language

[1] Preverbal nominal subjects are also common in BCSpaL1Spa. However, due to space limitations, this issue cannot be discussed in any greater depth.

varieties under study. In Table 3 only the most common patterns are reported. Regarding question intonation, first we would first like to highlight the crucial role of prosody in marking whether the subject is dislocated or not. Each dislocated element constitutes a tonal unit which is independent of the core sentence. Hence, an intonational contour made of a core sentence and two dislocated elements (e.g., *Vindrà, la Maria, demà?* Will come, Maria, tomorrow) is produced with three different tonal units. Most instances of dislocated subjects in our data are right dislocations. Right dislocated elements reproduce the intonation pattern of the core sentence but with some variation depending on whether the intonational contour is rising or falling (see [16]).

Table 3.*Frequency of observation of the variable* NUCLEAR PATTERN *for y/n and wh-questions across different language varieties.*

Language varieties	Y/n questions		Wh-questions	
	NC	Freq	NC	Freq
BCat	¡H+L* L%	35/43	H+L* H%	37/80
CCat	L* H%	29/37	H* L%	68/71
VCat	L* H%	37/38	H+L* L%	73/73
CanSpa	L+¡H* L%	14/32	H* L%	33/73
	L* H%	14/32		
CasSpa	L+¡H* L%[2]	17/38	L* H%	22/71
			H* L%	18/71
BCSpaL1Spa	L+¡H* HL%	18/37	L* L%	48/74
BCSpaL1Bas	L+¡H* HL%	17/38	L* L%	65/74

The intonation of yes-no questions in our data can be grouped into two different patterns according to the pitch tonal event associated with the nuclear syllable: falling/low or high. Whereas Catalan follows the first pattern, nuclear accents in Spanish yes-no questions tend to be high. As for the boundary tones, they can also be low (BCat and Spanish) or high (CCat, VCat and CanSpa). These are not absolute tendencies, since for instance CanSpa can also present a rising pattern, that is, a low nuclear accent followed by a high boundary tone (as can be seen in Table 3). These general tendencies regarding the intonational patterns are consistent with the data presented in [7], [8] and [9], [10] for Catalan and Spanish respectively. When putting the results together, one could hypothesize that the tonal variation found in the nuclear syllable reflects language-specific differences (Catalan vs. Spanish), whereas the tonal variation located at the final stretch of the contour is related to the syntactic marking of the interrogative modality. In other words, language varieties that use syntactic means such as subject inversion or dislocation to mark interrogativity tend to use low tones,[3] but language varieties which do not resort to syntactic markers in a compulsory way (such as VCat or CanSpa) have high final boundary tones available in their intonational grammars. Figure 1 shows an instance of the y/n question *És francesa, la dona del Joan?* (Is French, the wife of.PERS.ART. Joan) produced by a CCat speaker. The utterance presents a L* nuclear accent associated with the accented syllable of *Joan* followed by a H%. Figure 2 illustrates the y/n question *¿Trabaja Juan hasta tarde?* (work.pres.3SG Juan until late) produced by a speaker of CanSpa. The intonational pattern is characterized by a nuclear L+¡H* accent associated with the accented syllable of *tarde* and a low boundary tone (L%).

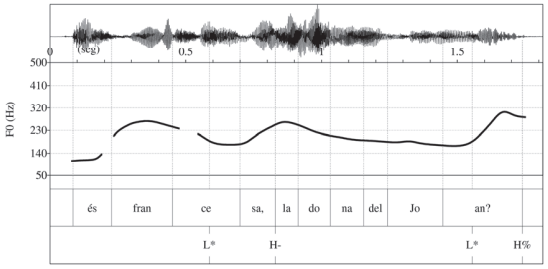


Figure 1: *Waveform and F0 contour of the y/n question with right dislocation of the subject És francesa, la dona del Joan? 'Is she French, Joan's wife?' produced by a speaker of CCat.*

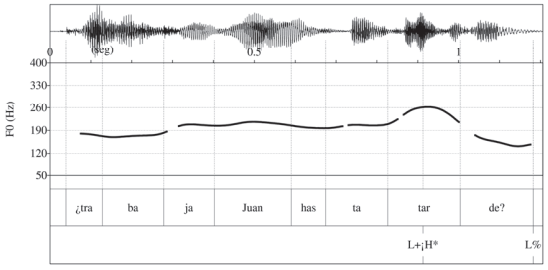


Figure 2: *Waveform and F0 contour of the y/n question ¿Trabaja Juan hasta tarde? 'Does Juan work until late?' produced by a speaker of CasSpa.*

As in the case of y/n questions, wh-questions can be grouped into two categories depending on the tone associated with the nuclear syllable: falling/low or high. Falling/low patterns are common in all the varieties except CCat, CanSpa and CasSpa, which produce high nuclear pitch accents. The final boundary tones tend to be low, although BCat and CasSpa can also present a final rising boundary tone (H%). This H+L* H% pattern in wh-questions has often been regarded as more polite compared to the falling pattern (H+L* L%) (see [17], [5], [7]). The more frequent presence of falling nuclear configurations in wh-questions could be explained by the fact that these questions are formally marked by a preposed wh-word and therefore they do not need to be marked intonationally. Figure 3 displays an example of the wh-question *Quan treballa el fill de la veïna?* 'When does the son of the neighbor work?' produced by a speaker of VCat and characterized by the H+L* L% nuclear configuration. Figure 4 shows the wh-question *¿Dónde nació Ana?* '¿Where was Ana born?' produced by a CanS speaker. The intonational contour includes a H* nuclear accent aligned with the accented syllable of *Ana* and a low final boundary tone.

An inter-transcriber reliability test of the ToBI transcription was conducted with a 10% sample of the data. The subset of data was selected on the basis that all the language varieties and question types were uniformly represented. Four transcribers labeled the subset using the Cat_ToBI and Sp_ToBI systems. Since there were more than two transcribers, the Fleiss' kappa statistic was used. We obtained a moderate agreement for the choice of pitch accents

[2] It should be noted that this result does not conform to the predictions made by traditional studies such as [17], [18]. However, our results agree with those of [19] in which y/n questions uttered in spontaneous speech are analyzed. A possible explanation for this inconsistency could be that L* H% contours are more restricted to formal speech situations ([20], [19]).

[3] As pointed out by a reviewer, CCat would not follow our generalization, since although y/n questions tend to be marked by subject inversion, 78% of the contours were pronounced by a H% boundary tone. The case of CCat deserves a more fine grained study since the alternation between rising and falling patterns could be related to subdialectal variation (see [7], Map 8.7.).

(0.62) and an almost perfect agreement for boundary tones (0.85) ([21]).

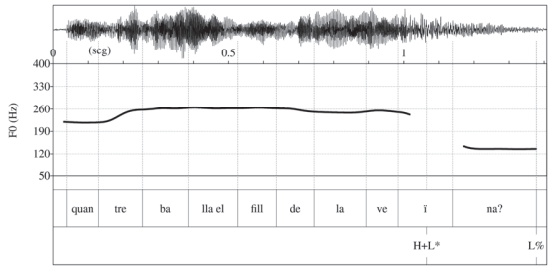


Figure 3: *Waveform and F0 contour of the wh-question Quan treballa el fill de la veïna? 'When does the son of the neighbor work?' produced by a speaker of VCat.*

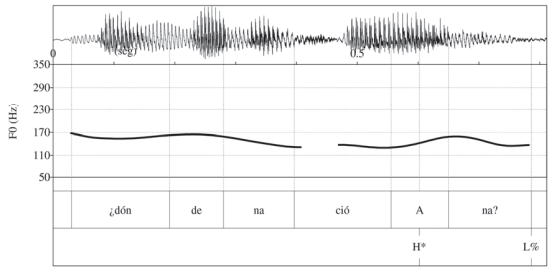


Figure 4: *Waveform and F0 contour of the wh-question ¿Dónde nació Ana? 'Where was Ana born?' produced by a speaker of CanSpa.*

4. Discussion and conclusions

This study represents a first step in exploring how prosody interacts with word order in the expression of interrogativity in Ibero-Romance. To this end, a production study was designed using the DCT methodology to elicit different question types. In addition, we controlled for verb and subject type, and degree of certainty about the response—here only subject type was included in the analysis. The data were annotated prosodically and syntactically with Praat ([14]). After performing a quantitative analysis using the dependent variable SUBJECT EXPRESSION AND POSITION, we conclude that there are three factors that play an important role in the expression of interrogativity by means of word order: question type (y/n question and wh-question), language variety and subject type (nominal, pronominal or *usted*). One of the most crucial differences obtained is between languages that can invert the subject to mark questions (VCat and Spanish) and language varieties that cannot (Eastern Catalan, i.e., BCat and CCat). The latter resorts to subject right/left dislocation. The exception to this general behavior concerns pronominal subjects (1st and 2nd person singular), which tend to occur in preverbal position. The formal pronoun *usted* behaves like a nominal subject in this respect. Our explanation for this particular behavior is based on the lack of a match between the features of the form *usted* (a second person, formal) and the features in INFL, inflection (third person). Our data advocates for the existence of a particular position for this form inside IP (inflectional phrase). VCat and CanSpa show a preference for preverbal subjects in yes-no questions including nominal subjects. These results for VCat and CanSpa should be related to the results obtained in a recent study we carried out (see [16]) which focused on the role of word order and prosody in the expression of not solely interrogativity but also different focus structures. The results revealed that especially CanSpa but also VCat, compared to other Catalan varieties such as BCat and CCat, are [+ plastic] (see [22], [23]) language varieties in the marking of focus, that is, the interaction between focus and prominence is achieved by means of prominence shift rather than changes in word order. This comes as no surprise since Caribbean varieties have been reported to have many commonalities with CanSpa (e.g., the use of the circumflex intonational pattern, possibility of lacking subject inversion in wh-questions, deletion of final *-s*) and they are known to be especially [+plastic] ([24] for Puerto Rican Spanish) and to exhibit a particular preference for SVO order in questions ([22] for Puerto Rican Spanish).

With respect to intonation, we interpret the absence of syntactic marking (wh-word, subject-verb inversion or subject dislocation) for questions as corresponding to a more salient (in terms of pitch height) intonational marking. Hence, wh-questions are mainly characterized by falling patterns, even though the nuclear accent can be low/falling or high. As for yes-no questions, they can be classified by the nuclear tone, low/falling for Catalan and high for Spanish, but also by the final boundary tone. Regarding the final boundary tone, the general tendency points to low tones for language varieties with subject inversion or dislocation, and to optionally high tones for those varieties that do not present syntactic marking in a mandatory way.

As some researchers have suggested, the lack of subject inversion in direct questions may be the syntactic manifestation of a larger shift towards languages becoming more rigid with respect to word order ([26], [27], [22]) with French being the most drastic case among Romance languages. Finally, we would like to highlight that together with syntactic characteristics such as the absence of subject drop, the presence of subject pronouns with infinitival forms (i.e., *para yo hacer esto* 'for me to do this', see [25], 26]) or the lack of subject inversion in questions, the use of a more salient intonational marking can offer insights into the dynamics of language change.

5. Acknowledgments

A preliminary version of this article was presented at the Forschungskolloquium der Romanistischen Sprachwissenschaft. We are grateful to the participants at that meeting for their helpful comments and suggestions. We thank Joan Borràs-Comes, Paolo Roseano and Rafèu Sichel-Bazin for participating unselfishly in the inter-transcriber reliability test of the ToBI transcription and Meghan E. Armstrong, Nick Henriksen and two anonymous reviewers for their comments and advice on the first written version of this paper. Thanks also go to the participants in our experiment as well as the people that helped us to contact potential participants, namely Gotzon Aurrekoetxea, Mercedes Cabrera, Verònica Crespo-Sendra, Irene de la Cruz, Gorka Elordieta, Leire Gandarias, Miriam Rodríguez and Paco Vizcaíno. The research assistant Anne-Kathrin Knecht deserves a special mention for her help in the preparation of this article. This research has been funded by the project FFI2011–23829/FILO awarded by the Spanish Ministry of Economy and Competitiveness.

6. References

[1] Zubizarreta, M. L., Prosody, focus, and word order, MIT Press, 1998.

[2] Gabriel, C., "On focus, prosody, and word order in Argentinean Spanish: a minimalist OT account", Revista Virtual de Estudos da Linguagem (ReVEL), Special issue 4 "Optimality theoretic Syntax", 183-222, 2010.

[3] Domínguez, L., "Analyzing unambiguous narrow focus in Catalan", in T. Ionin, H. Ko and Nevins, A. [Ed], The Proceedings of the Second Humit Conference, MIT Working Papers in Linguistics 43:17-34, 2002.

[4] Prieto, P. and Rigau, G., "The syntax-prosody interface: Catalan interrogative sentences headed by *que*", Journal of Portuguese Linguistics, 6(2):29-59, 2007.

[5] Escandell-Vidal, V., "Los enunciados interrogativos. Aspectos semánticos y pragmáticos", in I. Bosque and Demonte, V. [Ed], Gramática Descriptiva de la Lengua Española, 3929-3991, Espasa Calpe, 1999.

[6] Rigau, G., "La posición del sujeto en catalán", in J.M. Brañas, Ch. Nak-Won and Chan-Yong, Sh. [Ed.], Conexiones de la Sociedad Coreana y la Española. I Congreso Internacional de Coreanología, 71-82, UAB/Universitat Nacional de Chonbuk, Publicaciones Digitales S.A., 2002.

[7] Prieto, P., and Cabré, P. [Ed.], L'entonació dels dialectes catalans, Publicacions de l'Abadia de Montserrat, 2013.

[8] Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R. and Vanrell, M.M., "Intonational phonology of Catalan and its dialectal varieties", in S. Frota and Prieto, P. [Ed], Intonational variation in Romance, OUP, in press, to appear in 2014.

[9] Prieto, P. and Roseano, P. [Ed.], Transcription of Intonation of the Spanish Language, Lincom Europa, 2010.

[10] Hualde, J. I., and Prieto, P., "Intonational variation in Spanish: European and American varieties", in S. Frota and Prieto, P. [Ed], Intonational variation in Romance, Oxford: OUP, in press, to appear in 2014.

[11] Blum-Kulka, S., House, J. and Kasper, G., "Investigating cross-cultural pragmatics: An introductory overview", in S. Blum-Kulka, J. House and Kasper, G. [Ed], Cross-cultural pragmatics: Requests and apologies, 13-14, Norwood, NJ: Ablex, 1989.

[12] Billmyer, K. and Varghese, M., "Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests", Applied Linguistics, 21(4): 517-552, 2000.

[13] Félix-Brasdefer, C., "Data collection methods in speech act performance: DCTs, role plays, and verbal reports", in A. Martínez-Flor and Usó-Juan, E. [Ed.], Speech act performance: Theoretical, empirical, and methodological issues, 41-56, John Benjamins, 2010.

[14] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer [Computer program], Version 5.3.55. Online: http://www.praat.org, accessed on 2 Sep 2013.

[15] Fernández Soriano, O., "El pronombre personal. Formas y distribución. Pronombres átonos y tónicos", in I. Bosque and Demonte, V. [Ed], Gramática descriptiva de la lengua española, 1209-1275, Madrid: Espasa Calpe, 1999.

[16] Vanrell, M.M. and Fernández Soriano, O., "Variation at the interfaces in Ibero-Romance. Catalan and Spanish prosody and word order", Catalan Journal of Linguistics, 12: 253-282, 2013.

[17] Navarro-Tomás, T., Manual de entonación española, Hispanic Institute in the United States, 1944.

[18] Quilis, A., Fonética acústica de la lengua española, Gredos, 1981.

[19] Henriksen, N.C., Armstrong, M.E., and García-Amaya, L.J., "The intonational meaning of polar questions in Manchego Spanish spontaneous speech", in M.E. Armstrong, N.C. Henriksen, and Vanrell, M.M. [Ed.], Interdisciplinary approaches to intonational grammar in Ibero-Romance, Benjamins, accepted.

[20] Henriksen, N.C., "Style, prosodic variation, and the social meaning of intonation", Journal of the International Phonetic Association, 43: 153-193, 2013.

[21] Landis, J.R. and Koch, G.G., "The measurement of observer agreement for categorical data", Biometrics 33, 33(1): 159-174, 1977.

[22] Brown, E. L. and Rivas, J., "Subject-verb word order in Spanish interrogatives. A quantitative analysis of Puerto Rican Spanish", Spanish in Context, 8(1): 23-49, 2011.

[23] Vallduví, E., "The role of plasticity in the association of focus and prominence", Proceedings of the Eastern States Conference on Linguistics (ESCOL), 7: 295-306, 1991.

[24] Armstrong, M. "Puerto Rican Spanish intonation", in P. Prieto and Roseano, P. [Ed], Transcription of intonation of the Spanish language, Lincom, 2010.

[25] Vallduví, E., and Zacharski, R. "Accenting phenomena, association with focus, and the recursiveness of focus-ground", in P. Dekker and Stokhof, M. [Ed.], Proceedings of the Ninth Amsterdam Colloquium, 683-702, Institute for Logic, Language and Computation, 1994.

[26] Morales, A., "Hacia un universal sintáctico del español del Caribe: el orden SVO", Anuario de Lingüística Hispánica, 5: 139-152, 1989.

[27] Toribio, A. J., "Setting parametric limits on dialectal variation in Spanish", Lingua, 10: 315-341, 2000.

Prosodic Phrasing of SVO Sentences in French

Mathieu Avanzi[1], *George Christodoulides*[2], *Elisabeth Delais-Roussarie*[1]

[1]Laboratoire de Linguistique Formelle, UMR 7110, University Paris Diderot, France
[2]Institut Langage & Communication, Centre VALIBEL, UCLouvain, Belgium

mathieu.avanzi@gmail.com, george@mycontent.gr, elisabeth.roussarie@wanadoo.fr

Abstract

In the literature on prosody/syntax interface, syntactic information is usually considered as playing an important role in deriving the prosodic phrasing of an utterance. NP subjects, for instance, have often been claimed to phrase independently from the VP. It has nevertheless been shown that metrical factors could have an impact on phrasing, and that NPs could be phrased in the same prosodic phrase as the VP, or that the verb could be phrased with the subject. Several methods were used to measure metrical weight: number of syllables, of prosodic words, syntactic branchingness, etc. In order to determine which factors are more important, and how the all interact, we evaluate the weight that different metrical predictors have on prosodic phrasing. This is done by analyzing the phrasing of SVO structures in 200 sentences extracted from various French corpora. From the observation of the data that were semi-automatically annotated, it appears that subjects can be phrased independently or in the same prosodic phrase as the VP, and that objects are rarely isolated from the verb. The analysis reveals interesting results regarding the effect of articulation rate and number of syllables, whereas syntactic-branchingness didn't show any effect.

Index Terms: Prosody, metrical structure, phrasing, interface prosody-syntax, articulation rate

1. Introduction

Several theoretical and experimental studies have shown that multiple factors constrain prosodic phrasing, among which we may mention metrical constraints. Nevertheless, syntactic information are often considered as playing the most fundamental role, regardless of the approach used to account for the mapping between the prosodic structure and the syntactic one. In [1], for instance, it is assumed that edges of prosodic pharses should coincide with edges of syntactic phrase, whereas [2] insists on head-complement relations in determining the prosodic phrases (see for a review of the various approaches [3] and [4]).

Whatever approach is taken into account (end-based or relational), one would predict for French SVO sentences that NP subjects form an independent prosodic phrase, separated from the rest of the utterance, and more specifically the VP (see, for instance, [5] and [6]). On this basis, the sentence in (1) would be phrased in two prosodic groups, as shown in (2):

(1) Le gardien a vu le beau chien de ma voisine

(2) [Le gardien] [a vu le beau chien de ma voisine]

Nevertheless, it has been shown that mapping rules were not sufficient to predict phrasing. Factors such as length of the syntactic phrases could play a crucial role in the prosodic parsing of an utterance. As for SVO structure in French, [7] and [8] showed that NP subjects could be phrased in the same prosodic phrase as the VP when they are short (i.e. when they contain a small number of syllables), and that objects could be separated from the verbal head V to obtain well-balanced phrases. As a consequence the parsing in (3) is completely acceptable for the sentence (1), and maybe even better than the one proposed in (2):

(3) [Le gardien a vu] [le beau chien de ma voisine]

In the literature, the length of the constituents have been evaluated by several means: syntactic branchingness or complexity ([2]), number of prosodic units (generally number of prosodic words, as in [9]) or number of syllables ([10] and [8] among others). The way certain of these factors interact have been analyzed in details in some experimental studies (see [11] and [12] among others). In [12], where four different Romance languages were compared, it has been shown that SVO structures do phrase differently in the four Romance languages under investigation. Nevertheless, the (S)(VO) pattern was the most common for Catalan, Spanish and Italian, while (SVO) is the unmarked phrasing for European Portuguese. In addition, the authors showed that syntactic branchingness, number of prosodic words, and number of syllables had a different impact on phrasing decisions, and interacted in a different fashion across the four languages, leading for example to the production of patterns such as (SV)(O) in Catalan under certain length condition, a pattern which never was found for European Portuguese. A systematic study of the impact of these different factors has never been achieved for French. The goal of this paper is thus to investigate how all these factors interact in French. Due to a limited number of utterances and to the important number of parameters at play, the results presented here are only preliminary.

2. Methods

2.1. Data and participants

The analysis presented here is based on corpus data. We explored a total of 90 minutes of speech recorded by 18 speakers in four different speaking conditions. Among these 90 minutes, 20 minutes consist in the reading of fairy tales by professionals, 20 minutes in political speeches, and the remaining 50 minutes consist in different types of spoken data produced by 8 Parisian speakers recorded within the PFC project [13]. Following the data collection protocol used for the project, each speaker was asked to read carefully at a normal speech rate a journalistic text including 22 sentences (398 words), and then to converse freely in pair for 20 minutes. Among these data, the reading of the text and 3.30 minutes of conversation were taken into consideration for each of the 8 speakers in the present studies. Table 1 gives an overview of the number of speakers and of the duration of the samples.

Table 1: *Composition of the data set*

Speaking Style	Duration (min.)	Nb. of speakers
Fairy tales	19'28	6
Political speeches	21'08	4
Reading (PFC)	18'53	8
Spontaneous (PFC)	30'10	

We controlled for the sex of the speakers, selecting an equal number of male and female for each sub-corpus; however age was not a controlled variable. All participants speak a standard variety of French.

2.2. Labeling of the data and prosodic structure

The recordings were transcribed within Praat [14], and aligned in phones, syllables and words with EasyAlign [15]. All alignments were manually checked and corrected by hand by one of the authors. Then, the orthographic transcription of the data was annotated in PoS with the Dismo software [16]. This allows assigning a phonological status to the different words with respect to their ability to be accented (see among others [17] and [18]), and then segmenting the data in Phonological Words (henceforth PW).

The strength of the prosodic boundaries occurring at the end of prosodic words depends on their phonetic realization. When the last metrical syllable of a prosodic word is associated with a pitch movement, is lengthened, and even followed by a pause, it is considered as corresponding to a major prosodic boundary that may correspond to a major phrase or an IP boundary. In order to analyze intonational phrasing without going into details as for the distinction between major phrase or intermediate phrase on the one hand, and intonational phrase on the second (see [19] among others), the two levels have been grouped together and referred to as *major prosodic phrases* (MaP). The strength of the boundary occurring at the end of each prosodic word has been automatically detected with the Analor tool [20]. On the basis of four automatically measured acoustic parameters (relative syllabic duration, relative f0 average, slope contour amplitude and presence of an adjacent silent pause), the software estimates a degree of strength for the last syllable of each PW on a scale from 0 to 10 (from the least to the most prominent). The calculations rely on two fundamental principles. The first is a quantity principle: the greater the number of acoustic parameters involved in the identification of a prominence and the distance from predetermined thresholds, the stronger the prominence is perceived. The second is a compensation principle, which stipulates that if one of the classic parameters involved in the perception of prominence in French presents a low value and another presents a high value, there will be the same feeling of prominence as if the two parameters involved both presented a medium score. We considered that the right edge of a PW coincides with a major prosodic phrase boundary when the degree of strength associated to the last syllable of the PW reached a score of 4/10.

2.3. Extraction and prosodic encoding of the SVO structure

Using a manual annotation, we extracted from the entire corpus all sentences having a SVO pattern, where S, V and O were not separated by any other kind of linguistic material (appositive clauses, adverbs and other parentheticals), and where S and O consisted of a SN. In addition we restricted ourself to main clauses, leaving aside embedded SVO structure. In total, 198 sentences were extracted. Table 2 gives the distribution among the sub-corpus of the extracted sentences.

Table 2: *Nb of SVO sentences extracted from each sub-corpus*

Speaking Style	Nb. of sentences
Fairy tales	85
Political	28
Reading	67
Spontaneous	29

On the basis of the annotation described in section 2.2, it was possible to assign to each syntactic units of the SVO structure prosodic labels/ features(including F0 patterns, number of syllables, and number of PWs). Due to the uncontrolled character of the data (corpus data), the length of the various units vary greatly, as shown in **Erreur ! Source du renvoi introuvable.**.

Table 3: *Range and mean (standard deviation) nb. of syllables and of PWs for each syntactic constituent*

Synt. Const.	Nb. of Syll.		Nb. of PW	
S	Min.-Max Mean (SD)	1-17 4.8 (2.6)	Min.-Max Mean (SD)	1-4 1.7 (0.9)
V	Min.-Max Mean (SD)	1-8 3.1 (1.6)	Min.-Max Mean (SD)	1-3 1.1 (0.5)
O	Min.-Max. Mean (SD)	1-15 4.7 (2.7)	Min.-Max. Mean (SD)	1-4 1.7 (0.8)

In addition, the syntactic complexity of each S and O (simple nominal head or nominal head with branching complements) was manually coded. The articulation rate of the entire sentence was also automatically calculated by dividing the duration of the entire sentence (and by excluding the time of silent pauses) by the total number of syllables (it is therefore expressed in ms/syll, (see [21]).

3. Results

The analysis of the data, and more specifically of the boundary strength associated at the end of S, V, and O respectively, showed that S and O could be phrased independently from V, or phrased in the same MaP than V. It thus leads to four possible combinations, which were all obtained in our data. Yet, as table 4 shows, the four combinations do not occur with the same frequency.

Table 4: *Distribution of the phrasing obtained for the SVO structures*

Pattern	Count	%
(S)(V)(O)	16	8.1
(S)(VO)	92	46.5
(SV)(O)	4	2
(SVO)	86	43.3

The (S)(VO) pattern is the most frequently observed, followed by the (SVO) pattern. Taken together, these two patterns represent approximately 90% of the data. Cases where both S and O are isolated from V by a major prosodic phrase boundary are quiet rare, and cases where S and V are phrased in a single IP while O is in a distinct IP are even rarer (these two patterns taken together account for 10% of all SVO occurrences)

Due to the small amount of cases for the two last patterns, and in order to simplify the analysis and presentation of the results, we separated the data according to the way S and O were phrased in regards of V. Essentially, we compared the group of sentences where S was phrased with V (N= 90) and the group of sentences where S was phrased in a different IP than V (N=108), *cf.* §3.1; and we compared the group of sentences where O was phrased with V (N= 178) and the group of sentences where O was phrased in a different IP than V (N = 20), *cf.* §3.2.

Data were analyzed by mean of Nominal Logistic Regression within SPSS (v. 21.1). Due to the fact that the number of syllables and the number of PWs were strongly correlated for S ($r = 0.808$, $p < 0.01$), for V ($r = 0.718$, $p < 0.01$) and for O ($r = 0.786$, $p < 0.01$), and in order to avoid effects of co-variables, two different models were run to test the effect of prosodic branchingness. Note that due to the small amount of data, we did not include speaking style as a predictor.

3.1. S prosodic phrasing

First, a model was run with S phrasing status as the dependent variables (possible value: within an independent MaP/within the same MaP as V), and the following variables as predictors: syntactic branching of the element (branching/non-branching), number of syllables in S, number of syllables in V, number of syllables in O, the interaction between the number of syllables in S and the number of syllables in V, and the interaction between the number of syllables in V and the number of syllables in O.

Apart from syntactic constraints, results of the model revealed that the articulation rate has a significant impact on the phrasing of S ($\beta = 0.21$, $z = 20.114$, $p < 0.001$). As it can be seen in Figure 1 below, syllabic mean duration of the SVO sentence is greater (articulation rate slower) in cases where S is phrased as an independent MaP.

An effect of the number of syllables in the subject phrase was also found ($\beta = 0.404$, $z = 4.507$, $p < 0.05$). This shows that the more syllables are contained in S, the bigger the chances for S to be phrased as an independent MaP. In other words, an S containing 12 syllables stands in greater chance to be phrased in a separate MaP, separated from the verb, in comparison with an S containing 2 syllables. Prosodically embedded S constituents contain fewer syllables than prosodically independent S constituents; this can be seen by comparing the mean number of syllables in S when it is phrased as an independent MaPP and when it is phrased in the same MaPP as V (see Figure 2).

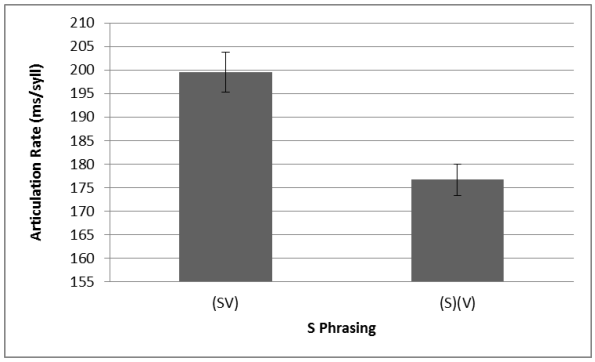


Figure 1: *Articulation rate (in ms/syll) by mean of S phrasing. Errors bars are standard errors from the mean.*

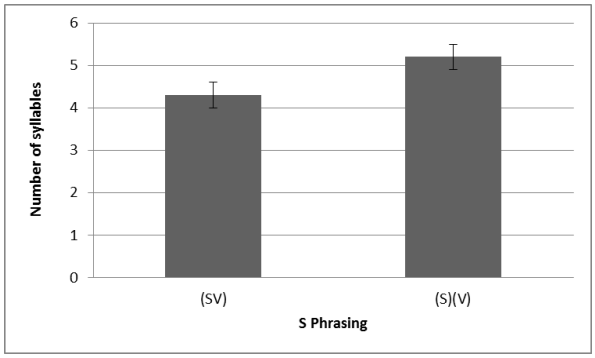


Figure 2: *Number of syllables by mean of S phrasing. Errors bars are standard errors from the mean.*

A marginal effect of the number of syllables in V was also observed ($\beta = 0.582$, $z = 2.896$, $p = 0.079$), revealing that the more syllables V contains, the greater the chances for S to be phrased in a single MaP. Nevertheless, it appeared that the number of syllables in S and the number of syllables in V interact ($\beta = 0.109$, $z = 6.543$, $p < 0.01$). The effect shows that when S is long and V is short, the chances to find a Major Prosodic Phrase boundary after S are stronger than when S is short and V is long. In the latter case, there are smaller chances for S to be followed by an MaP boundary.

The same model with the number of PWs instead of the number of syllables per constituent show similar results, with an effect of articulation rate ($\beta = 0.011$, $z = 6.087$, $p < 0.05$), an effect of the number of PWs in S ($\beta = 1.335$, $z = 6.520$, $p < 0.05$).

3.2. O prosodic phrasing

A first model was run with O phrasing status as the dependent variable (possible values: within a single MaP/within the same MaP than V), and the following variables as predictors: syntactic branchingness of the element (branching/non-branching), number of syllables in S, number of syllables in V, number of syllables in O, the interaction between the number of syllables in S and the number of syllables in V, and the

interaction between the number of syllables in V and the number of syllables in O.

Only the articulation rate appeared to be a significant predictor in this model. Articulation Rate had impact on O phrasing ($\beta = 0.21$, $z = 119.745$, $p < 0.001$), and, as can be seen in Figure 3 below, in the cases where O is phrased as an independent MaP, syllabic mean duration of the SVO sentence is greater (articulation rate slower).

The same model with the number of PWs instead of the number of syllables per constituent produced similar results, again showing solely an effect of articulation rate ($\beta = 0.021$, $z = 13.554$, $p < 0.001$).

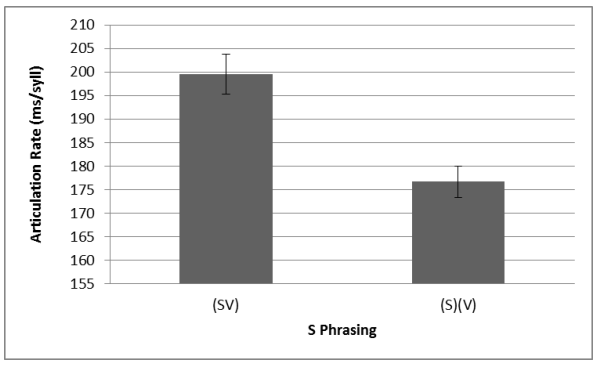


Figure 3: *articulation rate (in ms/syll) by mean of O phrasing. Errors bars are standard errors from the mean.*

4. Discussion

The analysis revealed interesting results regarding the phrasing pattern for SVO, and the effect of the different non syntactic factors constraining prosodic phrasing.

Regarding phrasing patterns, it appeared that S can be phrased with or without V with almost the same probability (in 54.1% of the data, S is phrased as an independent MaP, whereas in 45.6% of the data, it is in the same MaP than V). This is an important result, which indicates that the phrasing of S in a single MaP is not the default pattern in French, contrary to what has been claimed in the literature. In addition, this distribution of the various phrasing patterns shows that French differs from other Romance Language, such as Spanish and Catalan, where S is usually phrased in an independent MaP (see [11] and [12]). On the other hand, Object NP constituents seldom form a single MaPP: they are phrased with the V in more than 90% of the cases.

Regarding the factors constraining phrasing, our results show that syntactic complexity does not seem to be an important parameter to predict the phrasing of SVO structures. Indeed, the effect of syntactic branchingness of either O or S was never significant in our data: it does not help to explain why in some cases S or O are phrased separately or in the same MaP than the V. Articulation rate has a significant effect on phrasing of S and of O, supporting what has been found in the literature regarding the effects of articulation rate on phrasing (see [22] and [23]), i.e.: the faster a speaker articulates, the greater the chances for him to obliterate prosodic boundaries, and therefore to phrase the subject in the same MaP than the V. The number of syllables does not have any effect on the way the object NP is phrased: short or long Os are phrased in the same MaP or in an independent MaP because of different reasons. However, it clearly has an effect on the way the NP subject is phrased, supporting the idea the heavy NP subjects tend to be phrased in independent prosodic phrases, and light NPs to be embedded in one MaP along with the verb that follows (see [8]). It appeared that the number of syllables in V interacted with the number of syllables in S, confirming the idea that balance effects also have an importance on phrasing (see [8]). When calculated in term of prosodic branchingness, i.e. in terms of number of PWs, the lengths of V and O have not been observed to have any effect on phrasing.

It is necessary to conduct further analyses on a larger set of data in order to determine which parameters are the most robust for evaluating prosodic weight in French (number of syllables or number of prosodic words), in particular when analyzing phrasing patterns in SVO structure. According to [11], the choice of on parameter over the other is language-specific.

5. Conclusions

The aim of this paper was to test the impact of different phonological and syntactic factors on prosodic phrasing of French SVO structures. A set of approximately 200 sentences were extracted from 4 corpora of different speaking styles, and were semi-automatically annotated to study prosodic phrasing at the level of the Major Prosodic Phrase. The results have shown that NP subjects could be phrased within the same MaP than the verb, or in an independent MaP. The decision for one phrasing over the other is mainly correlated with articulation rate (the faster the speaker articulates, the greater the chances for the subject NPs to be phrased in an independent MaP) and on prosodic weight, be it calculated in terms of number of syllables or in terms of number of PWs in S (the longer the S, the greater the chances for S to be phrased as an independent MaP). Regarding the phrasing of O, results indicate that this constituent was rarely phrased autonomously. Furthermore, whenever it was realized as an independent MaP, the phrasing obtained was motivated by a slow articulation rate, and not by metrical properties (expressed in terms of number of syllables or in terms of number of PWs). Finally, for both S and O, no effect of syntactic branching was found. Further developments are needed to confirm these results, in particular in enlarging the set of analyzed data.

6. Acknowledgements

This paper is related to the work package “Prosodic phrasing and prosodic hierarchy: a data driven approach” of the Labex Empirical Foundations of Linguistics and is supported by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083). It is also partly achieved within the research project ANR-CONTINT 2011 PHOREVOX funded by ANR/CGI.

7. References

[1] Selkirk, E., “On derived Domains in Sentence Phonology “, Phonology 3, pp. 371-405, 1986.

[2] Nespor, M., & Vogel, I., Prosodic Phonology, Foris Publication, 1986.

[3] Inkelas, S., & Zec D., ‘Syntax-phonology interface’, in J. Goldsmith (ed), The Handbook of phonological theory, Blackwell Publishers, pp. 535-549, 1995.

[4] Selkirk, E., 'The Syntax-Phonology Interface', in J. Goldsmith, J. Riggle & A. Yu (Eds.), The Handbook of Phonological Theory, Blackwell Publishing, pp. 435-484, 2011.
[5] Delais-Roussarie, E., Pour une approche parallèle de la structure prosodique: Etude de l'organisation prosodique et rythmique de la phrase française, Thèse de Doctorat, Université de Toulouse-Le Mirail, 1995.
[6] Rossi, M., L'intonation, le système du Français : description et modélisation, Ophrys, 1999.
[7] Dell, F., 'L'accentuation dans les phrases en français', in F. Dell, D. Hirst & J.R. Vergnaud (Eds.), Forme sonore du langage: structure des représentations en phonologie, Hermann, pp. 65-122, 1984.
[8] Delais-Roussarie, E., "Phonological phrasing and accentuation in French", in M. Nespor & N. Smith (Eds.), Dam phonology : HIL phonology papers II, Holland Academic Graphics, pp. 1-38, 1996.
[9] Ghini, M., "Φ-formation in Italian: A new proposal", Toronto Working Papers in Linguistics 12, pp. 41-78, 1993.
[10] Martin, P., Prosodic and Rhythmic Structures in French. Linguistics, 25, 925-949, 1987.
[11] Elordieta, G., Frota, S. & Vigário, M., 'Subjects, objects and intonational phrasing in Spanish and Portuguese', Studia Linguistica 59, 110-143, 2005.
[12] D'Imperio, M., Elordieta, G., Frota, S., Prieto, P., and Vigário, M., "Intonational Phrasing in Romance: The role of prosodic and syntactic structure", in S. Frota, M. Vigário & M.-J. Freitas (eds), Prosodies, Mouton de Gruyter, 59-97, 2005.
[13] Durand, J., Laks, B., & Lyche, C. (Eds.), Phonologie, variation et accents du français, Hermes, 2009.
[14] Boersma, P., & Weenink, D., Praat (Version 5.3), Retrieved from http://www.fon.hum.uva.nl/praat/, 1995-2013.
[15] Goldman, J.-P. EasyAlign: an Automatic Phonetic Alignment Tool under Praat. Proceedings of Interspeech.
[16] Christodoulides, G., Avanzi, M., Goldman, J.-P. 'DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech'. Proceedings of LREC, 2014.
[17] Mertens, P., Goldman, J.-P., Wehrli, E., & Gaudinat, A., 'La synthèse de l'intonation à partir de structures syntaxiques riches', Traitement Automatique des Langues, 42, pp. 145-192, 2001.
[18] Goldman, J.-P., Auchlin, A., Roekhaut, S., Simon, A. C., & Avanzi, M., 'Prominence Perception and Accent Detection in French. A Corpus-based Account' in Proceedings of Speech Prosody, 2010.
[19] Frota, S., "Prosodic structure, constituents and their representations", in A. Cohn, C. Fougeron & M. Huffman (eds), The Oxford Handbook of Laboratory Phonology Oxford, Oxford University Press, Chapter 11, pp. 255-265, 2012.
[20] Avanzi, M., Obin, N., Lacheret-Dujour, A., & Victorri, B. (2011). Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree. Paper presented at the Proceedings of Interspeech.
[21] Miller, J.L., Grosjean, F., Lomato, C., "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications". Phonetica, 41, 215-225, 1984.
[22] Fougeron, C., & Jun, S. A. (1998). Rate Effects on French Intonation: Prosodic Organization and Phonetic Realization. Journal of Phonetics, 26, 45-69.
[23] Post, B. (2011). The multi-facetted relation between phrasing and intonation contours in French. In C. Gabriel & C. Lleó (Eds.), Intonational Phrasing in Romance and Germanic: Cross-linguistic and bilingual studies (pp. 43–74). Amsterdam: John Benjamins.

Intonational cues to item position in lists: evidence from a serial recall task

Michelina Savino[1], *Andrea Bosco*[1], *Martine Grice*[2]

[1] Dept. of Education, Psychology, Communication, University of Bari, Italy
[2] IfL-Phonetics, University of Cologne, Germany
michelina.savino@uniba.it, andrea.bosco@uniba.it, martine.grice@uni-koeln.de

Abstract

Intonation can convey information about how lists are structured into groups, as well as about specific item positions within a group. In Bari Italian, this function is expressed by three different tunes a) a rising contour, signalling that the list has not yet been completed; b) a high-rising contour, marking the penultimate item, i.e. signalling that the end of the list is approaching; c) a falling contour, marking the last item, i.e. cueing the end of the sequence. In this paper we explore the effects of such intonational information on working memory. In particular, we demonstrate that when listeners are requested to recall spoken nine-digit sequences by strictly following their serial order, their performance is significantly better when lists are characterised by tunes of the type described above, compared to sequences whose items are marked by a neutral, peak accent and/or are grouped by inserting a silent pause. We also observed that recall of items marked by specific contours at positions 3, 6 and 9 is particularly enhanced at these positions, whereas in sequences also containing intonational cues to items in penultimate position (2, 5 and 8) recall of those items is not equally improved. Therefore, it appears that in serial recall of spoken sequences, even when a large number of specific intonational cues to serial positions are available, listeners can use only a selection of them.

Index Terms: list intonation, serial recall, working memory

1. Introduction

In prosody research, the role of intonation in signalling discourse structure is widely acknowledged, as it cues hierarchical relationships among phrases within discourse units [1]. Specifically in lists, intonation can convey information about how they are structured into groups of items, as well as about specific item positions within a group. In this paper we explore the effects of the use of such intonational information on working memory, in particular in a serial recall task. This task consists in recalling lists of digits by following their strict serial order of presentation, and it is typically used as a test in psychology for assessing individual's short-term memory span [2]. In this research field, there is quite a large body of literature attesting the relationship between verbal serial recall and prosody, mainly consisting in the observation that lists of spoken items are better recalled when they are presented in groups (so called "grouping effect", see [3], [4], [5] among others). As to the specific role of intonation in enhancing serial recall, [6] and [7] provided evidence that it is limited to the grouping effect, i.e. it is equivalent to that triggered by pause insertion. However, in these studies results might be affected by the scarce control over the most suitable "position-informative" intonation patterns available for use in creating sequence stimuli. We hypothesise that if sequences to be recalled are produced by suitable intonation contours cueing specific positional information of items, listeners can fruitfully make use of them in terms of serial recall enhancement. In previous production and perception studies on Bari Italian, it has been shown that a rich inventory of tunes is available to speakers for signalling hierarchical relationships within discourse units at various levels ([8], [9], [10], see comparable strategies in Dutch [11] [12]), as well as specifically for cueing positional information in sequences. The most typical contours are:

- a rising contour, signalling that the list has not yet been completed ("non-final" contour, L* L-H%);
- a high-rising contour, marking the penultimate item, i.e. signalling that the end of the list is approaching ("pre-final" contour, H* H-H%);
- a falling contour, marking the last item, i.e. cueing the end of the sequence ("final" contour, H+L* L-L%).

The aim of this study is to explore the effect of positional information cued by those tunes in a serial recall task involving Bari Italian listeners.

2. Methodology

We identified two intonational patterns (we called 'Intonation contour A' and 'Intonation contour B') characterised by F0 shapes conveying hierarchical organisation of groups within a sequence, as well as positional information of items across and within groups. In a nine-digit sequence, we determined:

- 'Intonation Contour A', consisting of the "non-final" rising contour at positions 3 and 6, and a low-falling ("final") contour at position 9. Items at initial- and within-group positions (positions 1, 2, 4, 5, 7, 8) all have a peak accent, taken to be the neutral unmarked pattern. A scheme of intonation contour A is shown in Figure 1;
- 'Intonation Contour B', sharing the same intonational patterns of 'Intonation Contour A', except for a) a steep rising pitch accent (followed by a mid-fall) at positions 2 and 5, which pre-signals the end of the first and the second groups, i.e. penultimate position in the two non-final groups within the sequence; and b) a high-rising ("pre-final") contour at position 8 pre-signalling both the end of the third group *and* the end of the whole sequence. Intonation Contour B is schematised in Figure 2.

These two experimental conditions were compared with two further ones, namely:

- 'Grouped by Pause' sequences, where all digits have a peak contour, and sequences are grouped by inserting a pause at the end of each three-digit (sub)sequence, as schematised in Figure 3;
- 'Ungrouped' (Control) sequences, sharing the same intonation of the 'Grouped by Pause', but without pause grouping, as schematised in Figure 4.

We hypothesise that serial recall performance would be:
1) better in both Intonation Contours A and B and the 'Grouped by Pause' conditions than in the 'Ungrouped' (Control) condition (due to the grouping effect);
2) better in both 'Contour A' and 'Contour B' than in the 'Grouped by Pause' condition, because of the absence of intonational marking of item position in the latter condition. In

particular, at least items in positions 3 (last item in the first serial group), 6 (last item in the second serial group), and 9 (last item in the third serial group *and* in the whole sequence) should benefit in terms of recall enhancement;
3) better in 'Intonation Contour B' than 'Intonation Contour A' because of the enhanced hierarchical and positional information conveyed by intonation in certain positions in Contour B, namely: digits at positions 2 ("pre-final" contour= item at mid position in the first group), 5 ("pre-final" contour=item at mid position in the medial group), and 8 ("pre-final" contour=item at mid position in the last group *and* penultimate item in the whole sequence).

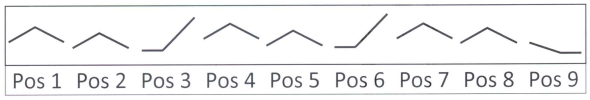


Figure 1: *Schematisation of sequence stimuli as realised according to the 'Intonation Contour A' condition.*

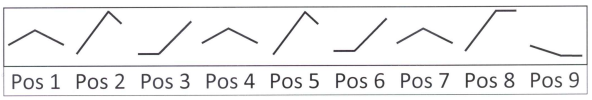


Figure 2: *Schematisation of sequence stimuli as realised according to the 'Intonation Contour B' condition.*

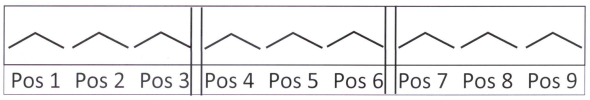


Figure 3: *Schematisation of sequence stimuli as realised according to the 'Grouped by Pause' condition, that is a neutral, peak contour on each item, and item-grouping realised by pause insertion.*

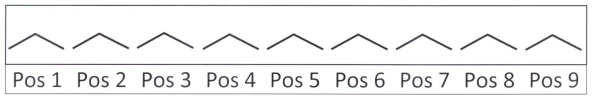


Figure 4: *Schematisation of sequence stimuli as realised according to the 'Ungrouped' (control) condition, that is a neutral, peak contour on each item, and no item-grouping.*

2.1. Stimuli

In order to produce sequences according to the conditions above, three types of stimuli for each digit (1-9) were created:

- type (a), where each digit in the sequence was realised with the unmarked, neutral F0 peak, as described above;
- type (b), where digit sequences were realised with Intonation contour A;
- type (c), where digit sequences were produced with Intonation contour B.

All series for each of the digits were produced by a trained native speaker of Bari Italian (author MS) in the same recording session. Therefore, nine sequences with the same digit (one for each digit) were produced with Intonation Contour A, nine with Intonation Contour B, and nine by realising each item with a neutral, "citation-form" intonation. In this way, all intonational realisations in each position (first, second, third, fourth, etc.) within each contour type were available for each digit. They were saved as individual audio files, and used as "building blocks" for creating all the nine-digit spoken sequences under the four conditions. Stimuli were created by concatenating the individual audio files into nine-digit sequences. In a post-editing step, care was taken that speech signal amplitude was homogeneous in all sequences. Spoken digit realisations of type (a) were used for creating sequences for the conditions 'Ungrouped', and 'Grouped by Pause', in the latter case by inserting a 310 ms silence after digits in positions 3 and 6. Spoken digit renditions of types (b) and (c) were used for creating sequences for 'Intonation contour A' and 'Intonation contour B', respectively. An example sequence for each of the experimental conditions is shown in Figures 5-8. For example, the stimulus for the 'Contour A' condition shown in Figure 7 (sequence: four-seven-three six-one-eight two-five-nine) was realised by concatenating the spoken digit "four" as realised in the first position of the whole 'digit four' sequence with Contour A, with the spoken digit "seven" as realised in the second position of the whole 'digit seven' sequence with contour A, and so on.

We created 68 nine-digit lists from pseudo-random permutation of the 1-9 digits, avoiding two adjacent digits in ascending or descending order, and making sure that a digit did not appear in the same position in consecutive lists. The concatenated nine-digit sequences were created on the basis of these lists, the duration of each sequence averaging 6.4 sec. We produced 17 stimuli for each of the four conditions, for a total amount of 68 stimuli (including 8 to be used for the training session, 2 per condition). All steps for the preparation of stimuli were carried out using Praat software tool for speech analysis [13].

2.2. Subjects

Twenty-nine informants (23 females and 6 males) took part in the experimental sessions. They were aged 20-45 (average =22.4), and reported no speech or hearing deficits. They were students of Psychology at the University of Bari, all born and living in the Bari dialectal area. None of them had a background in linguistics or prosody. They were given one exam credit as a reward for participating in the experiment.

Before starting the task, subjects were tested as to their short-term memory span by means of the standard Digit Span (DS) test of WAIS-R [14], which resulted homogeneous across groups (minimum DS=5).

2.3. Procedure

Participants were tested individually in a quiet laboratory, sitting in front of a computer and wearing a headset with headphones and microphone. They were instructed to listen to each sequence and recall all the nine digits orally by strictly following their order of presentation. Spoken responses were directly recorded to disk. Each list was preceded by a warning tone and 500 ms silence, and after each spoken response subjects could proceed with the next sequence by pressing the spacebar. They were allowed to pause whenever they wanted during the session, and they were encouraged to take a break after every block of 15 stimuli. Participants were asked to recall a total of 60 sequences (15 for each condition), preceded by a short (8 stimuli) training session. The order of presentation was balanced across the subjects. The whole session (i.e. including the DS test) lasted approximatly 40 min for each

informant. The experiment was implemented and run using SuperLab 2.0.

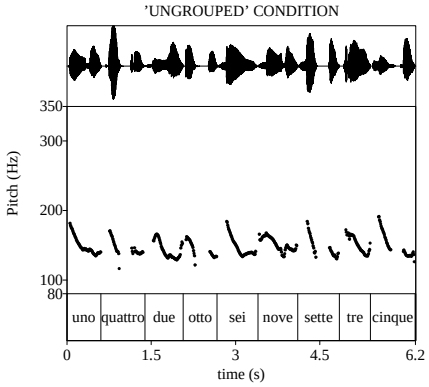


Figure 5: *Speech waveform and F0 contour of one of the stimuli for the 'Ungrouped' (Control) condition.*

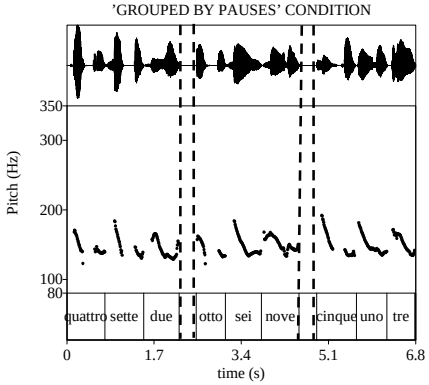


Figure 6: *Speech waveform and F0 contour of one of the stimuli for the 'Grouped by Pause' condition. Vertical broken lines mark silent intervals (pauses) between groups.*

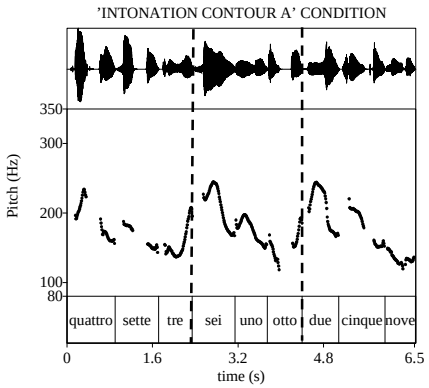


Figure 7: *Speech waveform and F0 contour of one of the stimuli for the 'Intonation contour A' condition. Vertical broken lines mark the right edge of each group (intonational phrase)*

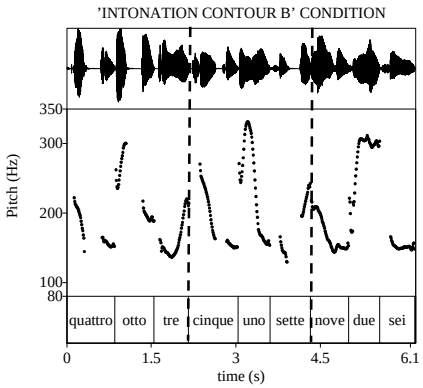


Figure 8*: Speech waveform and F0 contour of one of the stimuli for the 'Intonation contour B' condition. Vertical broken lines mark the right edge of each group (intonational phrase).*

3. Results and Discussion

A mixed factors general linear model was carried out, with: 1) Condition (4 levels: Ungrouped, Grouped by Pause, Intonation Contour A, Intonation Contour B), 2) Serial Group Within the Sequence (3 levels: first, second, third), 3) Within-Group Position (3 levels: first, second, third), as factors.

Results (Figures 9-12) show a very large effect of Condition: $F(3; 84) = 26.42$; $p<0.001$, in that sequences in the 'Ungrouped' (Control) condition are recalled significantly worse than those in the three remaining conditions. This confirms the general findings on serial memory of verbal sequences that they are recalled better when they are grouped prosodically ("grouping effect"). Most interestingly, lists produced with both Contours A and B show statistically better recall performance with respect to those produced by inserting silent pauses between serial groups (Fisher LSD post hoc test, $p<0.05$). This confirms our prediction that intonation plays a specific role in serial recall enhancement beyond the grouping effect. On the other hand, our hypothesis that Contour B sequences would perform better than Contour A sequences is not confirmed. We also found a significant second order interaction effect between Condition, Serial Group and Within-Group Position: $F(12; 336) = 3.06$; $p<0.001$; again, Contour A and Contour B conditions showed better recall performance than sequences Grouped by Pause, but without any difference between Contour A and Contour B sequences. We then looked at recall performance of specific intonationally-marked positions in the sequences, namely positions 3, 6 ("non-final" contour in both Contour A and Contour B conditions), positions 2, 5, 8 ("pre-final" contour in Contour B condition), and position 9 ("final" contour in both Contour A and Contour B conditions). We performed Planned Comparisons between these positions in the relevant conditions, namely:

- position 3 in Contours A/B vs. Grouped by Pause conditions: recall performance was significantly better in both Contour A and Contour B ($p<0.01$);
- position 6 in Contours A/B vs Grouped by Pause conditions: recall performance was significantly better in Contour B ($p<0.05$), but not in Contour A (only approaching statistical significance, $p=0.06$);

- position 9 in Contours A/B vs Grouped by Pause conditions: again, recall performance was significantly better in both Contours A and B ($p<0.01$);
- positions 2, 5, 8 in Contour A vs Contour B conditions: recall performance did not show any statistical difference.

These outcomes indicate that serial recall of items marked by "non-final" contours (i.e. those marking positions corresponding to the end of a serial group) and "final" (marking the end of a sequence) is particularly enhanced. Interestingly, the effect of positional information cued by a "final" F0 contour (position 9) is so strong that it significantly increases the recency effect. On the other hand, listeners did not make use of "pre-final" intonational information during the recalling phase, which is the reason why overall recall performance of Contour B sequences was not significantly better than those with Contour A, contrary to our hypothesis.

Therefore, it appears that there is a limit to the number of intonational cues to positions which listeners can use, and that those marking the end of a serial group (at positions 3 and 6) or the end of the sequence (at position 9) are the most likely to be used.

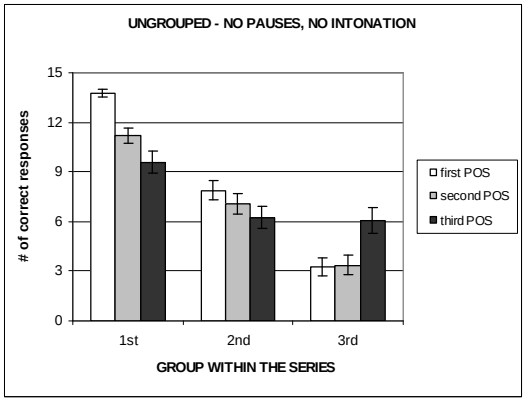


Figure 9: *Correct recall (mean values) across the 3 serial groups within a sequence, and the position within each serial group (first POS, second POS, third POS), for the 'Ungrouped' (Control) condition.*

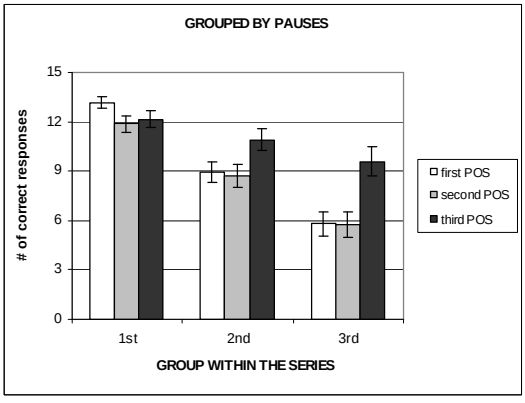


Figure 10: *Correct recall (mean values) across the 3 serial groups within a sequence, and the position within each serial group (first POS, second POS, third POS), for the 'Grouped by Pause' condition.*

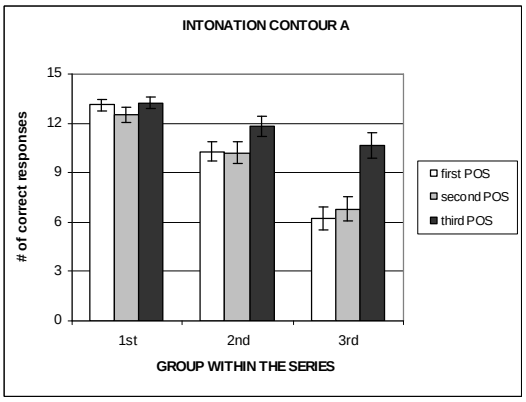


Figure 11: *Correct recall (mean values) across the 3 serial groups within a sequence, and the position within each serial group (first POS, second POS, third POS), for the 'Intonation Contour A'condition.*

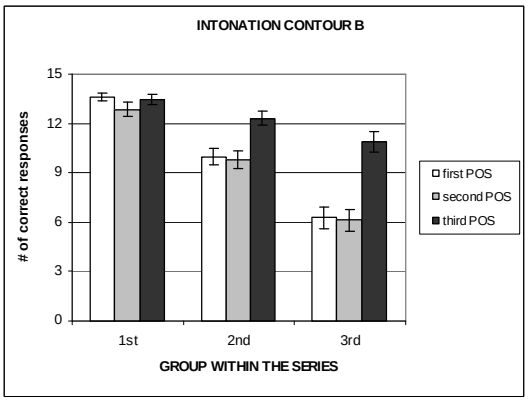


Figure 12*: Correct recall (mean values) across the 3 serial groups within a sequence, and the position within each serial group (first POS, second POS, third POS), for the 'Intonation Contour B'condition.*

4. Conclusions

In this paper, we explored the effects of positional information conveyed by intonation on serial memory. In a serial recall task, we observed that when listeners are presented nine-digit lists with F0 contours marking the hierarchical organisation of items within the sequence, their recall performance is significantly better than in cases where such intonational information is absent. However, it appears that intonational enhancement of serial recall has its limits, since we also observed that even when a large number of specific intonational cues to serial position are available, listeners are only able to make use of a selection of them. Such limitations might be due to cognitive processing needs imposed by working memory.

5. Acknowledgements

We would like to thank Ralf Rummer for helpful discussion and advice, and our student assistants in Bari Mirco Lacalandra and Gabriella Monticelli for help collecting the experimental data.

6. References

[1] Hirschberg, J. and Pierrehumbert, J. "The intonational structuring of discourse", Proceedings of the 24th Annual Meeting of the Association of Computational Linguistics, 136-144, 1996.

[2] Baddeley A., Eysenck M.W., Anderson, M. C., "Memory", Psychology Press, 2009.

[3] Frankish, C., "Modality-specific grouping effect in short-term memory", Journal of Memory and Language, 24: 200-209, 1985.

[4] Cowan, N. Saults, J. S., Elliott, E. M., and Moreno, M. V., "Deconfounding serial recall", Journal of Memory and Language 46:153-177, 2002.

[5] Reeves, C., Schmauder A. R., and Morris, R. K. "Stress grouping improves performance on an immediate serial list recall task", Journal of Experimental Psychology: Learning, Memory and Cognition, 26 (6): 1638-1654, 2000.

[6] Frankish, C. "Intonation and auditory grouping in immediate serial recall", Applied Cognitive Psychology, 9: 5-22, 1995.

[7] Saito, S., "Effects of articulatory suppression on immediate serial recall of temporarily grouped and intonated lists", Psychologia, 41: 95-101, 1998.

[8] Savino, M., "Non-finality and pre-finality in Bari Italian intonation: a preliminary account", Proceedings of the VII European Conference on Speech Communication and Technology, 2: 939-942, 2001.

[9] Savino, M., "Intonational cues to discourse structure in a variety of Italian", in P. Gilles and J. Peters [Eds], Regional variation in intonation, 161-187, Niemeyer, 2004.

[10] Savino M., Grice, M., Gili Fivela, B. Marotta, G., "Intonational cues to discourse structure in Bari and Pisa Italian: perceptual evidence", Proceedings of Speech Prosody (on CD-ROM), 2006.

[11] Geluykens, R., Swerts, M., "Prosodic cues to discourse boundaries in experimental dialogues", Speech Communication, 15: 69-77, 1994.

[12] Swerts, M., Collier, R., Terken, J., "Prosodic predictors of discourse finality in spontaneous monologues", Speech Communication, 15: 79-90, 1994.

[13] Boersma, P. and Weenink, D., "Praat, a system for doing phonetics by computer", Glot International, 5(9/10): 131-151, 2001.

[14] Wechsler, D., "Manual for WAIS-R" The Psychological Corporation, 1987.

Prosodic focus-marking in Chinese four- and eight-year-olds

Anqi Yang [1], *Aoju Chen*[1, 2]

[1] Utrecht Institute of Linguistics, Utrecht University, the Netherlands
[2] Max Planck Institute for Psycholinguistics, the Netherlands
a.yang1@uu.nl, aoju.chen@uu.nl

Abstract

This study investigates how Mandarin Chinese speaking children use prosody to distinguish focus from non-focus, and focus types differing in size of constituent and contrastivity. SVO sentences were elicited from four- and eight-year-olds in a game setting. Sentence-medial verbs were acoustically analysed for both duration and pitch range in different focus conditions. The children started to use duration to differentiate focus from non-focus at the age of four. But their use of pitch range varied with age and depended on non-focus conditions (pre- vs. post-focus) and the lexical tones of the verbs. Further, the children in both age groups used pitch range but not duration to differentiate narrow focus from broad focus, and they did not differentiate contrastive narrow focus from non-contrastive narrow focus using duration or pitch range. The results indicated that Chinese children acquire the prosodic means (duration and pitch range) of marking focus in stages, and their acquisition of these two means appear to be early, compared to children speaking an intonation language, for example, Dutch.

Index Terms: focus, tone, Mandarin Chinese; L1 acquisition

1. Introduction

The term 'focus' refers to an information structural category and is defined as the new information in a sentence to the receiver [e.g. 1, 2]. This study involves three types of focus, i.e. narrow focus, contrastive focus and broad focus. The former two differ from the latter in the size of the focus constituent, e.g. a lexical word (narrow focus, contrastive focus) vs. a whole sentence (broad focus). Narrow focus and contrastive focus differ in that the latter conveys an explicit contrast to alternatives in the context.

Prosodic focus-marking in adult Mandarin Chinese (hereafter Mandarin) has been extensively studied. It is generally agreed that a focused constituent has a longer duration, a higher pitch level and/or a wider pitch range than the same constituent in the broad focus condition [e.g. 3, 4, 5] Furthermore, the post-focus part of the sentence is usually compressed in pitch (i.e. spoken with a lower pitch level or a smaller pitch range) and duration, while the pre-focus part undergoes little change in pitch or duration [e.g. 4, 6, 7, 8]. However, the difference between narrow focus and contrastive focus is less conclusive. Some researchers have reported that contrastive focus induces a wider pitch range than narrow focus in sentence-initial position when the focused constituent has a certain tonal composition [5]. Yet [9] have found neither pitch range nor durational differences between narrow and contrastive focus.

In contrast, little is known on how Mandarin-speaking children use prosody to mark focus. Studies on other languages have revealed that children learn to use prosody to mark focus in their respective languages in stages [10]. For example, English-speaking children can use accentuation to highlight contrastive focus by the age of three, and from three to six this use of accentuation is further consolidated [11, 12]. Dutch-speaking children can use accentuation to mark focus at the age of four or five but become adult-like in choice of accent type only at the age of seven or eight [10, 13]. Further, they cannot vary the phonetic realisation of a pitch accent in terms of pitch range for focus-marking purposes until the age of seven or eight [14]. The use of duration for this purpose is still not acquired at the age of seven or eight [14].

The current study is a first study examining Mandarin-speaking children's use of pitch and duration in focus-marking. As Mandarin uses pitch not only to mark focus and express other sentence-level meanings but also to distinguishing lexical meanings, acquiring Mandarin entails that children have to learn both functions of pitch. The question that arises is whether Mandarin-speaking children follow a similar developmental trajectory to children speaking a non-tonal language in prosodic focus-marking. As a first step towards addressing this question, we have investigated (1) how Mandarin-speaking children use prosody to distinguish focus from non-focus, (2) how they distinguish focus in different constituent-sizes (narrow focus vs. broad focus), and (3) how they distinguish contrastive focus from non-contrastive focus.

2. Method

2.1 Target sentences

We aimed to elicit 160 SVO sentences from participants: (5 focus conditions x 4 tones in the verbs x 4 tones in the object-nouns x 2 types of verbs and object nouns). The five focus conditions were: (1) Narrow focus on the subject in sentence-initial position (NF-i); (2) Narrow focus on the verb in sentence-medial position (NF-m); (3) Narrow focus on the object in sentence-final position (NF-f); (4) Contrastive focus on the verb in sentence-medial position (CF-m); (5) Broad focus over a whole sentence (BF). Four lexical tones were used in the verbs and object-nouns. Two types of verbs and corresponding object nouns were included (Table 1). Four subject nouns (cat, bear, dog, and rabbit) were evenly distributed over the sentences. Crucially, all words were selected from the words that are acquired by Mandarin-speaking children by the age of three or four [15]. The 160 sentences were split evenly into two lists (List 1 & 2) of 80 sentences such that each list contained all target words and all tonal combinations but not all word combinations of the verbs and objects. Half of the participants produced the sentences on List 1 and the other half produced the sentences on List 2.

Verb – type1	Verb – type 2	Noun - type1	Noun - type 2
T1 rēng (throw)	T1 jiāo (water)	T1 shū (book)	T1 huā (flower)
T2 mái (bury)	T2 wén (smell)	T2 qíu (ball)	T2 lí (pear)
T3 jiǎn (cut)	T3 tiǎn (lick)	T3 bǐ (pen)	T3 cǎo (grass)
T4 yùn (transport)	T4 mài (sell)	T4 cài (vegetable)	T4 shù (tree)

Table 1: Two types of verbs and object nouns

2.2 Speech elicitation

To elicit the target sentences, question-answer dialogues between the experimenter and the child as illustrated in examples (1) to (5) were embedded in a picture-matching game adapted from [10].

(1) Exp: Look! A book, and the book is in the air. It looks like someone throws the book. Who throws the book?

Child: [The rabbit] throws the book. (NF-i)

(2) Exp: Look! A rabbit, and there is also a book. It looks like the rabbit does something to the book. What does the rabbit do to the book?

Child: The rabbit [throws] the book. (NF-m)

(3) Exp: Look! A rabbit, and its arm is stretched out. It looks like the rabbit throws something. What does the rabbit throw?

Child: The rabbit throws [the book]. (NF-f)

(4) Exp: Look! A rabbit, and a book. It looks like the rabbit will do something to the book. I will make a guess: The rabbit cuts the book.

Child: The rabbit [throws] the book. (CF-m)

(5) Exp: Look! This picture is very blurring. I cannot see anything clearly. What happens in the picture?

Child: [The rabbit throws the book]. (BF)

In the game, the child's task was to help the experimenter to put pictures in matched pairs. Three piles of pictures were used. The experimenter and the child each held a pile of pictures; the third pile laid around on the table in a seemingly 'messy' fashion. The experimenter's pictures always missed some information, e.g. the subject, the action, the object or all the three pieces of information. The child's pictures always contained all the three pieces of information. In every trial, the experimenter showed a picture of hers to the child, described the picture and asked a question about it, as illustrated in (1) to (5). The child took a look at the corresponding picture in his pile and answered the question or made a correction (in the CF condition). The experimenter could then look for the right picture in the messy pile and matched it with her own picture to form a pair. Crucially, as rules of the game, the child was asked to answer the experimenter's question in full sentences and not to reveal his pictures to the experimenter.

In order to familarise the child with the game procedure, the experimenter started the game with five practice trials involving all five focus conditions. Prior to the practice session, the experimenter conducted a picture-naming task to make sure that the children would use the intended words to refer to the entities in the pictures.

2.3 Participants

Thirty-six children from three age groups (four-five yrs, even-eight yrs, ten-eleven yrs, twelve per group) participated in the experiment. They were tested individually in a quiet room in their kindergartens or schools in Beijing. In addition, fifteen university students speaking Mandarin were tested as controls, following the same procedure. Considering children's limited concentration capacity, the 80 sentences on each list were elicited in two sessions of 20 – 35 minutes on two different days. The adults and children were both audio and video-recorded during the experiments. The current paper presents results from four four-year-olds and four eight-year-olds.

2.4 Annotation and acoustic analysis

The audio recording from each child was orthographically annotated at three levels using Praat: trial, question from the experimenter, and response from the child. Usable sentences were then carefully selected from the recordings. Responses deviating from the target sentences in choice of word or word order or produced with self-repairs and hesitations were considered unusable and excluded from further analysis. In total, 432 sentences from the eight children were included in the analysis.

The usable sentences were then acoustically annotated at the word level and at the pitch level. Landmarks indicating word-onsets and word-offsets and the locations of the maximum pitch and minimum pitch within each word were inserted in Praat textgrids for each sentence. It is worth noting that Mandarin is a tone language, and each tone has a particular target to reach. According to [16], the pitch contour approximates to or reaches at the target towards the end of a syllable. In this study the tonal targets were taken into account. For Tone 2 (rising tone) and Tone 4 (falling tone), it was presumed that their pitch contour approach to or reach at the high/low target respectively at the syllable offset. To be more specific, the maximum pitch of Tone 2 was always labeled and measured on the right side of its minimum pitch, even though sometimes there was an even higher pitch occurring on the left side due to the influence of the preceding tone. Similarly, the minimum pitch of Tone 4 was obtained on the right side of its maximum pitch. For Tone 1 (flat tone), its maximum and minimum pitch were obtained regardless of their relative order of occurrence. The pitch contour of Tone 3 varied most, and three patterns were observed in the data, namely, fall-rise, rise and fall. When Tone 3 was realised as a fall-rise, it was assumed to have two targets to approach, first the low target and then the high target. In this case, the maximum pitch was obtained on the right side of the minimum pitch. When Tone 3 was realised as a fall, it was assumed to have a low target to approach, and its minimum pitch was obtained at the syllable offset. When Tone 3 was realised as a rise, it was assumed to have a high target to approach, and its maximum pitch was obtained at the syllable offset.

In this paper, we concentrated on the sentence-medial verbs. The verbs were on-focus in the NF-m condition, pre-focus in the NF-f condition and post-focus in the NF-i condition and were thus ideal for direct comparisons between focus and pre-/post-focus. The duration and pitch range (the difference between the maximum pitch and the minimum pitch) of the verbs were calculated and analysed as dependent variables.

To address the first research question, namely, how focus differs from non-focus in child Mandarin, we compared the prosody of the verbs in the NF-m condition (focused) with that in the NF-i (post-focus) and NF-f (pre-focus) conditions. To address the question about size of focused constituent, we compared the prosody of the verbs in the NF-m and CF-m combined narrow focus condition with that of the BF condition. To address the question on contrastivity, we compared the prosody of the verbs in the NF-m (non-contrastive narrow focus) condition with that in the CF-m (contrastive narrow focus) condition.

3. Statistical analysis and results

Mixed-effect modeling was used to assess the effect of fixed factors and the effect of interactions between the fixed factors and the other fixed factors on the dependent variables, i.e.

duration and pitch range of the verbs. There are two kinds of fixed factors: those related to focus directly and the others. The focus-related fixed factors were FOCUS (focus vs. non-focus), SIZE (narrow focus vs. broad focus), and CONTRASTIVITY (contrastive focus vs. non-contrastive focus). The other fixed factors were AGE (four-year-olds and eight-year-olds), TONE OF VERB (four tones for verbs), and TYPE OF VERB (type1 and type2). The random factor was SPEAKER. In the analyses on the effect of the fixed factors, two models were built for each fixed factor, one with only the random factor, and one with both the random factor and the fixed factor. The two models were then compared to each other. A statistically significant difference between these two models indicated a main effect of the focus-related factor. We then looked at the interaction between the focus-related fixed factor and the other fixed factors.

3.1 Focus vs. non-focus

3.1.1 Duration

Regarding the comparison between focus (verbs in NF-m) vs. post-focus (verbs in NF-i), the mixed-effect modelling showed that the main effect of FOCUS was significant ($p < 0.05$). As can be seen in Figure 1, the focused verbs in the NF-m condition were longer than the same verbs in the NF-i (post-focus) condition. There was also significant main effects of TONE OF VERB ($p < 0.05$) and TYEP OF VERB ($p = 0.01$), but no significant main effect of AGE ($p = 0.4$). No significant interaction was found between FOCUS and AGE ($p = 0.80$) or between FOCUS and TONE OF VERB ($p = 0.33$), but was found between FOCUS and TYPE OF VERB ($p < 0.05$). We then used subsets of data to look at the effect of FOCUS within the type 1 verbs and the type 2 verbs separately, and found that the main effect of focus was significant for both the type 1 verbs ($p < 0.05$) and the type 2 verbs ($p < 0.05$). This suggested that the interaction was caused by a gradient difference between type 1 verbs and type 2 verbs. The durational difference between focus and post-focus was larger in the type 2 verbs (0.08s) than in the type 1 verbs (0.04s).

Regarding the comparison between focus and pre-focus (verbs in NF-f), the mixed-effect modelling showed that the main effect of FOCUS was significant ($p < 0.05$). There was also significant main effects of TONE OF VERB ($p = 0.01$) and TYPE OF VERB ($p < 0.05$), but no significant main effect of AGE ($p = 0.44$). No significant interaction was found between FOCUS and AGE ($p = 0.82$), between FOCUS and TONE OF VERB ($p = 0.08$), or between FOCUS and TYPE OF VERB ($p = 0.14$).

The above results indicated that the children realized the focused verbs with a longer duration than the post-focal and pre-focal verbs, regardless the tones or the types of the verbs.

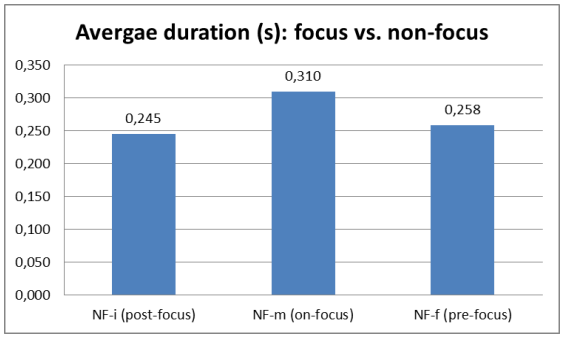


Figure 1: Verb duration in NF-i, NF-m, and NF-f

3.1.2 Pitch range

Regarding the comparison between focus and post-focus, the mixed-effect modelling showed that the main effect of FOCUS was significant ($p < 0.05$). The focused verbs in the NF-m condition had a wider pitch range (84Hz) than the same verbs in the NF-i (post-focus) condition (59Hz). The main effects of TONE OF VERB ($p < 0.05$) was also significant, but the main effect of AGE ($p = 0.53$) and TYPE OF VERB ($p = 0.1$) was not significant. A significant interaction was found between FOCUS and AGE ($p < 0.05$), and between FOCUS and TONE OF VERB ($p < 0.05$), but not between FOCUS and TYPE OF VERB ($p = 0.75$). We looked at the effect of FOCUS within each age, and found that the main effect of FOCUS was significant for the eight-year-olds ($p < 0.05$), but was not significant for the four-year-olds ($p = 0.72$) (Figure 2), so the eight-year-olds used pitch range to differentiate focus from post-focus, but the four-year-olds didn't. We then looked at the effect of FOCUS within each tone, and found that the main effect of FOCUS was significant for Tone 2 ($p < 0.05$) and Tone 4 ($p < 0.05$), but was not significant for Tone 1 ($p = 0.6$) or Tone 3 ($p = 0.28$).

Comparing focus with pre-focus, the mixed-effect modelling showed that the main effect of FOCUS was significant ($p < 0.05$). The focused verbs in the NF-m condition had a wider pitch range (84Hz) than the same verbs in the NF-f (pre-focus) condition (57Hz). The main effect of TONE OF VERB ($p < 0.05$) was also significant, but the main effects of AGE ($p = 0.17$) and TYPE OF VERB ($p = 0.43$) were not significant. A significant interaction was found between FOCUS and TONE OF VERB ($p < 0.05$), but not between FOCUS and AGE ($p = 0.22$) or between FOCUS and TYPE OF VERB ($p = 0.43$). We looked at the effect of FOCUS within each tone using subsets of data. It was found that the main effect of FOCUS was significant for Tone 2 ($p < 0.05$) and Tone 4 ($p < 0.05$), but was not significant for Tone 1 ($p = 0.5$) or Tone 3 ($p = 0.35$).

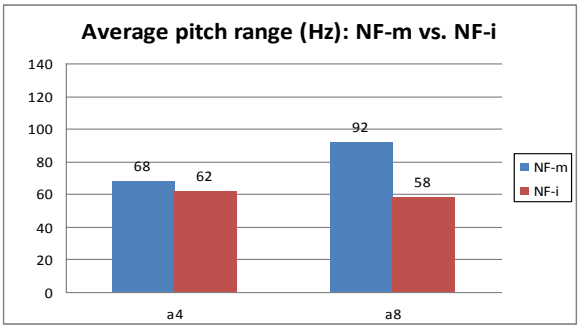


Figure 2: Pitch range in NF-m and NF-i for each age

The above results indicated that both the four- and eight-year-olds could use pitch range to differentiate focus from pre-focus, but only the eight-year-olds could use pitch range to differentiate focus from post-focus. In addition, without looking into each age group, we found that the children as a whole group used a wider pitch range to differentiate focus from post-focus and from pre-focus when the verbs were in Tone 2 and Tone 4, but not in Tone 1 and Tone 3. The four-year-olds' not using pitch range to distinguish focus from post-focus might be caused by their failure to use pitch range in tone 1- and tone-3-verbs.

3.2 Narrow focus vs. broad focus

To examine the realisation of narrow focus compared to that of broad focus, we grouped NF-m and CF-m together as a

combined narrow focus condition (hereafter, the "NF-m&CF-m" condition) with a small focal size, and compared it with the BF condition with a larger focal size. Mixed-effect modeling was adopted and the focus-related fixed factor was SIZE.

3.2.1 Duration

Comparing narrow focus with broad focus, the mixed-effect modelling showed that the main effect of SIZE was not significant ($p = 0.22$). In other words, children did not use duration to differentiate narrow focus from broad focus.

3.2.2 Pitch range

Regarding the comparison between narrow focus and broad focus, the mixed effect modelling showed that the main effect of SIZE was significant ($p < 0.05$). Figure 3 showed that the pitch range of the verbs in the NF-m&CF-m condition was larger than that in the BF condition. The main effect of TONE OF VERB ($p < 0.05$) was also significant, but the main effects of AGE ($p = 0.05$) and TYPE OF VERB ($p = 0.21$) were not significant. No significant interaction was found between FOCUS and AGE ($p = 0.35$), between FOCUS and TONE OF VERB ($p = 0.15$), or between FOCUS and TYPE OF VERB ($p = 0.05$). The results revealed that the four- and eight-year-olds used pitch range to differentiate narrow focus from broad focus, regardless of tones and types of verbs.

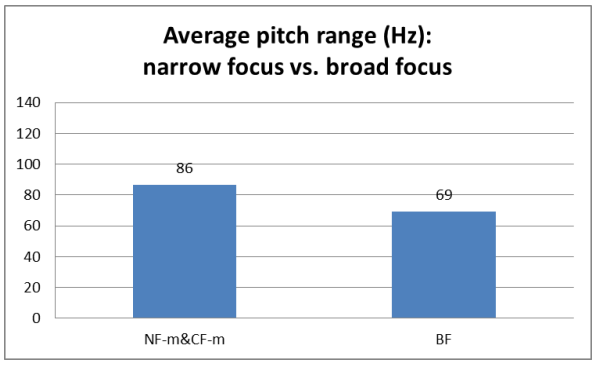


Figure 3: Pitch range of the verbs in NF-m&CF-m and BF

To sum up, children used pitch range but not duration to distinguish narrow focus from broad focus.

3.3 Contrastive focus vs. non-contrastive focus

Regarding the comparison between contrastive and non-contrastive focus, the mixed-effect modeling showed that the main effects of CONTRASTIVITY were not significant for duration ($p = 0.69$) or for pitch range ($p = 0.37$), indicating that the children did not use duration or pitch range to distinguish contrastive focus (CF-m) from non-contrastive (NF-m) focus.

4. Discussion and Conclusions

This study aimed at finding out how Mandarin-speaking children use pitch range and duration to mark focus in comparison with non-focus, how they encode focus with different constituent size, and how they differentiate contrastive focus from non-contrastive focus. With regard to focus, the children from both age groups produced the focused words with a longer duration than the non-focused ones. Further, both the four- and eight-year-olds used a wider pitch range for the focused verbs than for the pre-focal ones, but only the eight-year-olds used a wider pitch range for the focused verbs than for the post-focal ones. In addition, the children as a whole group used pitch range to differentiate focus from post-focus and pre-focus for the Tone 2 and Tone 4 verbs, but not for the Tone 1 and Tone 3 verbs. With regard to the size of the focal constituent, the children used pitch range but not duration to differentiate narrow focus from broad focus. With regard to contrastivity, children did not differentiate contrastive narrow focus from non-contrastive narrow focus using duration or pitch, similar to the findings on adult Mandarin [9].

The results had four implications. First, in previous studies on prosodic focus marking, the non-focus condition varies from study to study. Our results show that the defintion of the non-focus condition can influence the results on the use of prosody in distinguishing focus from non-focus in children. Second, to differentiate focus from non-focus, the children used duration regardless of lexical tone but used pitch range only in Tone 2- and Tone 4-verbs, while to differentiate narrow focus from broad focus they used only pitch range. As such selective uses of duration and pitch range have not been observed in adult Mandarin, these results may suggest that the children have not consolidated the use of pitch or duration. Third, as has been mentioned, to become adult like, Mandarin-speaking children have to acquire both the lexical and post-lexical functions of pitch. Previous studies on the acquisition of Mandarin tones showed that the production of Tone 4 is most adult-like in the production of Chinese 3-year-olds, followed by Tone 1, Tone 2, and Tone 3 [17]. However, in terms of focus-marking, we found that the children used pitch range to mark focus only when the verbs were in Tone 2 and Tone 4 but not in Tone 1 or Tone 3. These indicated that the acquisition of pitch range in focus marking may not be related to the order of tonal acquisition. However, To explicate children's use of pitch, not only pitch range but also the maximum and minimum pitch should be analyzed. Last, corss-linguistically, comparing to Dutch-speaking children, who have not acquired the use of duration in focus-marking at the age of seven or eight [14], the Chinese children acquired the use of duration for focus-marking quite early. Besides, the use of pitch range was in place in the Chinese four-year-olds, though not necessarily in all conditions. This suggested an earlier acquisition of pitch range as well in the Chinese children.

5. Acknowledgements

This study is supported by a VIDI grant (276-89-001) from the NWO (Netherlands Organisation for Scientific Research) to the second author. We would like to express our special gratitude to Min Zhu, Jun Bian, Yian Liang, and Shushuang Yu from Beijing 21st Century International Kindergarten and School, the children and their parents for their full cooperation. We would also like to thank Hua Shu from Beijing Normal University for her support, and Mei Ou, Mengting Huang, Yun Li, and Xingzhi Yao from Beijing Forestry University for administering the tests with the adults. We thank Paula Cox for drawing the pictures, Frank Bijlsma, Sjef Pieters, and Alex Manus for the technical support, and Mattis van den Bergh for statistical support. Last, we thank Xiaoli Dong, Mengru Han, René Kager, Zenghui Liu, Anna Sara Romøren and Wim Zonneveld for their input.

6. References

[1] Lambrecht, K. (1994). *Information structure and sentence form: Topics, focus, and the representations of discourse referents.* Cambridge: Cambridge University Press.

[2] Gundel, J. K. (1999). "On different kinds of focus". In P. Bosch, & R. van de Sandt (eds.) *Focus: linguistic, cognitive, and computational*. Cambridge: Cambridge University Press.
[3] Shih, C. (1988). Tone and intonation in mandarin. *Working Papers of the Cornell Phonetics Laboratory*, 3, 83-109.
[4] Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of Phonetics*, 27(1), 55-105.
[5] Chen, Y., & Braun, B. (2006). Prosodic realization of information structure categories in standard Chinese. Paper presented at the *Proceedings of Speech Prosody*, Dresden, Germany.
[6] Shih, C. (2000). A declination model of mandarin Chinese. *Intonation: Analysis, Modelling and Technology*, 243-268.
[7] Chen, Y. (2010). Post-focus F0 compression—Now you see it, now you don't. *Journal of Phonetics*, 38(4), 517-525. doi:10.1016/j.wocn.2010.06.004
[8] Xu, Y. (2011). Post-focus compression: Cross-linguistic distribution and historical origin. Paper presented at the *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong. pp. 152-155.
[9] Greif, M. (2010). Contrastive focus in mandarin Chinese. Paper presented at the *Proc. Speech Prosody*, Chicago, UAS.
[10] Chen, A. (2011a). Tuning information packaging: intonational realization of topic and focus in child Dutch. *J. Child Lang*, *38*, 1055-1083.
[11] Hornby, P. A. and Hass, W. A. (1970). Use of contrastive stress by preschool children. *J. of Speech and Hearing Research*, 13, 359-399.
[12] MacWhinney, B. and Bates, E. (1978). Sentential devices for conveying givenness and newness: A cross- cultural developmental study. *Journal of verbal learning and verbal behavior*, 17, 539-558.
[13] Chen, A. (2011b). The developmental path to phonological focus-marking in Dutch. In *Prosodic Categories: Production, Perception and Comprehension* (pp. 93-109). Springer Netherlands.
[14] Chen A. (2009). The phonetics of sentence-initial topic and focus in adult and child Dutch. In M. Vigário, S. Frota and M. J. Freitas (eds.), *Phonetics and Phonology: Interactions and interrelations* (pp. 91-106). Amsterdam: Benjamins.
[15] Li, P., Zhao, X., Liu, Y. and Levine, S. *Chinese Single-character Word Database*. [http://www.personal.psu.edu/pul8/psylin_norm/psychnorms.html]
[16] Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33: 319-337.
[17] Wong, P. (2012). Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *Journal of Phonetics*, *40*(1), 141-151

Prosodic effects on vowel spectra in three Australian languages

Simone Graetzer, Janet Fletcher, John Hajek

Phonetics Laboratory, School of Languages and Linguistics, University of Melbourne

{simone.graetzer, j.fletcher, j.hajek}@unimelb.edu.au

Abstract

In this paper, the spectral properties of vowels in three Australian languages are examined with the aim of determining whether prosodic prominence and domain-edge effects on formant frequencies, formant variability and vowel space dispersion can be identified. It is shown that these vowel systems are sufficiently dispersed, with an anchoring of the system by the open central vowel. It is also shown that for Burarra but not for Gupapuyngu or Warlpiri there is some evidence of prosodically-driven hyper-articulation. Finally, the data indicate pre-boundary lengthening in all three languages, which in some cases appears to be associated with changes in vowel quality.

Index Terms: vowel, vowel space, variability, acoustics, dispersion, duration, Australia

1. Introduction

The focus of this paper is on vowel spectra and the systemic variability and dispersion of vowels in disyllabic words in three Australian languages: Burarra, Gupapuyngu and Warlpiri. The paper addresses the interaction and effects of factors such as vowel quality, and prosodic prominence and position in the prosodic domain (word-final or pre-boundary, and word-medial).

1.1. Adaptive Dispersion Theory

Adaptive Dispersion Theory (DT) proposes that vowel contrasts are systemic and relational rather than local and absolute [1]. Each vowel acts as a repeller in a dynamic system. As such, it was originally claimed that adjacent vowels should be roughly equidistant in a system, *i.e.*, (perceptual) vowel contrasts should be maximal, and the acoustic distance between adjacent vowels should reduce as the system size increases. DT also predicted that languages with the same vowel system will exhibit identical F1 distances between adjacent vowels. However, the maximal contrast hypothesis was later modified by Lindblom [1] to allow for the possibility of merely sufficient dispersion, emphasizing sufficient contrast combined with economy of effort. A number of previous analyses of Australian languages, including Burarra and Warlpiri, have indeed found that these vowel spaces are sufficiently, rather than maximally, dispersed [2] [3].

1.2. Prosodic effects on vowel realisation

In commonly studied languages such as English, prosodically prominent vowels are typically less variable and more peripheral in the F2 x F1 space than prosodically weak vowels. Vowels at higher levels of prosodic prominence tend to be produced with greater acoustic expansion or peripherality or with more extreme gestures [4] - [6]. For example, in English, accented /i/ tends to be produced with a more anterior constriction than unaccented /i/ [5], while in French, the stronger the prosodic boundary associated with the vowel /a/, the higher the F1 frequency [7]. Furthermore, prosodic domain-initial (e.g., phrase-initial) positions seem to be 'generally characterised by more "forceful" articulatory gestures' [8] (p. 232). This type of behaviour has been termed prosodically driven 'hyper-articulation' [4] [9]. Typically, vowels are associated with less variability and greater dispersion when prosodically prominent, as part of an overall articulatory expansion of prominent syllables [4] [10]. Stressed vowels are also typically more resistant to coarticulation by adjacent segments than are unstressed vowels [9] [11]. Additionally, there are prosodic domain-final effects on vowels: pre-boundary or phrase-final lengthening can occur [12].

In the context of Australian languages, Burarra vowels are known to undergo some magnitude of vowel reduction in unstressed syllables [13]. In older speakers of the Central Australian language, Arrernte, the low central vowel appears to involve a higher F1 (increased openness) when stressed [14]. In other Australian languages, little evidence has been found of the predicted effect of prosodic prominence on vowels in the F2 x F1 plane, but there is some evidence of an effect of pre-boundary or phrase-final position on vowel realisation. Some vowels in some Australian languages are longer and more peripheral in phrase-final positions [3], while others are lengthened phrase-finally but do not change in quality [14] [15]. In their study of Bininj Gun-wok and Dalabon, Fletcher and Butcher [16] did not find evidence of reduced vowel peripherality in phrase-final position, perhaps due to duration-related expansion of the vowel space in this position. Similarly, Fletcher and Butcher [3] found for a female speaker of another three vowel language with a length distinction, Kayardild, that close vowels tended not to show effects of prosodic context but rather of vowel length phrase-finally (or an interaction between such prosodic effects and length). In Kunwinjku, there is an effect of prosodic prominence on variability, which is greater in unaccented vowels [17]. Finally, in Warlpiri, it appears to be the medial consonant rather than the prominent vowel that undergoes medial strengthening (and lengthening), and this consonant may carry stress [18].

1.3. Research questions and parameters

In the current paper, we set out to determine how vowels in Burarra, Gupapuyngu and Warlpiri differ in F1 and F2 in disyllabic CV1CV2 words. Secondly, we wanted to examine whether vowels differ in F1 and F2, formant variability and vowel space dispersion according to prosodic prominence and to position in the word (V1 or V2). Finally, we wanted to determine whether there is any effect of vowel position on vowel duration in the three languages. As all three vowel systems are relatively small, the issue of an effect of inventory size on variability or dispersion is not addressed here; see [19]. Of the languages considered here, Burarra is a non-Pama-Nyungan language and Gupapuyngu and Warlpiri are members of the Pama-Nyungan family, which includes most

indigenous Australian languages. Burarra and Gupapuyngu are spoken in central and north-eastern Arnhem Land, respectively, and Warlpiri is spoken to the north-west of Alice Springs. These languages have many place and few manner consonantal contrasts and small vowel inventories. Gupapuyngu and Warlpiri have three vowels with a length contrast, while Burarra has five vowels: /i/, /ɛ/, /ɐ/ (hereafter 'a'), /o/ and /u/.

2. Methods

The subjects of this study were nine adult female speakers of Burarra (speakers DP, KF, MW), Gupapuyngu (AM, BT, EG) and Warlpiri (BP, KR, RR) aged between 30 and 65. Typically, three tokens of each word type were elicited. The corpus was collected and digitised by Andrew Butcher. To avoid effects of gender, we limited our analysis to the data available for female speakers in the corpus. All acoustic measurements were carried out in the EMU Speech Database System [20]. The onsets of the vowels were marked at the onset of periodicity and the offsets at the offset of periodicity. F1 and F2 values were measured at the vowel midpoint. All relevant segmentation procedures are described in full by Graetzer [19]. The prosodic prominence of vowels was determined on the basis of published prosodic descriptions, e.g., [13] [18], an auditory impressionistic analysis and an acoustical analysis; we assumed that the vowel carrying a sharp F0 rise to a peak somewhere in or around the syllable rhyme was accentually prominent. Main stress was word-initial in these data. Words were associated with post-lexical (or phrasal) prosodic prominence; therefore, prosodic effects applying to both the utterance/phrase level and the word level were relevant. Statistical and graphical procedures were run in R version 2.14.0 [21] using the EMU-R package. F1 and F2 frequencies (Hz) were extracted from the vowel midpoint ($V1_{MID}$ and $V2_{MID}$) in CV1CV2 words, in which V1 is prominent and V2 is not. In order to compare vowel systems in the three languages, formant frequencies were submitted to the normalisation procedure used in similar studies in the context of vowels [22]: Nearey vowel-extrinsic normalisation [23]. This procedure was conducted on the raw F1 and F2 data by means of a formula and function outlined by Harrington [24]. The distribution of the vowel categories is given in Table 1 (note that no long vowels were present for Warlpiri; however, short and long Warlpiri vowel counterparts have been found to be similar in quality [25]; note also differences in Burarra distribution between V1 and V2). Linear Mixed Model (LMM) procedures were used for investigations of F1 and F2 frequencies, in which the effects of vowel quality and word-medial consonant place of articulation (not discussed here), language group and vowel position (V1/V2; random factor: speaker) were examined. The non-normalised F1 and F2 LMM results were then compared to the normalised results and any differences in significance levels were reported.

Euclidean distances (here termed 'distances') - acoustic straight line distances between vowels in the F2 x F1 plane - were calculated after, e.g., [1] [24], as a measure of the magnitude of hyper-articulation in the vowel space. This procedure is similar to that used by Recasens and Espinosa [22]. The distances were then treated as the dependent variable in an LMM, with the fixed factors of language group and vowel position (random factor: speaker). Levene-type t-tests applied to deviations of observations from the median were run per speaker and formant (F1 and F2) to test for equality of variance between V1 and V2 for Gupapuyngu and Warlpiri, which have three vowel qualities in V1 and V2 position, but not for Burarra, as it has five vowel qualities in V1 position and only three in V2. Wilcoxon signed rank tests for paired samples were calculated for V1 and V2 durations per speaker in order to determine whether there was lengthening of V1 or V2 (α=0.05). Pearson's product moment correlation tests were run for vowel duration and distances, *i.e.*, vowel space expansion, collapsing V1 and V2 categories. Raw and normalised results were not compared in the context of the Levene's and Wilcoxon tests as these were intra-speaker.

Table 1. *The distribution of vowels where BUR = Burarra, GUP = Gupapuyngu, WAR = Warlpiri.*

	V	BUR	GUP	WAR
V1	a	167	152	219
	a:	N/A	107	N/A
	ɛ	36	N/A	N/A
	i	66	48	102
	i:	N/A	50	N/A
	o	84	N/A	N/A
	u	97	110	124
	u:	N/A	87	N/A
V2	a	422	281	199
	i	23	108	110
	u	5	165	136

3. Results

3.1. Normalised and raw F1 and F2 frequencies, variability and euclidean distances

3.1.1. Normalised frequencies

In Figure 1, 95% confidence ellipses in Nearey-normalised formant data extracted at vowel midpoints are shown per language group and vowel position (V1, L; V2, R). The typically close proximity and overlap of the vowel ellipses indicates sufficient rather than maximal dispersion. For Gupapuyngu and Warlpiri, a 'canonical' realisation of a triangular vowel space is observed, with the low central vowel as 'anchor' (see, e.g., Butcher's [2] use of this term). Gupapuyngu displayed greater dispersion than Burarra and Warlpiri in both vowel positions. For Burarra, /a/ acted as anchor in V2, while /ɛ/ and /a/ were equally close to the grand centroid ('X') in V1. In the F2 x F1 space, in both V1 and V2, the difference between Gupapuyngu and Warlpiri mainly involved an upwards shift but also a retraction of /i a/ in Warlpiri relative to Gupapuyngu. In Warlpiri, /i u/ were realised with a lower F1 (increased closeness) than in the other languages. Within each vowel system, point vowels were similarly spaced relative to one another overall. However, when frequency distances between point vowels were compared for Gupapuyngu and Warlpiri, which have identical vowel systems, in F1, in V1 and V2, /a/ was closer to /i u/ in Warlpiri than in Gupapuyngu (there was a relative reduction in the F1 range for Warlpiri), while /i u/ were similarly distant. In F2 in V1, in Gupapuyngu, /a/ was further from /i/ and closer to /u/ than in Warlpiri. F2 distances in V1 were similar in the two languages, but slightly higher (more anterior) in Gupapuyngu. Burarra vowels /ɛ o/ were approximately equidistant between

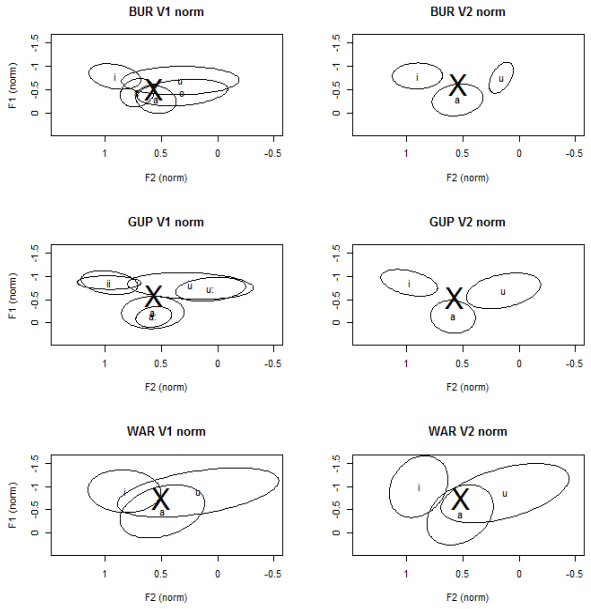


Figure 1: *Nearey-normalised F2 x F1 plots in V1 (L) and V2 (R) conditions (upper: Burarra; middle: Gupapuyngu; lower: Warlpiri); axes in Nearey units.*

the open and close vowels and were very similar in height. Burarra /o u/ were very similar in F2, while /ɛ/ was much lower in F2 (less anterior) than /i/. In V2, Burarra differed from Gupapuyngu and Warlpiri in that /i/ was associated with a similar F1 to /u/, while for the other languages /i/ was associated with a much lower F1 than /u/, *i.e.*, /i/ was a closer vowel than /u/ in these systems. It is clear that /u/ can be realised with little or no rounding given the degree of F2 variability in /u/ and its realisation relative to /i a/ in the vowel space. This is the case particularly for Burarra and Gupapuyngu in V1, in which position there were more /u(:)/ tokens in the corpus.

3.1.2. Raw frequencies - LMM

LMM procedures were used to investigate the effect of vowel position within the word, and thus also prosodic prominence, on F1 and F2 formant frequencies at vowel midpoints in CV1CV2 words. The factor of language group was also included in the model. In the F1 condition, Warlpiri was associated with lower frequencies than Burarra (BUR:WAR, Estimate=-57, SE=28, t=-2, $p<0.05$). Other comparisons were not significant. Vowels in V1 position were typically associated with slightly lower F1 values than vowels in V2 (V2:V1, Estimate=-23, SE=4, t=-5.8, $p<0.0001$). This difference was located in the Gupapuyngu and Burarra data. For Gupapuyngu this pattern appears to be due to a slightly lower F1 (more close vowel) in V1 /a u/, while in Burarra it appears to be due to a lower F1 in V1 /a/. For Burarra it cannot be excluded that this result is due to differing vowel distribution in V1 and V2. Tukey's post-hoc tests indicated that Gupapuyngu, like Warlpiri, was associated with lower F1 frequencies (more close vowels) than Burarra (z=-3.6, $p<0.001$). In a LMM analysis of F2, there was no effect of language group or vowel position. These F1 and F2 results were consistent with those for the normalised data.

3.1.3. Variability (raw frequencies)

Across all vowels, variability tended to be higher in magnitude in F2 - presumably because of coarticulatory effects exerted by surrounding consonants - than in F1, but particularly for /i u/ (Figure 1). Typically, close front vowels varied less than the other point vowels in F1. This was not the case in F2, in which /a/ varied least and /u/ tended to vary most. With the exception of Burarra (but see Table 1), for each point vowel, F1 variability tended to be higher in V2 and F2 variability tended to be slightly higher in V1. According to Levene's tests for Gupapuyngu and Warlpiri, in F1, only one speaker of each language group produced a difference in overall variance between V1 and V2 conditions: EG and KR respectively (GUP EG, $F(1,194)=7.2$, $p<0.01$; WAR KR, $F(1,342)=33.9$, $p<0.0001$). In F2, only one speaker of Warlpiri, BP, produced a difference between conditions ($F(1,232)=9$, $p<0.005$). In the three cases, V1 was associated with less variance than V2, indicating that more variability existed in the prosodically weak, word-final vowel, as would be predicted.

3.1.4. Euclidean distances (raw frequencies)

Across the language groups, as indicated by the distance of vowels from the grand centroid in Figure 1, /a/ tended to be associated with smaller distances of 200 to 300Hz, while /i/ was associated with larger distances of 520 to 775Hz, and /u/ was associated with intermediate to relatively large distances of 440 to 690Hz. In Gupapuyngu, long vowels tended to be more peripheral than short vowel counterparts. For Burarra speakers DP and MW, significantly greater dispersion occurred in V1 than V2 when all vowel qualities were included in V1 (DP, V=8844, $p<0.005$; MW, V=20133, $p<0.05$). The same pattern did not achieve significance for speaker KF (V=2928, p=0.17), perhaps because there were no V2 /u/ tokens for that Burarra speaker. When vowel qualities were reduced to /i a u/ only, to facilitate comparison between V1 and V2, similar results were obtained (DP, W=13806, $p<0.0001$; KF, W=3384, p=0.23; MW, W=13830, $p<0.05$). This pattern of greater expansion in V1 was mainly located in /i u/. When Gupapuyngu short vowels only were considered, distances overall did not differ according to vowel position (AM, V=10414, p=0.4; BT, V=13259, p=0.11; EG, V=2449, p=0.94), which indicated a lack of vowel space expansion in V1 relative to V2. With regard to the Warlpiri speakers, there was greater dispersion in V2 for at least two speakers (BP, V=2200, $p<0.005$; KR, V=5896, $p<0.05$), and this pattern was particularly evident in /i u/, but the difference did not achieve significance for the third speaker, RR (V=5646, p=0.4). In order to compare overall dispersion across languages, an LMM was run with the fixed factors of language group and vowel position with interactions. All main and interaction effects were significant at $p<0.005$ or below. Vowel positions differed strongly at $p<0.0001$ (V2:V1, Estimate=78, SE=37, t=8.3); V1 was typically associated with larger distances; however, this difference was in fact primarily located in Burarra. Gupapuyngu was associated with most dispersion and Burarra with least, while Warlpiri dispersion was intermediate (BUR:GUP, Estimate=191, SE=52, t=3.7, $p<0.0005$; BUR:WAR, Estimate=167, SE=52, t=3.2, $p<0.005$). These results were again consistent with those for the normalised data.

3.2. Vowel duration

Typically in Burarra, /a/ was longer in duration than /i/ and /u/ by >30ms. In Warlpiri, there was a trend in the same direction but in Gupapuyngu, there was little difference overall between vowels. For the majority of speakers, vowels /a i u/ tended to be longer in duration in V2 position (by 40ms in Burarra and Warlpiri and by 90ms in Gupapuyngu, on average), indicating phrase-final or pre-boundary lengthening. For Burarra speakers DP and MW, vowel duration was greater in V2 (DP, V=2310, $p<0.0001$; MW, V=1454, $p<0.0001$). For KF, V1 and V2 did not vary (V=2161, $p=0.44$), but recall that there were no KF /u/ tokens in V2. For all Gupapuyngu and Warlpiri speakers, durations were consistently greater in V2 at $p<0.0001$. The Gupapuyngu short/long vowel distinction in V1 tended to show a 1:1.5 – 1:1.7 ratio, in which the long vowel was 1.5 to 1.7 times the length of its short counterpart. Correlations of vowel duration and distances per speaker (collapsing vowel positions) were performed to test whether expansion related meaningfully to phrase-final lengthening. Correlations were in most cases low or very low at $r < 0.25$. However, four of the nine /i/ tests resulted in low to moderate correlations of >0.4, and two of the nine /u/ tests resulted in low correlations of 0.3. That is, under certain conditions for some speakers, there was a weak positive correlation between vowel space expansion and vowel duration; such correlations were more likely to occur for /i/ than /a u/.

Table 2. *Summary of vowel position results*

Procedure	V1		V2	Comment
LMM F1	V1	<	V2	GUP & BUR
LMM F2	V1	=	V2	No effects
Variability F1	V1	=<	V2	GUP & WAR
Variability F2	V1	=<	V2	GUP & WAR
Distances BUR	V1	>	V2	Esp. in /i u/
Distances GUP	V1	=	V2	No effects
Distances WAR	V1	<	V2	Esp. BP & KR; /i u/
LMM Distances	V1	>	V2	Located in BUR
Duration	V1	<	V2	Except BUR KF

4. Discussion and conclusions

Our first research question concerned how vowels in these languages differ in F1 and F2 in disyllabic CV1CV2 words. Clear evidence was provided of sufficient rather than maximal dispersion in Burarra, Gupapuyngu and Warlpiri vowel spaces, as has been found previously for Burarra [2] and other Australian languages such as Kunwinjku and Dalabon [17]. Typically, for the three languages considered here, /a/ acted as vowel space anchor or pivot. Some differences between the languages' vowel spaces became apparent. The Warlpiri normalised vowel space was found to be slightly more compact and also slightly lower in F1 (more close) than those of Gupapuyngu and especially Burarra. In Burarra, the close vowels tended to be similar in F1, but in Gupapuyngu and Warlpiri, /i/ was lower in F1 than /u/. When F1 distances between vowels were compared in Gupapuyngu and Warlpiri, which have identical vowel systems, F1 distances tended to be slightly larger in Gupapuyngu. These results disconfirm the strong version of the DT hypothesis that languages with the same vowel system will exhibit identical F1 distances between adjacent vowels. In general, /i/ tended to vary least in F1, indicating strong articulatory requirements associated with tongue dorsum raising and fronting and bracing against the hard palate [22] [26]. The finding of high variability in /u/ in F2, which is typically associated with this vowel [22] [26], suggests some fronting and lip unrounding. /a/ tended to vary in both F1 (height) and F2 (anteriority). The large amount of F2 variability and overlap in these vowel spaces can be related to the 'place of articulation imperative' [13]. This constraint requires that, given few vowel contrasts and many (coronal) consonant place constrasts, cues to place are prioritised over vowel quality cues in Australian languages. The role of consonant place in vowel variability and dispersion in these languages will be investigated in future research.

The second research question was whether vowels differ according to prosodic prominence and position in the word (V1 or V2) in F1 and F2, formant variability and vowel space dispersion. Table 2 presents a summary of the relevant findings. Across languages, phrase-final vowels (V2) tended to be slightly higher in F1 (more open), but this pattern in F1 was mainly located in the Burarra data in /a/, and in the Gupapuyngu data in /a u/. There were no effects in F2. Some evidence was presented for some Gupapuyngu and Warlpiri speakers of greater F1 and F2 variability in the word-final vowel, consistent with findings for Kunwinjku [17]. Given the finding of increased duration in V2 across languages, it is likely that any increased variability in V2 is due to phrase-final lengthening rather than the absence of accentuation. Regarding dispersion, typically, for the Burarra speakers, greater dispersion occurred in the prosodically prominent vowel, consistent with previous claims that Burarra exhibits vowel reduction in unstressed syllables, *i.e.*, V2 in this experiment [13]. For two Warlpiri speakers, greater dispersion occurred in the phrase-final, lengthened, vowel. We found in Gupapuyngu that /a u/ were slightly more open in V2 than in V1, indicating increased jaw opening and sonority, which is likely to be the result of lengthening. However, the effect of vowel position on overall dispersion in Gupapuyngu did not reach significance for any speaker. In general, dispersion in the vowel space was not greater in Burarra than in Gupapuyngu in V1 (or V2), despite the former language having a larger inventory size. In fact, Gupapuyngu was associated with the largest magnitude of dispersion overall. The finding that a language with a larger inventory does not necessarily show a larger magnitude of dispersion is consistent with previous findings in European languages, e.g., [22]. Regarding the question of whether there is any effect of vowel position on vowel duration, as mentioned, our results confirm that there is pre-boundary lengthening in Burarra, Warlpiri and Gupapuyngu, in accordance with observations for other Australian languages, e.g., [14]. Typically, vowel duration did not correlate with expansion. However, for some speakers, /i/ showed a weak positive correlation. Further research is needed to separate the effects of position in word and prosodic prominence and to comprehensively investigate prominence effects on word-medial consonants in these languages.

5. Acknowledgements

We would like to thank Andrew Butcher for providing the corpus. We would also like to thank the speakers involved. This research was funded by an Australian Postgraduate Award granted to the first author.

6. References

[1] Lindblom, B., "Phonetic universals in vowel systems", in J. J. Ohala, and J. J. Jaeger [Eds.], Experimental Phonology, 13-44, Orlando, Academic Press, 1986.

[2] Butcher, A. R., "On the phonetics of small vowel systems: evidence from Australian languages", in R. Togneri [Ed.], Proceedings of the 5th Australian International Conference on Speech Science and Technology, 28-33, Canberra, Australian Speech Science and Technology Association, 1994.

[3] Fletcher, J. and Butcher, A., "Local and global influences on vowel formants in three Australian languages", in D. Recasens, M. -J. Solé, and J. Romero [Eds.], Proceedings of the 15th International Congress of Phonetic Sciences, 905-908, Barcelona, 2003.

[4] de Jong, K., "The supraglottal articulation of prominence in English: linguistic stress as localized hyper-articulation", Journal of the Acoustical Society of America, 97(1):491-504, 1995.

[5] Cho, T., "Prosodic strengthening and featural enhancement: evidence from acoustic and articulatory realisations of /a u/ in English", Journal of the Acoustical Society of America, 117(6):2867-2878, 2005

[6] Mooshammer, C. and Fuchs, S., "Stress distinction in German: Simulating kinematic parameters of tongue tip gestures", Journal of Phonetics, 30(3):337-355, 2002.

[7] Tabain, M., "Effects of prosodic boundary on /aC/ sequences: Acoustic results", Journal of the Acoustical Society of America, 113(1):516-531, 2003.

[8] Fujimura, O., "Methods and goals of speech production research", Language and Speech, 33(3):195-258, 1990.

[9] Cho, T., "Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English", Journal of Phonetics, 32(2):141-176, 2004.

[10] Beckman, M. E., Edwards, J. and Fletcher, J., "Prosodic structure and tempo in a sonority model of articulatory dynamics", in G. J. Docherty and D. R. Ladd [Eds.], Papers in Laboratory Phonology II, 68-86, Cambridge University Press, 1992.

[11] Bombien, L., Mooshammer, C., Hoole, P. and Kühnert, B., "Prosodic and segmental effects on EPG contact patterns of word-initial German clusters", Journal of Phonetics, 38:388-403, 2010.

[12] Fletcher, J. and Vatikiotis-Bateson, E., "Prosody and intra-syllabic timing in French", in Proceedings of the 3rd Australian Speech Science and Technology Conference, 318-323, 1990.

[13] Butcher, A. R., "Australian Aboriginal languages: consonant salient phonologies and the 'place-of-articulation imperative", in J. M. Harrington and M. Tabain [Eds], Speech Production: Models, Phonetic Processes and Techniques, 187-210, Psychology Press, 2006.

[14] Tabain, M. and Breen, G., "Central vowels in Central Arrernte: a spectrographic study of a small vowel system", Journal of Phonetics, 39(1):68-84, 2011.

[15] Bishop, J., "'Stress Accent' without Phonetic Stress: Accent Type and Distribution in Bininj Gun-wok", in Proceedings of the 1st International Conference of Speech Prosody, 179-192, Aix-en-Provence, 2002.

[16] Fletcher, J. and Butcher, A. R., "Vowel dispersion in two northern Australian languages: Dalabon and Bininj Gun-Wok", in C. Bow [Ed.], Proceedings of the 9th International Conference on Speech Science and Technology, 343-348, Melbourne, 2002.

[17] Fletcher, J., Stoakes, H., Loakes, D. and Butcher, A., "Spectral and durational properties of vowels in Kunwinjku", in J. Trouvain and W. J. Barry [Eds.], Proceedings of the 16th International Conference of Phonetic Sciences (ICPhS), 937-940, Pirrot, 2007.

[18] Butcher, A. R. and Harrington, J. M., "An acoustic and articulatory analysis of focus and the word/morpheme boundary distinction in Warlpiri", in S. Palethorpe and M. Tabain [Eds.], Proceedings of the 6th International Seminar on Speech Production, 19-24, Sydney, 2003.

[19] Graetzer, S., An acoustic study of coarticulation: consonant-vowel and vowel-to-vowel coarticulation in four Australian languages. PhD thesis, University of Melbourne, 2012.

[20] Cassidy, S. and Harrington, J., "EMU README; the EMU speech database system". Online: http://emu.sourceforge.net/, accessed on 1 Dec 2013.

[21] R Development Core Team, R: a language and environment for statistical computing, R Foundation for Statistical Computing. Online: http://www.R-project.org, accessed on 1 Nov 2013.

[22] Recasens, D. and Espinosa, A., "Dispersion and variability in Catalan five and six peripheral vowel systems", Speech Communication, 51(3):240-258, 2009.

[23] Nearey, T., Phonetic feature systems for vowels. PhD thesis, University of Alberta. Reprinted by the Indiana University Linguistics Club, 1978.

[24] Harrington, J., The phonetic analysis of speech corpora. Chichester, John Wiley and Sons, 2010.

[25] Butcher, A. R., The sounds of Australian languages. Unpublished manuscript, 1993.

[26] Recasens, D., "Lingual coarticulation", in W. Hardcastle and N. Hewlett [Eds.], Coarticulation: theory, data and techniques, 80-104, Cambridge University Press, 1999.

Rhythmic Correspondence between Music and Speech in English Vocal Music

Xi Chen, Peggy Pik Ki Mok

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
chenxi1099@gmail.com, peggymok@cuhk.edu.hk

Abstract

This study investigates the rhythmic structures of music and speech, and the possible corresponding rhythmic patterns between the two domains in English vocal music. With fifteen English songs as samples, lexical stress of multi-syllabic words is compared with three musical dimensions—metrical stress, duration, and pitch respectively. It was found that in the chosen English songs, there is a good mapping between the metrical stress of music and the lexical stress of lyrics. In addition, the duration and the pitch patterns not only generally match the lexical stress patterns most of the time, but also serve to manifest the prominence of the primary lexical stress on one hand, and to reflect the weakness of the unstressed syllables on the other. In addition to a general good match in rhythm, this study also shows match differences within the three comparisons. Match degrees vary according to different meter patterns. Moreover, pitch takes priority over duration in their respective match with lexical stress of the lyrics. Finally, the primarily stressed syllables match duration and pitch patterns much better than the unstressed ones do.

Index Terms: Musical rhythm, Speech rhythm, English songs

1. Introduction

Widespread consensus exists among musicians and linguists that music and speech have strong parallels in many aspects. Among them, rhythm is widely acknowledged as a crucial prosodic feature in both domains, and rhythmic correspondence is an important locus of interest in this interdisciplinary field. Lerdahl & Jackendoff [1] adopted linguistic framework to compare prosodic structures of music and speech so as to investigate how the prosodic elements such as duration, pitch and intensity create structured rhythmic and melodic patterns in the two domains. Their theory of metrical structures was later developed by Todd's [2] wavelet model analyses to rhythm in music and speech. Recent empirical studies [3, 4] found that the rhythmic features of the language in a geographic location would leave imprint on its local musical instrument. And even for the different dialects of the same language, they have different rhythmic characteristics which are reflected in the local musical styles [5]. Palmer and Kelly [6] discussed the relations between linguistic prosody and musical meter in songs, and found that compound word and nuclear stress rules coincide with musical rules of metrical accent. A very recent study [7] has investigated the degree of stress-meter alignment in French vocal music and found that stronger metrical stress tends to fall on the final syllables of poly-syllabic words and mono-syllabic content words. Thus the two parallel stresses are more likely to be consistent. The author also estimated that this alignment in English vocal music might be better than in French because the lexical stress in English is less controversial than in French, but so far very few studies have systematically examined the rhythmic correspondence in English songs. The present study is designed to fill this gap.

This study investigates the underlying rhythmic correspondence between music and speech in English vocal music. Word stress is a manifestation of rhythm in speech, and there are three categories of word stress in English: primary stress, secondary stress, and unstress [8]. In music, rhythm can be realized in the variation of intensity, duration and pitch [9, 10]. First, musical accent can be presented by metrical structure. Different meter patterns are created by endowing a sequence of beats with different metrical stress, and the stronger one takes the accent [11]. In addition, previous studies showed that as a reliable cue of melodic contour, duration can also present the stress by changing the length of the notes [9] [12][13], and longer notes can emphasize the musical event [14]. Besides, pitch patterns also create stress distinctions by making the pitch of a note higher than its surrounding ones [1] [12] [15]. According to Jones' idea [15] of interval accent, a note higher than its surrounding notes is supposed to be stressed, and a note with lower pitch is supposed to be unstressed. In the present study, lexical stress in multi-syllabic words in lyrics is chosen, and is respectively compared with three dimension of musical rhythm—metrical stress, duration, and pitch. These three comparisons can reveal 1) the situation of stress-meter alignment in English vocal music; 2) whether the duration patterns and pitch patterns can manifest word stress patterns.

2. Method

2.1. Materials

Fifteen English songs were selected. They are famous folk songs, golden oldies, ballads and popular movie songs which are written in different time periods with diverse backgrounds. The styles of the songs are supposed to be simple, smooth, and traditional so as to ensure that they can be easily recognized and sung by most people. The fifteen songs cover the most common meter patterns: 2/2, 2/4, 4/4, 3/4, and 6/8, and there are three songs for each meter. The alignment analysis was based on score analysis, not the sound tracks. The scores with the assigned lyrics were found from the Internet and two university libraries in Hong Kong. For ease of reference, each song is numbered. The information of the selected songs is as following:

2/2: (1) ***Red River Valley*** (Traditional Canadian folk song)
(2)***Take Me Home, Country Roads*** (Lyrics and music by John Denver, Bill Danoff & Taffy Nivert)
(3) ***Tie a Yellow Ribbon round the Ole Oak Tree*** (Lyrics and Music by Russellbrown & Irwinlevine)

2/4: (4) ***Yankee Doodle Boy*** (From *Yankee Doodle Dandy*. Lyrics and music by George M. Cohan)
(5) ***Oh Susanna*** (Lyrics and Music by Stephen Foster)
(6) ***Do-Re-Mi*** (From *Sound of music*. Lyrics by Oscar Hammerstein II; music by Richard Rodgers)

4/4: (7) ***Sound of Silence*** (Lyrics and Music by Paul Simon)
(8) ***Yesterday once more*** (Lyrics and Music by Richard Carpenter and John Bettis)

(9) ***My Heart Will Go On*** (Lyrics by Will Jennings, Music by James Horner)

3/4: (10) ***Green Sleeves*** (English traditional folk song)
(11) ***Moon River*** (Lyrics by Jonny Mercer, music by Henry Mancini)
(12) ***My Favorite Things*** (From *Sound of Music*. Lyrics by Oscar Hammerstein II; music by Richard Rodgers)

6/8: (13) ***Silent Night*** (Lyrics by Joseph Mohr; music by Franz Gruber)
(14) ***We Are the Champions*** (Lyrics by Freddie Mercury, Music by Queen)
(15) ***Dulcinea*** (From *Man of La Mancha*. Lyrics by Joe Darion, Music by Mitch Leigh)

2.2. Procedure

This study includes three parts in which the lexical stress of the lyrics is respectively compared with metrical stress, duration patterns and pitch patterns of music. Since in English, distinguishable lexical stress cannot be found in mono-syllabic words [8], this study only focuses on di-syllabic and multi-syllabic words.

2.2.1. Match between metrical stress and lexical stress

On the score of each song, syllables of disyllabic and multi-syllabic words are circled with their aligned notes, and the categories of lexical stress of the syllables and the metrical stress of the aligned notes are identified and marked.

In the analysis, the three categories of metrical stress of music (major stress, minor stress, and unstress) are respectively represented by capital letters A, B and C, and the three categories of lexical stress in English (primary stress, secondary stress and unstress) are respectively represented by small letters a, b, and c. We follow the convention of western music system for the standard of metrical stress patterns in music [11]. And the English lexical stress patterns are based on the *Oxford Advanced Learner's Dictionary, 8th edition*, published by the Oxford University Press ELT.

In the analysis, each target is a combination of two types of stresses since it includes a syllable which takes a lexical stress, and a syllable-aligned note which is attached to a particular metrical stress. There are altogether nine possibilities of such combinations. Among them, in the cases of Aa, Bb, and Cc, the two types of stresses are consistent, so they match perfectly. In Ac and Ca, the two types of stresses are contradictory. In Ab, Ba, Bc and Cb, the match between the two types of stresses is not as perfect as that in the first group, but is still better than that in the mismatched group. In this way, the nine possibilities can be divided into three match groups—perfect match (Aa, Bb, Cc), secondary match (Ab, Ba, Bc, Cb) and mismatch (Ac, Ca).

After labeling the stress combination patterns of each target, the match degree of each category can be calculated by the following formula: P %= (N_P/ N_t) * 100%, where P stands for any match category; N_P stands for the number of syllables that belong to a particular match category; N_t is the total number of syllables of disyllabic and multi-syllabic words.

2.2.2. Duration analysis

In this section, the target notes are represented by a short underline "_". The relative duration of the two notes adjacent to the target notes is examined as well. Compared with the target note, if the duration of the adjacent note is shorter, then it is marked as "S"; if longer, then it is marked as "L"; if identical, then it is marked as "I". If there is a rest, or if there is no adjacent note (the left position of the initiate note or the right position of the last note), a short dash "-" is used. There are fifteen possibilities as follows:

S_S, S_-, -_S, S_I, I_S, S_L, L_S, I_-, I_I, -_I, L_I, I_L, L_L, L_-, -_L

A capital letter X is used to represent an arbitrary metrical stress. Xa stands for a target in which the syllable takes the primary lexical stress. Xb is a target in which the syllable takes the secondary lexical stress, and Xc is a target in which the syllable is unstressed. In the analyses of both duration and pitch patterns, it was found that the number of Xb cases is quite small (12 out of 685 for duration patterns and 10 out of 687 for pitch patterns), so we only focus on Xa and Xc in this paper.

For the group of Xa, if the two adjacent notes are shorter than the target note, then the prominence of the target note is highlighted auditorily. In a similar way, for the group of Xc, if the duration of the two adjacent notes is longer than the target note, then the target note is heard as weakened. In these two cases, the duration patterns manifest the lexical stress patterns well, and the two types of stresses are in perfect match. The match degree decreases when the adjacent notes are identical to, or worse still, are longer than the target note. The match patterns are classified in Table 1.

Table 1. *Match patterns in duration analysis.*

Duration Patterns	**S_S, -_S, S_-**	**S_I, I_S**	**L_I, I_L**	**L_L, -_L, L_-**
Xa	Perfect match	Moderate match	Poor match	Mismatch
Xc	Mismatch	Poor match	Moderate match	Perfect match

Note: Xa: target with primarily stressed syllables; Xc: target with unstressed syllables

2.2.3. Pitch analysis

In this section, the analysis is similar to that in the duration comparison. If the pitch of the adjacent note is higher than the target note, then it is marked as "H (high)", if lower, then it is marked as "L (low)", if identical, then it is marked as "I (identical)". There are fifteen possibilities of patterns:

L_L, L_-, -_L, L_I, I_L, L_H, I_-, I_I, -_I, H_L, I_H, H_I, H_H, H_-, -_H

The pitch patterns for Xa and Xc are classified as in Table 2 below:

Table 2. *Match patterns in pitch analysis.*

Pitch Patterns	**L_L, -_L, L_-**	**L_I, I_L**	**H_I, I_H**	**H_H, -_H, H_-**
Xa	Perfect match	Moderate match	Poor match	Mismatch
Xc	Mismatch	Poor match	Moderate match	Perfect match

Note: Xa: target with primarily stressed syllables; Xc: target with unstressed syllables

It should be noted that in both duration and pitch analyses, for cases with the patterns of I_-, I_I, and -_I, the target note has the same duration/pitch as its adjacent notes. Thus the duration/pitch pattern has no influence on the manifestation of the lexical stress. In the patterns of S_L, and L_S in the duration analysis, and the patterns of L_H and H_L in the pitch analysis, each case is an ascending scale or descending scale. Thus, the influence of the duration/pitch pattern on the

lexical stress is not clear. The number of such cases is 253 out of 683 in the duration analysis and 269 out of 687 in the pitch analysis. The present study will not focus on these patterns.

3. Results

3.1.1. Stress match

Table 3 shows the degrees of stress match. Regarding the average match degrees of these fifteen songs, the percentage of perfect match is 67.71%. It is 9.05% for secondary match, and 23.39% for mismatch. The degree of perfect match for each song is the highest among the three match categories, and also the percentage is higher than 50% (except 7 and 8).

Concerning the specific meter patterns, songs with 3/4 meter have the highest degree of perfect match whereas songs with 4/4 meter have the lowest degree. However, songs with 4/4 meter do not have a high mismatch degree since they still have a considerable degree in secondary match. It is found that songs with 2/4 and 2/2 meter have the highest mismatch degree (and is much higher than that of the songs with the other three meter patterns), showing that their general match are not as good as other songs. By contrast, songs with 6/8 meter have the lowest mismatch degree, showing that they have very good general match.

Table 3. *Degrees of stress match.*

Meter Patterns	Songs	Perfect Match	Average for each meter	Secondary Match	Average for each meter	Mismatch	Average for each meter
2/2	1)	82.61%	65.84%	0	0.43%	17.39%	33.73%
	2)	62.82%		1.28%		35.90%	
	3)	52.08%		0		47.91%	
2/4	4)	53.25%	57.31%	0	4.00%	46.75%	38.69%
	5)	52.00%		12.00%		36.00%	
	6)	66.67%		0		33.33%	
4/4	7)	45.37%	54.78%	28.70%	26.10%	25.93%	19.12%
	8)	47.54%		24.59%		27.87%	
	9)	71.43%		25.00%		3.57%	
3/4	10)	85.71%	85.14%	0	1.04%	14.29%	13.82%
	11)	81.25%		3.13%		15.63%	
	12)	88.46%		0		11.54%	
6/8	13)	100%	76.81%	0	13.89%	0	9.30%
	14)	51.11%		24.44%		24.44%	
	15)	79.31%		17.24%		3.45%	
Average		67.97%		9.09%		22.93%	

3.1.2. Duration patterns

In this section, the cases in perfect match plus moderate match in Table 1 are regarded as general match, and the cases in poor match plus mismatch are regarded as general mismatch. The data is reported in Table 4 which shows the specific match cases of each song in the duration analysis.

It is clear in Table 4 that in the group of Xa, 9 songs have more cases in general match than those in general mismatch. 6 songs have more cases in general mismatch than those in general match. If we take all fifteen songs as a whole, for Xa, there are 64 cases in perfect match, 46 cases in moderate match, 58 cases in poor match and 32 cases in mismatch. Thus there are 110 cases in general match, and 90 in general mismatch. As a result, the general match cases outnumber the general mismatch cases.

In Xc, 7 songs have more cases in general match than those in general mismatch, 5 songs have more cases in general mismatch than those in general match, and 3 songs have the same number of cases in general match and in general mismatch. If we view the fifteen songs as a whole, for Xc, there are altogether 36 cases in perfect match, 80 cases in moderate match, 26 cases in poor match and 82 cases in mismatch. Thus, there are 116 cases in general match and 108 cases in general mismatch. As a result, the general match cases again outnumber the general mismatch cases.

Table 4. *Matching cases in duration analysis.*

Duration Analysis	Xa		Xc	
	Song number	Total	Song number	Total
GM > GMi	1), 4), 5), 6), 9), 10), 12), 13), 14)	9	1), 4), 5), 6), 7), 9), 12)	7
GM = GMi	--	0	2), 10), 14)	3
GM < GMi	2), 3), 7), 8), 11), 15)	6	3), 8), 11), 13), 15)	5

Note: GM: general match; GMi: general mismatch.

3.1.3. Pitch patterns

Table 5 shows the data concerning the specific matching cases of each song in the pitch analysis.

Table 5. *Matching cases in pitch analysis.*

Pitch Analysis	Xa		Xc	
	Song number	Total	Song number	Total
GM > GMi	1), 2), 3), 5), 6), 7), 9), 10), 11), 12), 13), 14), 15)	13	3), 4), 6), 7), 10), 11), 12), 14), 15)	9
GM = GMi	--	0	2), 9), 13)	3
GM < GMi	4), 8)	2	1), 5), 8)	3

Note: GM: general match; GMi: general mismatch

According to Table 5, for Xa, 13 songs have more cases in general match than those in general mismatch. 2 songs have more cases in general mismatch than those in general match. With regard to the whole picture of the match cases of the 15 songs, there are 82 cases in perfect match, 59 cases in moderate match, 43 cases in poor match and 49 cases in mismatch. Thus for Xa, there are 141 cases in general match and 92 cases in general mismatch. So there are more general match cases than general mismatch cases.

As for Xc, 9 songs have more cases in general match than those in general mismatch, 3 songs have more cases in general mismatch than those in general match, and 3 songs have the same number of cases in both general match and general mismatch. For a more general view, in Xc, there are 74 cases in perfect match, 26 cases in moderate match, 43 cases in poor match and 35 cases in mismatch. As a result, there are 100 cases in general match and 78 cases in general mismatch. Again, the general match cases outnumber the general mismatch cases.

4. Discussion

From the results of the degrees of stress match, it is clear that for all the songs except 7) and 8), the percentage of perfect match significantly outnumbers the other two match categories. Besides, all the songs have fewer than half of the cases with mismatched stress patterns. Therefore, it can be concluded that, the general good match situations between metrical stress in

music and lexical stress in lyrics are dominating in English songs. Concerning the five specific meter patterns, songs with 3/4 and 6/8 meters have the best stress mapping, whereas songs with 2/2 and 2/4 meter patterns have the most unsatisfactory stress mapping. The reason for this may be due to the difference of the characteristics of the meters. 3/4 and 6/8 meters are in relatively free and relaxing style whereas 2/2 2/4 and 4/4 meters are more compact and strict. Consider that many rhymes and marching songs are written in 2/2, 2/4 and 4/4 meters whereas waltz and barcarolles which are commonly used to dance to are written in 3/4 and 6/8 meters. 6/8 meter is often considered as the double sets of 3/4 meter. Since it has more beats (6 beats) as well as one minor stress per measure, its meter space allows more flexibility in accommodating lexical stress.

Songs with 2/2 and 2/4 meters have the highest percentages of mismatch degree. Since in most cases, songs with 2/2 and 2/4 meters can be naturally written into 4/4 meters due to their very similar auditory features, our selected 2/2 and 2/4 meter songs can be analyzed as 4/4 meter songs. Under such condition, the match situation is totally different as Table 6 shows. The mismatch degrees for these songs sharply decline. This is because when compared with 2/2 and 2/4 meters, 4/4 meter features the strictness like them on the one hand, and has a minor stress which mediates the major stress and the unstress on the other, thus adding the possibilities of secondary match and providing this meter with more flexibility. This is also the reason why the stress match degree for 4/4 meter ranks between that of 2/2 and 2/4 meters.

It should be noted that a song can be written in different meter signatures. It is common that songs of a binary meter can be written in any other binary meters but resulting little auditory difference for the audience (same situation for songs of triple meters). Besides, composers may develop the original piece to a special version by adding syncopations and anacruses to arouse feelings of out of expectation. These factors may cause a different matching result. In this study, however, the selected songs are original versions or the accepted earliest English versions to avoid the above factors.

Table 6. *2/2 and 2/4 meter songs analyzed as 4/4 meter songs in stress match.*

Song number	Meter pattern	Perfect match	Secondary match	Mismatch	Mismatch average
1)	**2/2**	47.92%	35.42%	16.66%	6.84%
2)		83.33%	12.82%	3.85%	
3)		69.57%	30.48%	0.00%	
4)	**2/4**	70.13%	24.68%	5.19%	2.81%
5)		83.87%	12.90%	3.23%	
6)		58.33%	41.67%	0.00%	

As for the duration analysis and pitch analysis, it is found that most of the selected songs have good mapping between duration/pitch patterns and lexical stress distributions for both the primarily stressed syllables and the unstressed syllables. In these songs, the duration/pitch patterns of the musical notes correspond with lexical stress patterns very well to highlight the primary lexical stress and weaken the unstressed syllables auditorily. However, there exist variations of match degrees between the two analyses. Table 7 incorporates the general match situations in duration analysis and pitch analysis. The ratio between general match and general mismatch shows the match degrees. The higher the ratio is, the higher the match degree is. The result shows that pitch patterns have higher match ratios than those of duration patterns for both primarily stressed syllables (1.53 vs. 1.22) and unstressed syllables (1.28 vs. 1.05). In this sense, it can be inferred that in English songs, the pitch match is more prominent than the duration match. Phonetically, pitch is a more important cue than duration in the manifestation of English lexical stress [16] [17]. Our data show the same patterns in English songs. Thus, it can be concluded that vocal music shares common rhythmic features with speech.

In addition, it can be seen from Table 7 that cases of primarily stressed syllables have higher match degrees than those of unstressed syllables in both duration analysis (1.22 vs.1.05) and pitch analysis (1.53 vs. 1.28). So the primarily stressed syllables have better mapping between lexical stress and duration patterns as well as pitch patterns than that of unstressed syllables. It should be noticed that duration analysis and pitch analysis have unequal total numbers of general match cases plus general mismatch cases. This is because this study has excluded the cases in which the patterns have no explicit influence on the lexical stress (see the end of 2.2.3).

Table 7. *Ratios between general match and general mismatch.*

Analysis	Patterns	GM	GMi	GM/GMi
Duration analysis	**Xa**	110	90	1.22
	Xc	116	110	1.05
Pitch analysis	**Xa**	141	92	1.53
	Xc	100	78	1.28

Note: GM: general match; GMi: general mismatch; Xa: target with primarily stressed syllables; Xc: target with unstressed syllables

It is also noticed from Table 3, Table 4 and Table 5 that, the song *Yesterday Once More* (No. 8) is found to have very poor rhythmic match in all three aspects (the perfect match degree is low in stress match, and the general mismatch cases outnumber the general match cases in both duration and pitch analyses). However, this song is classic and popular. *Yesterday Once More* used to be the highest-debuting single in 1973 (according to *Cash Box*, a music trade magazine). And until now, it is still among the most popular golden oldies in many places. What makes this song universally loved by so many people even though it has such a poor rhythmic mapping? It is likely that there are some other songs in similar situation. Are there any special rhythmic patterns or rhythmic interactions which make them auditorily harmonious despite their poor rhythmic match? These questions await investigation in further studies.

In summary, the current study shows that there is a good correspondence between metrical stress in music and English lexical stress. Besides, most of the selected songs have good mappings between duration/pitch and English word stress, and the duration patterns/pitch patterns in music can manifest English word stress in most cases. The pitch match is more prominent than duration match in the analysis. In addition, primarily stressed syllables have better mapping between lexical stress and duration patterns/pitch patterns than unstressed syllables. The good rhythmic correspondence between lyrics and music may be because that the composers and lyricists take special care to match the rhythms in melody with the stress in lyrics when they compose melody for pre-existing lyrics or write lyrics for a pre-existing tune. However, there might be variations of match degrees for the two types of work. It remains unclear which one has better rhythmic mapping in English vocal music. This is also an interesting question for further research.

5. References

[1] Lerdahl, F., and Jackendoff, R., "A generative theory of tonal music", Cambridge, Mass.: MIT Press, 1983.

[2] Todd, N. P. M., "Segmentation and stress in the rhythmic structure of music and speech: A wavelet model", Journal of the Acoustical Society of America, 93, (2363), 1993.

[3] Patel, A. D., and Daniele, J. R., "An empirical comparison of rhythm in language and music", Cognition, 87, B35–B45, 2003.

[4] Patel, A. D., Iversen, J. R., and Rosenberg, J. C., "Comparing the rhythm and melody of speech and music: The case of British English and French", Journal of the Acoustical Society of America, 119, 3034–3047, 2006.

[5] Mcgowan, R. W., and Levitt, A. G., "A comparison of rhythm in English dialects and music", Music Perception, 28(3), 307–313, (2011).

[6] Palmer, C., and Kelly, M. H., "Linguistic prosody and musical meter in song", Journal of memory and language, 31, 525–5 42, 1992.

[7] Temperley, N., "Stress-meter alignment in French vocal music", Journal of the Acoustic Society of America, 2013.

[8] Ashby, P., "Understanding phonetics", London: Hodder Education, 2011.

[9] Antley, B. R., "The rhythm of medieval music: a study in the relationship of stress and quantity and a theory of reconstruction with a translation of John of Garland's de Mensurabili Musica", Ann Arbor, Mich.: University Microfilms International, 1983.

[10] Orbach, J., "Sound and Music: for the pleasure of the brain", Lanham: University Press of America, 1999.

[11] Henry, E., "Music Theory", Englewood Cliffs, New Jersey: Prentice-Hall, INc, 1985.

[12] Huron, D., and Royal, M., "What is melodic accent? Converging evidence from musical practice", Music Perception, 13, 489–516, 1996.

[13] Hannon, E. E., and Snyder, J. S., "The role of melodic and temporal cues in perceiving musical meter", Journal of Experimental Psychology: Human Perception and Performance, 30 (5), 956–974, 2004.

[14] Drake, C., and Palmer, C., "Accent structures in music performance", Music Perception, 10, 343–378, 1993.

[15] Jones, M. R., "Dynamics of musical patterns: How do melody and rhythm fit together?" In Tighe, T. J. & Dowling, W. J. (Eds.), Psychology and music: The understanding of melody and rhythm. Hillsdale, N.J.: L. Erlbaum. 67–92, 1993.

[16] Lehiste, I., "Suprasegmentals", Cambridge, Mass.: M.I.T. Press, 1970.

[17] Clark, J., Yallop, C., and Fletcher, J., "An introduction to phonetics and phonology", Oxford: Blackwell Pub., 2007.

Speech rhythm and vowel raising in Bulgarian Judeo-Spanish

Christoph Gabriel [1], *Elena Kireva* [1,2]

[1] University of Hamburg, [2] University of Osnabruck (Germany)

christoph.gabriel@uni-hamburg.de, elena_kireva2004@yahoo.de

Abstract

The study investigates selected prosodic characteristics of (Sofian) Bulgarian Judeo-Spanish, a diaspora variety of Spanish spoken by descendants of the Jews expelled from Spain, all of them bilingual speakers with Bulgarian as their dominant language. While exhibiting some few relics from Old Spanish on the segmental level, Judeo-Spanish shows a puzzling similarity with Bulgarian with respect to speech rhythm and vowel raising. It is shown that the two languages spoken by the bilinguals, Bulgarian and Judeo-Spanish, pattern alike in displaying almost the same rhythmic values (except for %V) ([1], [2], [3], [4], [5]) and that raising of unstressed /a/ and /o/ as is typical of the variety of Bulgarian spoken in Sofia also regularly occurs in the Judeo-Spanish data. Our findings show that Judeo-Spanish is crucially influenced by Bulgarian, thus suggesting that it has largely converged toward the surrounding language on the phonological level.

Index Terms: Judeo-Spanish, Spanish, Bulgarian, language contact, speech rhythm, vowel raising

1. Introduction

Judeo-Spanish (JUSPA) emerged in the Middle Ages in a socio-political context marked by the contact of several languages. The Sephardic Jews living in Spain formed an ethno-sociological group different in customs and beliefs from the non-Jewish population. Their main vernacular was (Medieval) Spanish, while they used Hebrew in ritual and educational contexts. After the expulsion from Spain in 1492, their vernacular developed independently of Iberian Spanish due to contact with the new surrounding languages, among them Bulgarian. The Bulgarian variety of JUSPA is spoken by a rather small group of about 250-300 (at least) bilingual speakers with Bulgarian (BULG) as their dominant language, the JUSPA community of Sofia being even much smaller. The oldest living native speakers were born around 1920, the youngest in the 1960ies ([6], [7]). The use of JUSPA is restricted to informal communication within the community; in 1998, *Club ladino*, a community center for Sephardic Jews, was founded in the city center of Sofia.

Apart from the remarks included in general presentations (e.g. [8]), the literature on JUSPA phonology is rather sparse. Both the segmental and the tonal properties of the variety of Judeo-Spanish spoken in İstanbul (Turkey) have recently been investigated ([9] and [10]; the only study explicitly dealing with the phonemic system of Bulgarian JUSPA stems from the mid 1970ies ([11]). The prosody, in particular speech rhythm and vowel raising, of the variety in focus has not been investigated until now.

2. Phonological properties of JUSPA

This section highlights selected properties of JUSPA in comparison with the surrounding language BULG on the one hand and (Castilian) SPA on the other.

SPA exhibits an unmarked vowel system with the five phonemes /i, e, u, o, a/; BULG., in addition, has a phonemic schwa /ə/ ([12]) that also forms part of the JUSPA vowel system, albeit restricted to word-medial stressed syllables in Bulgarian loan words ([11]). A striking feature that has a crucial impact on speech rhythm is the presence or absence of vowel reduction ([1]). While completely absent from SPA ([13]), Bulgarian presents reduction (or rather raising) of the vowels /o, e, a/ that are realized as [u, i, ə] in unstressed syllables, e.g. [ˈrabutə] *работа* 'work' vs. [rəˈbotnik] *работник* 'worker' ([12], [14], [15], [16]). However, in the variety of BULG spoken in Sofia the raising of vowels commonly only affects /a/ and /o/, which surface in unstressed position as [ə] and [u], respectively, while the front vowel /e/ is hardly ever reduced to [i] ([14]). Interestingly, (Sofian) JUSPA also presents the feature of vocalic reduction ([11]), presumably as a consequence of its long-lasting contact with Bulgarian. Note in this context that vowel reduction has not been attested in other varieties of Judeo-Spanish (see, e.g., [9] concerning the variety spoken in İstanbul). Figure 1 depicts the process of vowel raising as is attested in Sofian BULG and JUSPA.

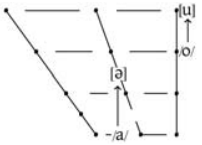


Figure 1: *Vowel raising in Sofian BULG and JUSPA.*

Regarding its consonantal system, JUSPA has preserved some features from Medieval Spanish, e.g. the sibilants /ʃ, ʒ/ (e.g., *bajo* [ˈbaʃo] vs. SPA [ˈbaxo], *mujer* [muˈʒeɾ] vs. SPA [muˈxeɾ], the [±]voiced contrast of the alveolar fricatives /s/ vs. /z/ ([11]), [17]). A reason for the maintenance of the rich sibilant system in JUSPA might be seen in the fact that the BULG consonant system also includes the phonemes /s z ʃ ʒ/ ([18]).

Concerning its rhythmic properties, (Castilian) SPA is traditionally classified as a typical syllable-timed language that strongly prefers regular sequences of CV syllables and completely lacks vowel reduction ([19]). Seen from the phonetic side, SPA is characterized by a high proportion of vocalic material (%V) and rather low values for the durational variability of both V and C intervals, as compared to stress-timed languages such as English ([1], [4], [20]). BULG, as opposed to SPA, exhibits complex syllable structures, allowing for up to three consonants in both the onset and the coda, i.e. (CCC)V(CCC), and presents vowel raising (see above), but usually no complete deletion of unstressed vowels as occurs, e.g., in English ([21], [22]). It is thus "less stress-timed" than English; its speech rhythm may thus be characterized as being of a mixed type ([23]). Until now, there is no work investigating the speech rhythm of JUSPA. For the current study, we hypothesize that the bilingual speakers (at least partially) transfer the rhythmic values from BULG to JUSPA, i.e. we assume that the rhythmic values shown in the

speech production of both of their languages are situated between the ones for SPA on the one hand and for Bulgarian, produced by monolingual speakers, on the other. Concerning the production of unstressed vowels, we expect the bilingual speakers to raise unstressed /a/ and /o/ in their production of JUSPA, due to transfer from BULG.

3. Methodology

3.1. Speakers

We collected data from five female bilingual JUSPA/BULG speakers, aged 80 to 88 (recordings Sofia, September 2012). Although they were born and raised in different places in Bulgaria (Kyustendil, Pazardzhik, Kazanlak, Samokov, and Karnobat), all of them have been living in Sofia for more than 60 years and speak the Sofian variety of BULG. The bilingual subjects were recorded in both of their languages (JUSPA and BULG_B(ilingual)). Five monolingual BULG speakers (1 ♂, 4 ♀, ages 26–34) and five monolingual speakers of (Castilian) SPA (3 ♂, 2 ♀, ages 24–34) serve as control groups. The Bulgarian subjects were born and raised in Sofia; the Spanish speakers were all born in the Castilian dialect area and raised in Madrid. The monolingual BULG speakers were recorded in Sofia in September 2012 (BULG_M(onolingual)); the (Castilian) SPA control data were gathered in Madrid in September 2011.

3.2. Material

The material gathered for the analysis of vowel raising and speech rhythm consisted of reading of the fable *The North Wind and the Sun*, recorded in BULG_M, BULG_B (*Северният вятър и слънцето*), JUSPA (*El ayre del norte i el sol*), and (Castilian) SPA (*El viento norte y el sol*), respectively. The data were recorded with a Marantz hard disk recorder (PMD671) and a Sennheiser microphone (ME64) and analyzed using *Praat* ([24]).

3.3. Analysis of vowel raising

In order to demonstrate that Judeo-Spanish also exhibits vowel raising like the contact language Bulgarian, both an auditory and an acoustic analysis were performed. First, all unstressed /a/ and /o/ occurring in the data recorded from the bilingual speakers (i.e. JUSPA and BULG_B) were transcribed by the second author (who is a native speaker of Bulgarian) according to their auditory properties as being reduced or unreduced. In a second step, two raters, both native speakers of Bulgarian, were asked to determine every single unstressed /a/ and /o/ as being realized as [a] or [ə] for /a/, and as [o] or [u] for /o/. For the final results of the auditory analysis, every vowel which was defined as being reduced by the second author and by at least one of the raters was counted as being reduced. The transcription agreement between the three raters amounts to 86 % for JUSPA and to 89% for BULG_B.

Subsequently, all stressed and unstressed /a/ and /o/ occurring in the data collected from the bilinguals were analyzed acoustically. In order to compare the formant frequencies of the unstressed vowels with the ones of the stressed ones in both of their languages, we measured the formants of all stressed and unstressed /a/ and /o/ in the JUSPA and the BULG_B data. All /u/ and /ə/ (we refer here to the Bulgarian /ə/ that occurs in both stressed and unstressed position) were also taken into account in order to compare the formant frequencies of the unstressed /a/ and /o/ with the respective values of /u/ and /ə/. Material produced with creaky voice or disfluencies was excluded from the acoustic analysis.

The first two formants of the above mentioned vowels were extracted using the *Praat* function Formant Track and running a script that provides three scores for F1 and F2 of each vowel (measured at the 25%, 50%, and 75% temporal points of the vocalic duration). Since the two data sets (BULG_B and JUSPA) differ with respect to the occurrences of syllable structures (CV, CVC, CCV, etc.) and regarding the segments that precede the vowels as syllabic onsets (plosives, nasals, liquids, etc.), we used only the values obtained from the measurements in the middle of the vowel in order to avoid co-articulation effects.

The values obtained from the formant tracker were checked randomly by the second author; incorrect values were respectively changed in the overall results, following [25]. We calculated the mean F1 and F2 values for stressed and unstressed /a, o/ occurring in both JUSPA and BULG_B. The occurrences of stressed and unstressed /u/ were grouped together for both languages, since /u/, as opposed to /a/ and /o/, is not expected to exhibit stress-dependent qualitative differences in Bulgarian. The same holds for stressed and unstressed /ə/.

3.4. Analysis of speech rhythm

For the rhythmic analysis, the whole data recorded in BULG_M, BULG_B, JUSPA, and SPA were segmented into vocalic and consonantal intervals.

Following [2, 26], the boundaries between V and C intervals were determined on the basis of formant structure and pitch period and set at the point of zero crossing of the waveform. Pre-pausal and phrase-final intervals were considered for the analysis since possible effects of final lengthening were likely to be reflected in the measures ([2], [4]). According to [4], we treated glides as belonging to the V intervals if there was no friction attested in the data. For plosives and affricates following a stretch of silence (pause) the beginning was placed at 0.05s prior to the burst, given that their boundaries can hardly be determined on the basis of the aforementioned criteria ([27]). Silent pauses and material affected by any kind of speech disfluency were excluded from the analysis.

Using the software *Correlatore* ([28]), we calculated both the proportion of vocalic material in the speech signal (V%) and the durational variability of vocalic and consonantal intervals as expressed by the variation coefficient VarcoV/C and the Pair-wise Variability Index (VnPVI, CnPVI, CrPVI). VarcoV/C is a speech rate normalized version of ΔV/C, which expresses the standard deviation of vocalic and consonantal intervals; the PVIs differ from both ΔV/C and VarcoV/C in computing the durational variability in successive V/C intervals instead of calculating the variability of V/C intervals over the whole acoustic signal; see [1], [2], [3], [4], and [5].

4. Results

The analysis of vowel raising and speech rhythm showed that the two languages spoken by the bilingual speakers, i.e. JUSPA and BULG_B, pattern alike with respect to both vowel raising (similar formant frequencies) and speech rhythm (comparable values for the rhythm metrics).

4.1. Vowel raising

Table 1 represents the results of the auditory analysis and shows the occurrences of reduced /a/ and /o/ in JUSPA and BULG_B.

Table 1. *Occurrences of reduced and unreduced /a, o/ in unstressed positions in JUSPA and BULG_B.*

	JUSPA		BULG_B	
	№ of /a/ and /o/	%	№ of /a/ and /o/	%
reduced /a/	135 /a/	75.5%	108 /a/	84.5%
unreduced /a/		24.5%		15.5%
reduced /o/	107 /o/	36%	70 /o/	71%
unreduced /o/		64%		29%

According to the outcomes of our auditory analysis, the bilingual speakers realized /a/ as [ə] in 75.5% of the cases in JUSPA and in 84.5% in BULG_B. The realization of /o/ as [u] amounts to 36% in the JUSPA data and to 71% for BULG_B. Summarizing, raising of /a/ and /o/ occurs more frequently in BULG_B than in JUSPA.

Tables 2 and 3, below, present the formant frequencies for the data gathered from the bilingual informants. The results of the acoustic analysis clearly show that the bilingual speakers exhibit vowel raising in both JUSPA and BULG_B.

Table 2. *Mean formant frequencies for JUSPA (Hz).*

	stressed /a/	unstressed /a/	stressed /o/	unstressed /o/	/u/
№	51	135	77	107	48
F1	877	633	602	413	348
F2	1491	1635	1067	1067	904

Table 3. *Mean formant frequencies for BULG_B (Hz).*

	stressed /a/	unstressed /a/	stressed /o/	unstressed /o/	/u/	/ə/
№	29	108	70	70	28	84
F1	847	520	578	385	330	490
F2	1650	1721	1094	1066	1032	1617

As can be seen in Tables 2 and 3, the material recorded comprises different numbers of stressed and unstressed vowels (i.e. 51 stressed /a/ and 135 unstressed /a/ for JUSPA, but 29 stressed /a/ and 108 unstressed /a/ for BULG_B, etc.). As for the formant frequencies, while the mean F1 values for stressed /a/ are almost the same in both the JUSPA and the BULG_B productions (877 Hz for JUSPA and 847 Hz for BULG), this is not the case for the mean F2 values which are 1491 Hz for JUSPA and 1650 Hz for BULG_B. Regarding /a/, it can be said that unstressed /a/ is frequently produced as [ə] in both bilingual languages (mean F1 value 633 Hz for JUSPA and 520 Hz for BULG_B). Thus, we observe a statistically significant difference between the F1 scores for stressed and unstressed /a/ in both JUSPA and BULG_B (dependent t-test: D=244.8±20.5, t(4)=26.709, p<0.001 for JUSPA; D=347.7±34.2, t(4)=22.745, p<0.001 for BULG_B).

Nevertheless, unstressed /a/ is more likely to be reduced (or rather: raised) in BULG_B (F1=520 Hz and F2=1721 Hz) than in JUSPA (F1=633 Hz and F2=1635 Hz), since the values for BULG_B are closer to those of stressed or unstressed /ə/ (see Table 3: F1=490 Hz and F2=1617 Hz). These findings confirm the outcomes of the auditory analysis in which the raters defined /a/ as [ə] in 84.5% of the cases for BULG_B and in 75.5% of the cases for JUSPA. The picture doesn't change when the realization of unstressed and stressed /o/ is taken into account: The mean F1 and F2 values for the stressed /o/ are quite similar in JUSPA (F1=602 Hz and F2=1067 Hz) and in BULG_B (F1=578 Hz and F2=1094 Hz). Although, according to the auditory analysis, unstressed /o/ seems to be realized as [u] twice more in the BULG_B speech than in the JUSPA material, the mean F1 and F2 scores for both bilingual varieties are nearly equal (F1=413 Hz and F2=1067 Hz for JUSPA; F1=385 Hz and F2=1066 Hz for BULG_B). The mean formant frequencies for unstressed /o/ pattern with the F1 and F2 values for /u/ rather than with the F1 and F2 scores for stressed /o/ in both varieties. However, the differences between the mean F1 and F2 scores for unstressed /o/ and for /u/ are higher in the JUSPA productions than the same ones in the BULG_B data. Nevertheless, the difference between the F1 scores for stressed and unstressed /o/ in both varieties is statistically significant (D=189.3±30.9, t(4)=13.697, p<0.001 for JUSPA; D=196.4±25.1, t(4)=17.482, p<0.001 for BULG_B).

4.2. Speech rhythm

Table 4 summarizes the values obtained from the analysis of the fable *The North Wind and the Sun* for SPA, JUSPA, BULG_B, and BULG_M. The scores for the two varieties spoken by the bilingual speakers are quite similar (except for %V), the values for VarcoV und VnPVI largely being situated between those obtained from the analysis performed on the SPA and the BULG_M data.

Table 4. *Mean values of six rhythm metrics for SPA, JUSPA, BULG_B, and BULG_M.*

	%V	VarcoV	VarcoC	VnPVI	CrPVI	CnPVI
SPA	40.4	43.3	39.8	36.6	40.6	46.3
JUSPA	45.6	45.3	42.3	43.6	60.3	45.7
BULG_B	38.6	44.1	41.1	44.3	58.1	47.5
BULG_M	33.7	49.7	38.1	49.8	50.9	45.1

While BULG_M displays the lowest proportion of vocalic material, BULG_B, SPA, and JUPSA show higher values for %V. As for the variability of vocalic intervals, the VarcoV and VnPVI scores for JUSPA and BULG_B are almost the same: More precisely, they are situated between those obtained from the analysis performed on the data produced by the two control groups (i.e. BULG_M and SPA). The differences between SPA on the one hand and JUSPA, BULG_B, and BULG_M on the other for the VnPVI values are statistically significant (SPA vs. JUSPA p=0.036, SPA vs. BULG_B p=0.018, and SPA vs. BULG_M p<0.001). Regarding the variability of consonantal intervals, both bilingual varieties exhibit quite similar values for VarcoC, CrPVI, and CnPVI. Considering the normalized metrics (VarcoC and CnPVI), it can be said that all four varieties pattern together showing similar variability of consonantal intervals; however, taking into account the values for the non-normalized or raw Pair-wise Variability Index for consonantal intervals (CrPVI), BULG_M and SPA display a lower variability of consonantal intervals. Figure 2 represents the distribution of the four varieties under discussion over the %V/VnPVI plane.

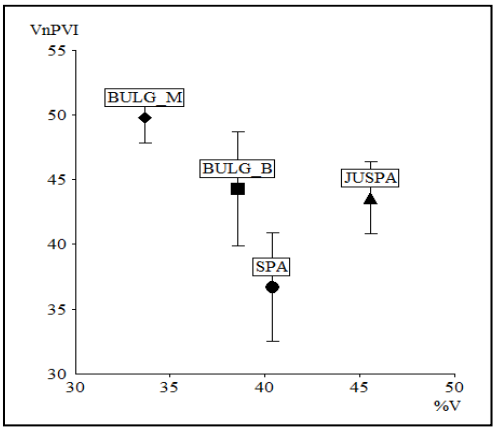


Figure 2: *Distribution of BULG_M, BULG_B, JUSPA, and SPA over the %V/VnPVI plane.*

5. Discussion

We interpret the results of our rhythmic analyses as follows: Regarding the proportion of vocalic material (%V) and the variability of V intervals (VarcoV and VnPVI) for BULG_M and SPA, it is expected that BULG_M displays a greater variability of vocalic intervals, but lower percentages for vocalic material (%V) than SPA, due to the presence of vowel reduction in Bulgarian. These expectations are confirmed by the results shown in Table 4 (see also Figure 2). Regarding the variability of C intervals (as is expressed by VarcoC, CrPVI, and CnPVI, respectively), the results confirm previous findings reported in, e.g., [21], who showed that Bulgarian (a language that presents complex consonant clusters and thus long C intervals almost throughout the whole speech signal) exhibits a variability of consonantal intervals similar to those of syllable-timed languages such as Spanish (a language with simple structures and thus presenting continuously short C intervals). This similarity becomes obvious when the values for the speech-rate normalized PVI (CnPVI) are taken into account (46.3 for SPA and 45.1 for BULG_B).

Interestingly, BULG_B patterns with JUSPA rather than with BULG_M in displaying almost the same rhythmic values (except for %V), which can be explained by the influence from Judeo-Spanish, the other language used by the bilingual speakers. The (unexpectedly) high %V values for JUSPA can be explained by the fact that the speakers read the fable in Judeo-Spanish slower as compared to their reading of the Bulgarian version of the text (regarding the influence of speech rate on rhythm see [21], [29], [30]). The reason for this might be the fact that their predominant language is Bulgarian and they are not accustomed to the use of Judeo-Spanish in its written form. Regarding the variability of V intervals, expressed by VarcoV and VnPVI (see Figure 2 for the latter), both BULG_B and JUSPA display intermediate scores situated in between those of SPA and BULG_M. The lower variability of V intervals in the BULG_B as compared to BULG_M once again might be interpreted as an effect of 'syllable-timed influence' from JUSPA in the bilingual speakers. The higher VnPVI value for JUSPA as compared to the one for SPA, in turn, suggests 'stress-timed influence' from Bulgarian and may be interpreted as an effect of vowel reduction (or rather: vowel raising), a phonological feature that is completely absent from (Castilian) SPA and has presumably been transferred to JUSPA from the dominant language Bulgarian in the bilingual speakers. Regarding the variability of C intervals, the differences between the languages investigated are less clear (see Table 4). However, the two varieties spoken by the bilingual speakers, JUSPA and BULG_B, once again display similar values.

To sum up, the phonological shape of the diaspora variety JUSPA patterns with its contact language Bulgarian in several respects: It presents vowel reduction (or rather: vowel raising) in the same way as the Sofian variety of Bulgarian does (raising of /a/ and /o/ to [ə] and [u] in unstressed position). Regarding the variability of vocalic intervals, it exhibits intermediate values, located in between those of the variety of Bulgarian spoken by monolingual speakers in the capital of Sofia (BULG_M) and Castilian Spanish (SPA). These findings can be attributed to the long-standing contact with Bulgarian and to convergence of two phonological systems (Bulgarian and Spanish) in the bilingual speakers.

6. Conclusion

Sofian Judeo-Spanish phonology is characterized by two opposite aspects: Some segments, e.g., the sibilants /ʃ/ and /ʒ/ mentioned in section 2, above, are maintained from Medieval Spanish, thus, at least partly, attributing a conservative character to the variety investigated here. The innovative features, though, are more striking, since both the feature of vowel reduction (or rather: vowel rasing) and the rhythmic properties are (at least partially) transferred from the surrounding language Bulgarian to Judeo-Spanish. To put it bluntly, it may be stated that the bilingual JUSPA/BULG_B speakers practically use the same phonology for both of their languages − at least regarding the aspects investigated in the present study. This might be due to the fact that the first Sephardic Jews who arrived in the Ottoman Empire (today: Bulgaria) and started acquiring Bulgarian as an L2 drew on the resemblances between the two phonological systems in order to avoid high cognitive costs in language processing (see [31]). During the initial period, they might have had two distinct phonologies, while the segments that belong to both of the systems, such as the sibilant phonemes, increasingly converged with respect to their concrete phonetic realization. In a further step, the two systems might have completely converged, insofar that phonological features not belonging to the Judeo-Spanish system (such as vowel raising) were integrated as well. It is even conceivable that the contemporary bilingual JUSPA/BULG_B speakers dispose of one phonological system only, i.e. the Bulgarian one, as a result of an entire convergence on the phonological level.

All things considered, we argue that the variety of Judeo-Spanish nowadays spoken in Sofia (JUPSA) is a typical contact variety that exhibits features of the languages involved in the situation of linguistic contact. Our results by and large confirm the view that the sound shape of a given language is more likely to adopt features from other languages in contact situations than is the case for core-syntactic properties such as, e.g., the ordering of verb and object (OV vs. VO) [32].

7. Acknowledgements

We would like to express our gratitude to A. Benet and S. Cortés for providing us with the *Praat* script we used and to A. Bakardzhiev for modifying it according to our needs. Further thanks you to V. Druchkiv (UKE Hamburg-Eppendorf, Germany) for his help with the statistical analyses.

8. References

[1] Ramus, F., Nespor, J. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition 73:65-192, 1999.

[2] White, L. and Mattys, S. L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics 35:501-522, 2007.

[3] Dellwo, V. and Wagner, P., "Relations between Language Rhythm and Speech Rate", in M. Solé, D. Recasens and J. Romero [Eds], Proceedings of the 15th International Congress of Phonetic Sciences, 2693-2696, Casual Productions, 2003.

[4] Grabe, E. and Low, E.L., "Durational variability in speech and the Rhythm Class Hypothesis", in C. Gussenhoven and N. Warner [Eds], Papers in Laboratory Phonology 7, 515-546, Mouton de Gruyter, 2002.

[5] Kinoshita, N. and Sheppard, C., "Validating acoustic measures of speech rhythm for second language acquisition", in W.-S. Lee and E. Zee [Eds], Proceedings of the 17th International Congress of Phonetic Sciences, 1086-1089, City University of Hong Kong, 2011.

[6] Studemund-Halévy, M. and Fischer, S., "What happens when a language ceases to be used by its speakers? Documentation of Bulgarian Judezmo", in M. Studemund-Halévy, C. Liebl and I. Vucina Simovic [Eds], Sefarad an der Donau: Lengua y literatura de los Sefardíes en tierras de los Habsburgos, 407-424, Tirocinio, 2013.

[7] Schelling, A., Judenspanisch in Bulgarien: Eine bedrohte Minderheitensprache, Master Thesis, Universität zu Köln, 2005.

[8] Hetzer, A., Sephardisch: Judeo-español. Djudezmo. Einführung in die Umgangssprache der südosteuropäischen Juden, Harrassowitz, 2001.

[9] Hualde, J. I. and Şaul, M., "Istanbul Judeo-Spanish", Journal of the International Phonetic Association 41:89-110, 2011.

[10] Romero, R., "Palatal east meets velar west: Dialect contact and phonological accommodation in Judeo-Spanish", Studies in Hispanic and Lusophone Linguistics 6:279-300, 2013.

[11] Kanchev, I. V. (Кънчев, И. В.), Fonética y fonología del judeoespañol de Bulgaria, Ph. D. dissertation, University of Sofia, 1975.

[12] Danchev, A., "On the Contrastive Phonology of the Stressed Vowels in English and Bulgarian", Papers and Studies in Contrastive Linguistics 25:156-175, 1989.

[13] Martínez Celdrán, E., Análisis espectrográfico de los sonidos del habla, Ariel, 2007.

[14] Wood, S. and Pettersson, T., "Vowel Reduction in Bulgarian: The Phonetic Data and Model Experiments", Folia Linguistica 22:239-262, 1988.

[15] Crosswhite, M., "Sonority-Driven Reduction", in L. Conathan, J. Good, D. Kavitskaya, A. Wulf and A. Yu [Eds], Proceedings of the Twenty-Sixth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Aspect, 77-88, Berkeley, 2000.

[16] Gulian, M., Escudero, P. and Boersma, P., "Supervision hampers distributional learning of vowel contrasts", in J. Trouvain and W. J. Barry [Eds], Proceedings of the 15th International Congress of Phonetic Sciences, 1893-1896, 2007.

[17] Ariza, M., Fonología y fonética históricas del español, Arco libros, 2012.

[18] Ternes, E. and Vladimirova-Buhtz, T., "Bulgarian", Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 55-57, 1999.

[19] NGRAE = Real Academia Española / Asociación de Academias de la Lengua Española: Nueva gramática de la lengua española: Fonética y fonología, Espasa Libros, 2001.

[20] Benet. A., Gabriel. C., Kireva. E. and Pešková, A., "Prosodic transfer from Italian to Spanish: Rhythmic Properties of L2 Speech and Argentinean Porteño", in Q. Ma, H. Ding and D. Hirst [Eds], Proceedings of the 6th International Conference on Speech Prosody, 438-441, Tongji University Press, 2012.

[21] Barry, W., Andreeva, B., Russo, M., "Dimitrova, S. and Kostadinova, T. Do rhythm measures tell us anything about language type?", in M. Solé, D. Recasens and J. Romero [Eds], Proceedings of the 15th International Congress of Phonetic Sciences, 2693-2696, Casual Productions, 2003.

[22] Barry, W., Andreeva, B. and Koreman, J., "Do rhythm measures reflect perceived rhythm?", Phonetica 66:78-94, 2009.

[23] Dimitrova, S., "Bulgarian Speech Rhythm. Stress-timed or syllable-timed?" Journal of the International Phonetic Association 27:27-33, 1998.

[24] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" (Version 5.3.53), 2013. [computer program]. Retrieved 9th July 2013, <from http://www.praat.org/>.

[25] Cortés, S., Lleó, C. and Benet, A., "Gradient merging of vowels in Barcelona Catalan under the influence of Spanish", in K. Braunmüller and J. House [Eds], Convergence and Divergence in Language Contact Situations, 185-204, John Benjamins, 2009.

[26] Peterson, G. E. and Lehiste, I., "Duration of syllable nuclei in English", Journal of the Acoustical Society of America, 32:693-703, 1960.

[27] Mok, P. and Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", in P. Barbosa, S. Madureira and C. Reis [Eds], Proceedings of the 4th Conference on Speech Prosody, 423-426, Editora RG/CNPq, 2008.

[28] Mairano, P. and Romano, A., "Un confronto tra diverse metriche ritmiche usando Correlatore", in S. Schmid, M. Schwarzenbach and D. Studer [Eds], La dimensione temporale del parlato, Proceedings of the V National AISV Congress, 79-100, EDK, 2010.

[29] Dellwo, V., Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence, PhD thesis, RFW Univesity, Bonn, 2010.

[30] Beňuš, Š. and Šimko, J., "Rhythm and tempo in Slovak", in Q. Ma, H. Ding and D. Hirst [Eds], Proceedings of the 6th International Conference on Speech Prosody, 502-505, Tongji University Press, 2012.

[31] Matras, Y., Language Contact, Cambridge University Press, 2009.

[32] Thomason, S. G. and Kaufman, T., Language contact, creolization, and genetic linguistics, University of California Press, 1998.

The role of stress perception in the assignment of written accent in Spanish

Sandra Schwab[1], *Carla V. Jara Murillo*[2]

[1] ELCF, Université de Genève, Switzerland
[2] Escuela de Filología, Lingüística y Literatura, Universidad de Costa Rica, Costa Rica
Sandra.Schwab@unige.ch, Carla.Jara@ucr.ac.cr

Abstract

The aim of this investigation is to examine whether the adults' difficulty in placing the written accent in Spanish words is related to their ability in perceiving stress. The following variables were also taken into account in this study: the participant's education level (academic and non-academic), the stimulus lexical status (words and non-words), accentual pattern (proparoxytone, paroxytone and oxytone words) and length (2, 3 and 4 syllables). Participants performed a stress identification task and a word spelling task. Besides the effects of lexical status, education level and accentual pattern, results show an effect of the stress perception in the assignment of the written accent: stimuli with a correctly identified stress were more likely to be correctly written (i.e. with or without written accent) than the incorrectly perceived stimuli. This finding reinforces the idea that there is a relationship between prosodic and written skills.

Index Terms: stress perception, written accent, Spanish

1. Introduction

In Spanish, lexical stress is distinctive and therefore its placement is variable. As it is well known, stress distinguishes Spanish words such as *límite* (['limite] *limit*), *limite* ([li'mite] *(that) I limit*) and *limité* ([limi'te] *I limited*). Thus, it can appear on the last syllable of the word (oxytone word), on the penultimate syllable (paroxytone word) or on the antepenultimate syllable (proparoxytone word)[1]. However, the paroxytone pattern is by far the most general. It has been shown in [1] that in a corpus of 9219 Spanish words including propaxytone, paroxytone and oxytone words (excluding monosyllabic and unaccented words), 80% of the words were paroxytone, whereas approximately 17% of the words were oxytone and 3% were proparoxytone. The paroxytone pattern can thus be considered as the default accentual pattern in Spanish.

The normative spelling system establishes the following rules for placing Spanish written accent. In oxytone words, where stress is on the last syllable, a written accent is required if the word ends with -n, -s or with a vowel (e.g. *educación*, *voté*). In paroxytone words, where stress is on the penultimate syllable, a written accent is required if the word ends with a consonant other than -n and –s (e.g. *cárcel*, *carácter*). Finally, in proparoxytone words, where stress is on the antepenultimate syllable, a written accent is always required (e.g. *tránsito*).

It has been frequently noted that in Spanish orthography, the most common spelling mistake is the omission of the written accent. This comes out as the result of several investigations on different Spanish written corpora (e.g. [2], [3], [4]). In the case of Costa Rica (the variety under study in this investigation), similar results were found in a study with children (primary school) [5]. As far as adults are concerned, [6] found that 84% of the spelling errors in an adult Costa Rican written Spanish corpus (COCAE) were due to the omission of the written accent. Besides other reasons (e.g. education level, lack of attention, etc.), the difficulty in placing the written accent might be explained by the difficulty in perceiving stress. In that respect, [7] mentioned that 46% of the participants (students at the very beginning of their studies in *Pedagogy in Spanish*) had difficulties in identifying the stressed syllable. The possible relationship between spelling errors and stress perception (among other variables) was examined in 10-year children in [8]. They found that what they called the "stress awareness" (i.e. identification of the stressed syllable in non-words) was a good predictor of the spelling errors. They concluded that the prosodic skill (i.e. stress sensitivity) influenced the performance in word spelling task in 10-years children.

The aim of this investigation is to examine whether the adults' difficulty in placing the written accent is related to their ability in perceiving stress. The participant's education level (academic and non-academic) is also taken into account in this study. Moreover, as far as the stimuli are concerned, we consider the lexical status (words and non-words), the accentual pattern (proparoxytone, paroxytone and oxytone words) and the length (2, 3 and 4 syllables).

2. Method

2.1. Participants

Thirty-two Costa Rican participants (all from the San José region) took part in this experiment. They were divided into two groups. The first group was composed of 16 non-academic participants (henceforth "non-academic") (5 males and 11 females, age range: 18-70 years, mean age = 42.75). All of them had attended middle- or high-school level education, where they were taught Spanish grammar and orthography. The second group was composed of 16 students and professors from the faculty of Arts and from the faculty of Sciences of the Universidad de Costa Rica (henceforth "academic") (6 males and 10 females, age range: 19-51 years, mean age = 29.19).

2.2. Material

The auditory stimuli used in this experiment consisted of 96 items (48 real words and 48 non-words), recorded by a Costa Rican female speaker. The stimuli were prepared according to the following criteria. The real words were 48 Spanish words of two, three and four syllables, divided in proparoxytone (PP;

[1] Note that some adverbs ending in *-mente*, have two accented syllables (e.g. *comúnmente* [ko,mun'mente], *commonly*) and that some words, due to enclitics, present a stress on the forth to last syllable (e.g. *comiéndoselo* [ko'mjendoselo], *eating it*).

e.g. *únicos*)[1], paroxytone (P; e.g. *acabo*) or oxytone words (O; *además*)[2]. Twenty-one words required a written accent (e.g. *líder*) and 27 did not (e.g. *rosas*). The words with written accent were mainly taken from the COCAE corpus, as they generated a large number of errors ([6]). As far as the non-words were concerned, they were stimuli of two, three and four syllables (with PP, P or O accentual pattern) that were created by combining (in a random way but following the Spanish phonotactics) the syllables of the real words. This set comprised 27 non-words with written accent (e.g. *tébar*) and 21 without it (e.g. *meper*)[3].

2.3. Procedure

The experiment was run online with the *Labguistic* platform (www.labguistic.com). It was divided into two parts: *Perception* and *Spelling*. In the first part (Perception), participants heard a stimulus and had to indicate the position of stress (on the last syllable (O), the penultimate syllable (P) or the antepenultimate syllable (PP)). In the second part (Spelling), participants heard a stimulus and had to write it (with the instruction to place the written accent, if necessary).

The exact same stimuli were played in both parts (although not in the same order). Each part was composed of three sections, each one corresponding to the two, three or four syllable stimuli. Words and non-words were mixed together and presented in a different random order for each participant. The participants could hear each stimulus twice, if needed.

2.4. Data analysis

The responses given in the *Spelling* part were corrected according to the following criteria. We considered as correct responses the words (and non-words): 1) with two possible spellings, as there was no way to decide which spelling was the correct one (e.g. v/b; *boté* vs *voté*); 2) with a substitution of a letter with no implication for the presence of the written accent (e.g. *tía* instead of *día*); 3) with a substitution of a letter with an implication for the presence of the written accent (e.g. *rodes* instead of *ródez*); 4) with the addition of a letter with no implication for the presence of the written accent (e.g. *púdiro* instead of *údiro*), 5) with the omission of a letter with no implication for the presence of the written accent (e.g. *pasesa* instead of *pasesas*). We excluded the words (and non-words) with the addition or omission of a letter with an implication for the presence of the written accent (e.g. *mapovol* instead of *mapovo*; *simeracta* instead of *simeráctar*). Moreover, we collected the correct/incorrect responses of the *Perception* part.

We analyzed the data by means of mixed-effects logistic regression models in which participants and stimuli were entered as random terms ([10], [11]). For the sake of clarity, the results and figures are presented in percentages, although all statistical analyses have been performed on raw data.

[1] Note that disyllabic words could not be proparoxytone.

[2] We made sure that the lexeme frequency (taken from [9]) was not only similar across the three pattern ($F(2, 45) = 0.93$, n.s), but also in the two, three and four syllable words ($F(2, 45) = 0.88$, n.s).

[3] Although non-words do not exist in Spanish, they would follow the same Spanish accentuation rules as real words. For example, the oxytone non-word [me'peɾ] is written *meper*, and the paroxytone non-word ['teßaɾ] is written *tébar*.

3. Results and discussion

3.1. Spelling performance

A first model was run with the spelling correct/incorrect response as a dependent variable and with the following predictors: the participant's education level (academic/non-academic), the stimulus lexical status (words/non-words), the stimulus accentual pattern (PP, P, O), the stimulus length, and the participant's perception response (correct/incorrect). Given that the stimulus length had no effect and did not interact with other variables, a new analysis was run without it and without all non significant interactions.

3.1.1. Role of lexical status

As can be seen in Figure 1, results showed an effect of lexical status, with more correct responses for real words than for non-words ($F(1, 3033) = 33.08$, $p < .001$), independently of the other variables. Thus, it seems easier to correctly spell (i.e. to place or not the written accent) the real Spanish words than the invented words. It is also worth it noting that the scores are very high (88.36% in average), which suggests that the task is not particularly difficult for the participants. Moreover, it is important to observe the absence of an interaction between the lexical status and the educational level, which means that the difference between words and non-words is similar in academic and non-academic participants.

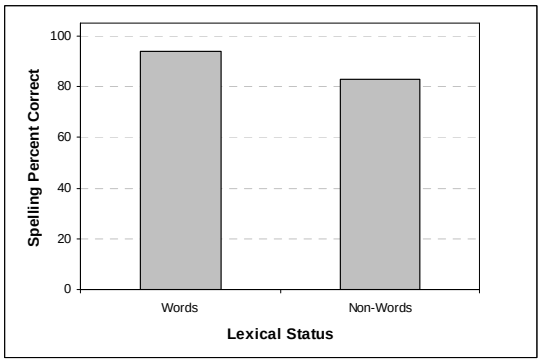


Figure 1: *Spelling percent correct as a function of lexical status.*

3.1.2. Role of education level and accentual pattern

Figure 2 presents the spelling percent correct as a function of the accentual pattern and education level. Despite the very good global performance, results show an effect of education level ($F(1, 3033) = 12.66$, $p < .001$): academic participants present more correct responses (93.38%) than non-academic participants (83.34%)[4]. We also observed an effect of accentual pattern ($F(1, 3033) = 4.26$, $p < .05$). Post-hoc analyses show more correct responses for PP (90.60%) and O (91.64%) than for P (85.33) ($p < .05$).

[4] Given that age and gender were not similar across academic and non-academic participants, we ran a model with age and gender as predictors to make sure that these variables were not responsible for the differences observed between academic and non-academic participants. None of the variables (or their interaction) has an effect on the spelling response.

More interestingly, we note an interaction between the education level and the accentual pattern ($F(1, 3033) = 4.08$, $p < .05$). As can be seen in Figure 2, the difference between academic and non-academic is more important in PP than in P or O ($p < .05$). Moreover, academic participants present more correct responses for PP and O than for P, whereas non-academic participants present more correct responses for O than for PP or P. In other words, it seems that academic participants have more difficulties with P stimuli, while non-academic participants have more difficulties with PP and P stimuli.

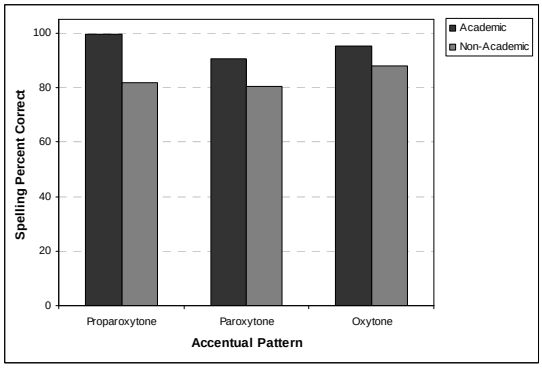


Figure 2: *Spelling percent correct as a function of accentual pattern and education level.*

3.1.3. Role of Stress perception

Interestingly, we observe an effect of the perception response on the spelling responses ($F(1, 3033) = 14.13$, $p < .001$). As can be seen in Figure 3, the stimuli with a correctly identified stress are more likely to be correctly written (i.e. with or without the written accent) than the incorrectly perceived stimuli.

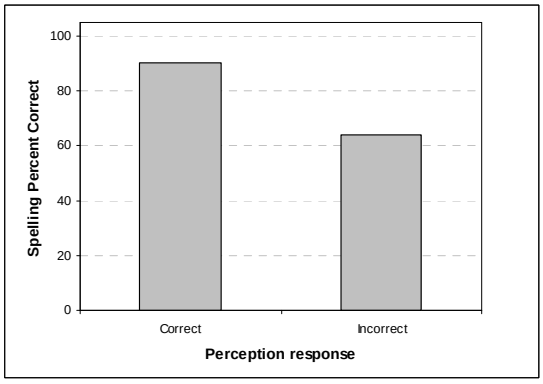


Figure 3: *Spelling percent correct as a function of the perception response.*

However, as shown in Figure 4, the effect of the perception response is modulated by the accentual pattern ($F(1, 3033) = 8.06$, $p < .001$). In fact, the effect of perception is smaller in P than in PP or O (independently of the education level) ($p < .05$). Finally, results show an interaction between the perception response and the education level ($F(1, 3033) = 17.28$, $p < .001$).

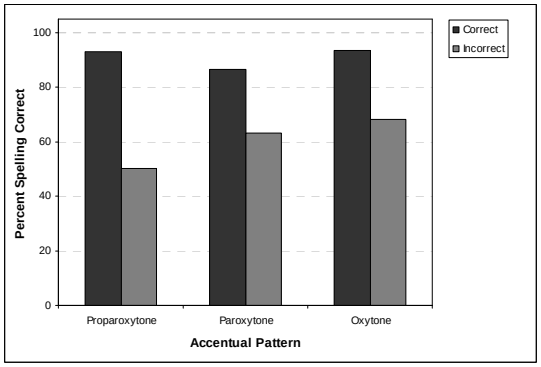


Figure 4: *Spelling percent correct as a function of the accentual pattern and the perception response.*

As can be seen in Figure 5, the difference between correct and incorrect perception responses is bigger in academic than in non-academic participants ($p < .05$). In other words, the perception of stress seems to play a more important role in academic than in non-academic participants.

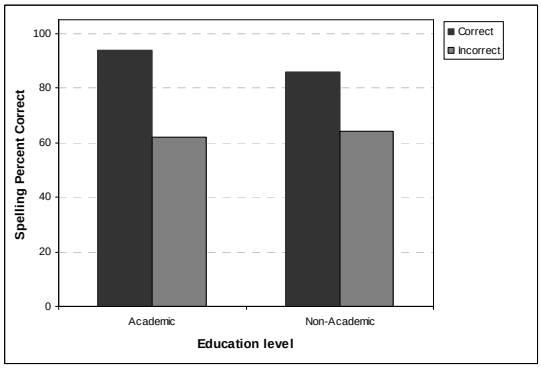


Figure 5: *Spelling percent correct as a function of the education level and the perception response.*

In order to explain the bigger importance of stress perception in academic than in non-academic participants, we ran an analysis with the perception response as the dependent variable and with the following predictors: the participant's education level (academic/non-academic), the stimulus lexical status (words/non-words), the stimulus accentual pattern (PP, P, O), the stimulus number of syllables. The results show, among other effects[1], an effect of education level: academic participants are better in identifying stress position than non-academic participants (98.63% and 88.54% respectively).

Figure 6 presents the spelling percent correct as a function of perception percent correct and education level. As can be seen, academic participants present less variability in their perception responses than non-academic participants (the standard deviation for academic participants is 1.65 and it is 15.73 for non-academic participants).

[1] Besides the effect of education level ($F(1, 3040) = 3.39$, $p < .001$), results of stress perception show an effect of lexical status, with more correct responses for words than for non-words ($F(1, 3040) = 5.29$, $p < .05$), an effect of the number of syllables, with more correct responses for disyllabic stimuli ($F(1, 3040) = 16.44$, $p < .001$), but no effect of accentual pattern.

Therefore, the fact that the academic participants' perception performance is not only higher, but also more invariable than the non-academic participants' performance may explain why perception plays a more important role in academic than in non-academic participants.

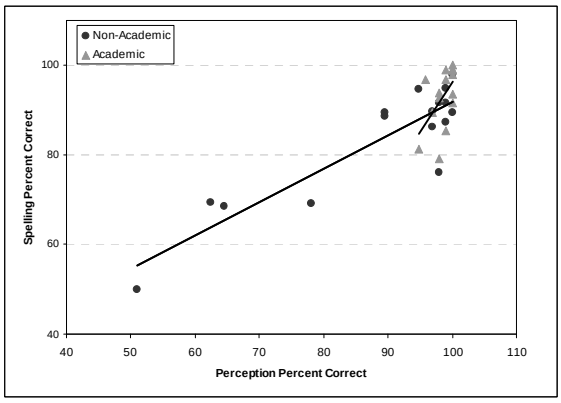


Figure 6: *Spelling percent correct as a function of perception percent correct and education level.*

3.2. Error analysis

In the error analysis, we examined whether the errors (N = 354) came from the addition of the written accent (henceafter "Added Written Accent"; e.g. *fluidéz* instead of *fluidez*) or from the absence of the written accent (henceafter "Missing Written Accent"; e.g. *carcel* instead of *cárcel*). Results show first that *Missing Written Accent* errors (N = 236) are more frequent than *Added Written Accent* errors (N = 118) ($\chi^2(1, N = 354) = 39.33$, $p < .001$). In other words, participants missed more written accents than they added.

Then, an analysis was run with the error type (Missing/Added Written Accent) as dependent variable and with the following factors: education level (academic/non-academic), lexical status (words/non-words) and accentual pattern (PP, P, O) and the interactions between these variables. No main effect was significant. Only the interaction between education level and lexical status was significant ($F(1, 350) = 6.40$, $p < .05$). Figure 7 presents the percent *Added Written Accent* error as a function of lexical status and education level.

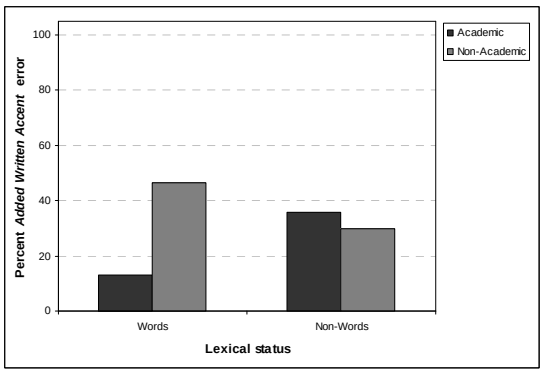


Figure 7: *Percent "Added Written Accent" error as a function of lexical status and education level.*

As can be seen, non-academic participants present more *Added Written Accent* errors than academic participants in words, while they present in non-words the same number of *Added Written Accent* errors as the academic participants. In other words, non-academic participants tend to erroneously add a written accent to a greater extent than academic participants, especially in real words.

4. General discussion

The aim of the investigation was to examine whether the adults' difficulty in placing the written accent in Spanish words was related to their ability in perceiving stress. We also took into account the following variables: the participant's education level, the stimulus lexical status, accentual pattern and length.

While results showed no effect of the stimulus length, they showed an effect of lexical status (i.e. more correct spellings for words than for non-words). This finding might suggest that the presence of the written accent is stored in the mental representation of the words. However, further research is still needed to confirm this hypothesis. An effect of education level was also observed, which indicates that academic participants mastered to a greater extent the Spanish written accentuation rules than non- academic participants. More interestingly, we found an effect of accentual pattern, with less correct responses for paroxytone stimuli than for proparoxytone or oxytone stimuli. We already mentioned that the paroxytone pattern is the default accentual pattern in Spanish. Yet, determining how frequently the paroxytone words require a written accent may explain why this pattern presents more errors. To do this, we analyzed a corpus composed of formal written Spanish texts (CODIMEP-CR-XXI). It contained 124'000 tokens, which corresponded to 11'370 types (i.e. different words). We found that 16% of the words have a written accent. Among those, 53% were oxytone, 31% were proparoxytone and 16% were paroxytone. It is important to note that, among the paroxytone words with written accent, only 15% corresponded to words ending in consonants other than -n and -s. (e.g. *dólar*), while 85% corresponded to words ending with hiatus (e.g. *día*). Yet, our paroxytone stimuli with written accent always showed this infrequent spelling (e.g. *líder*, *cárcel*, *fácil*, etc.). Moreover, taking into consideration that the default pattern in Spanish is paroxytone [1] and that only 16% of Spanish words require written accent, from which 85% are proparoxytone or oxytone (corpus CODIMEP-CR-XXI), we can assume that Spanish paroxytone words with written accent are much less frequent that paroxytone words without written accent. Yet, our stimuli included the same number of paroxytone items with and without written accent (words and non-words). This might explain why there are more errors in the spelling of paroxytone stimuli.

Finally, we observed an effect of the stress perception in the assignment of a written accent: stimuli with a correctly identified stress are more likely to be correctly written (i.e. with or without written accent) than the incorrectly perceived stimuli. However, this effect is weaker in paroxytone words and in non-academic participants. Taken together, these results confirm the hypothesis that the adults' difficulty in placing the written accent in Spanish is related to their ability in perceiving stress. In other words, there is a relationship between prosodic and written skills, as it was showed for reading in [8].

5. References

[1] Quilis, A., "Tratado de fonética y fonología españolas", 2nd. ed., Gredos, Madrid, 1999.

[2] Mesanza, J., "Palabras que peor escriben los alumnos (inventario caográfico)", Escuela Española, Madrid, 1990.

[3] Pujol, M., "Análisis de errores grafemáticos en textos libres de estudiantes de enseñanzas medias". Doctoral dissertation, Departament de Didàctica de la Llengua i la Literatura, Universitat de Barcelona, 1999. Online: http://hdl.handle.net/2445/41392, accessed on 4 Dic 2013.

[4] Pujol, M., "La ortografía", in: S. Torner and M. P. Battaner [Eds.], El corpus PAAU 1992: estudios descriptivos, textos y vocabularios, 29-65, Universitat Pompeu Fabra, Barcelona, 2005.

[5] Murillo Rojas, M., "Vocabulario cacográfico. Pautas para la enseñanza de la ortografía en la escuela primaria costarricense". Káñina, Revista de Artes y Letras de la Universidad de Costa Rica, 30(1):59-70, 2006.

[6] Jara Murillo, C. V., "COCAE: Corpus Cacográfico Adulto del Español de Costa Rica", Research Report, Project No. 745-B2-A13, Universidad de Costa Rica, 2013. Online: http://www.kerwa.ucr.ac.cr/handle/10669/8928, accessed on 4 Dec 2013.

[7] Henry, E., "Dificultades en la percepción del acento", Revista Lingüística Teórica y Aplicada 21, Concepción, Chile, 1983.

[8] Defior, S., Gutiérrez-Palma, N. and Cano-Marín, M. J., "Prosodic awareness skills and literacy acquisition", Journal of Psycholingusitic Research, 41: 285-294, 2012.

[9] Alameda, J. R. and Cuetos, F., "Diccionario de frecuencias de las unidades lingüísticas del castellano". Oviedo: Servicio de Publicaciones de la Universidad de Oviedo, 1995.

[10] Baayen, R. H., Davidson, D. J., and Bates, D. M., "Mixed effects modeling with crossed random effects for subjects and items", Journal of Memory and Language, 59: 390-412, 2008.

[11] Bates, D. M. and Sarkar, D., "lme4: Linear mixed-effects models using S4 classes", R package version 2.6., www.r-project.org, 2007.

Parameterization and automatic labeling of Hungarian intonation

Uwe D. Reichel[1], *Alexandra Markó*[2], *Katalin Mády*[3]

[1]Institute of Phonetics and Speech Processing, University of Munich, Germany
[2]Eötvös Loránd University Faculty of Humanities, Budapest, Hungary
[3]Hungarian Academy of Sciences, Budapest, Hungary
reichelu@phonetik.uni-muenchen.de, marko.alexandra@btk.elte.hu, mady@nytud.hu

Abstract

In Hungarian intonation research a common framework developed by Varga (2002; [1]) is to categorize the intonation within the domain of accent groups by *character contours*. We propose a linear parameterization of a subset of these contours derived from polynomial stylization. These parameters were used to train classification trees and support vector machines for contour prediction. Parameter extraction and training was carried out on the original F0 contours of spontaneous speech data as well as on three differently normalized variants suppressing fundamental frequency level and range effects. The highest accuracies were obtained for classification trees and F0 residuals after midline subtraction, but the overall performances were rather poor. Nevertheless, a significant improvement of the results was achieved by a Hidden Markov model to predict the correct label sequence from the partly erroneous classification output.

Index Terms: intonation, Hungarian, character contours, stylization, labeling

1. Introduction

An established approach in Hungarian intonation research is to describe fundamental frequency (F0) curves in terms of *character contours (CC)*. This framework was developed by Varga [2, 1] and follows the tradition of contour-based intonation representations [3, 4] that focus on the contour properties of F0 rather than treating it as a sequence of tone targets [5].

Varga [1, p. 33] defines a CC as a *"discrete, meaningful speech melody"* with a *"characteristic shape"*. Its domain is a syllable sequence consisting of an initial accented (*"major-stressed"* [1, p. 33]) syllable and all following syllables till the next accented one or till the end of the intonation phrase. We refer to this sequence by the term *accent group* (AG) in the following.

According to [1] in Hungarian eleven character contours (and an appended contour that is not necessarily related to an AG) can be distinguished. The nine main contours are illustrated in Figure 1. They can be divided into three major classes: *i. front falling* (left column) *ii. sustained* (middle column), and *iii. end-falling* (right column). The abstract meanings assigned to these major classes are *self-contained*, *forward-pointing*, and *yes-no interrogative*, respectively.

For contour-based as well as for tone-sequence intonation models numerous machine learning techniques have been developed to derive the intonation representation automatically from the signal. Among the established techniques to learn categorical intonation labels from acoustic features are neural networks [6], decision trees and predicate logic learning rules [7], Hidden Markov models [7, 8] classification and regression trees [9], support vector machines, and instance-based learning [10]. [11] and [12] extract intonation categories in an unsupervised bottom-up fashion by means of clustering. The extraction of non-categorical parametric intonation representations is generally treated as an optimization task in an analysis-by-synthesis framework to minimize the distance between the observed F0 contour and the contour generated by the parameterization [13, 14, 15].

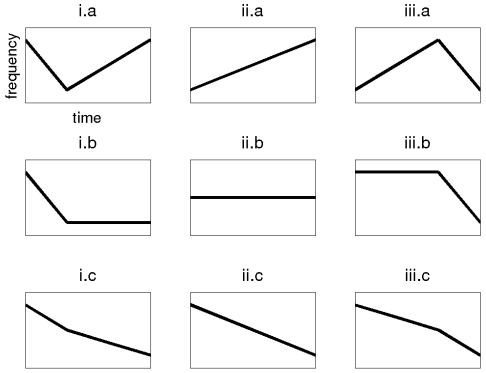


Figure 1: *Nine of the eleven character contours proposed by Varga [1, p. 43].*

Despite of this extensive literature on automatic intonation labeling, to our knowledge no approach has yet been published to extract the described character contours from the signal automatically. The aims of this study are thus to develop an appropriate F0 parameterization linking the signal to the character contour inventory and to train classifiers for automatic F0 labeling.

2. Data and preprocessing

The examined data consists of 50 Hungarian spontaneous speech utterances from collaborative dialogs by 10 Hungarian speakers. Each utterance forms a single intonation phrase (IP). Within each IP the accent groups were manually segmented following the definition of section 1, and the character contours were manually labeled by phonetic expert native speakers (the second and the third author of this study). In total, the data contains 140 AGs each linked to a character contour.

F0 was extracted by autocorrelation (Praat 5.3, sample rate

100 Hz). Voiceless utterance parts and F0 outliers were interpolated by piecewise cubic splines. The contour was then smoothed by Savitzky Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value. This base value was set to the F0 median below the 5th percentile of an utterance and serves to normalize F0 with respect to its overall level.

2.1. Normalization

To reduce the influence of F0 register on the local character contour shapes we generated three F0 residual variants. To capture the register in terms of its level and range [16] we fitted a base-, a mid-, and a topline for the IP. The baseline and the midline represent different aspects of the F0 level, whereas the F0 range information is provided by the time-varying span between the base- and topline. The robust fitting procedure that is motivated and explained in greater detail in [17] has already been applied for boundary strength [17] and prosodic phrase examinations [18]. The procedure is illustrated in the left panel of Figure 2 and consists of the following steps:

- A window of length 200 ms is shifted along the F0 vector with a stepsize of 10 ms.
- Within each window the F0 median is calculated
  - of the values below the 10th percentile for the baseline,
  - of the values above the 90th percentile for the topline, and
  - of all values for the midline.

  This gives 3 sequences of medians, one for the base-, the mid-, and the topline, respectively.
- Within each median sequence outliers are replaced by linear interpolation.
- Finally, for all three median sequences linear polynomials are fitted.

From this register stylization three F0 residuals were generated as can be seen in the right panel of Figure 2:

- the **baseline residual** by subtraction of the baseline from the F0 contour,
- the **midline residual** by subtraction of the midline from the F0 contour, and
- the **range residual** by normalizing each F0 value to the range between base- and topline at the corresponding position. These local reference points are set to 0 (baseline) and 1 (topline), respectively.

For the first two residuals the influence of register level is suppressed, whereas the impact of range is reduced for the third residual.

3. Parameterization

3.1. Method

Within each accent group segment we parametrized the original F0 contour as well as all residuals to derive a character contour representation. As can be seen in Figure 1 all contours can prototypically be represented by single lines or line pairs. We derived these lines the following way:

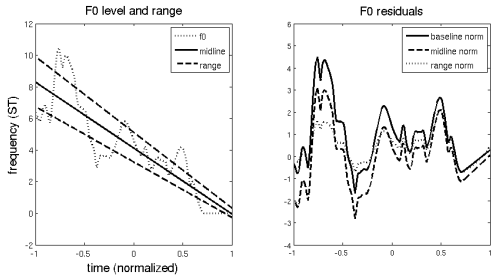


Figure 2: **Left:** *Extraction of a base-, mid- and topline within an intonation phrase to account for F0 level and range varying over time.* **Right:** *F0 residuals after level subtraction (baseline or midline) and range normalization.*

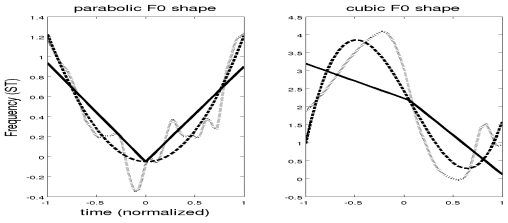


Figure 3: *Third-order polynomial stylization of F0 shapes.* **Left:** *Parabolic shape indicated by the absence of a turning point within the given time interval. The extreme value is selected as the split point* **Right:** *Cubic shape, where the turning point serves as the split point.*

- A third order polynomial was fitted to the F0 contour within an AG. Time was normalized to the interval from -1 to 1.
- The extreme values and the turning point of this polynomial within the normalized time interval were calculated from the polynomial's first and second derivative, respectively.
- If the polynomial neither contains a turning point nor an extreme value in the examined time interval, a **linear** F0 shape can be deduced that can sufficiently be characterized by a single line.
- If the polynomial contains only an extreme value but no turning point, this indicates a **parabolic** F0 shape in the examined time window. The extreme value is taken as the split point for subsequent line pair fitting. An stylization example for parabolic F0 shape is shown in the left panel of Figure 3.
- The existence of a turning point gives indication for a **cubic** shape. Since for these shapes two extreme values can occur in the examined interval, instead of choosing one of them in an ad-hoc manner, we select the turning point as the split point for subsequent line pair fitting. The right panel of Figure 3 shows a stylization example for cubic F0 shapes.
- The split point serves to divide the polynomial into two parts. Separately for the first and the second part a straight line is fitted passing through the split point by means of linear regression with zero intercept.

From this parameterization single line character contours can be described by the slope s of the regression line. Two-line

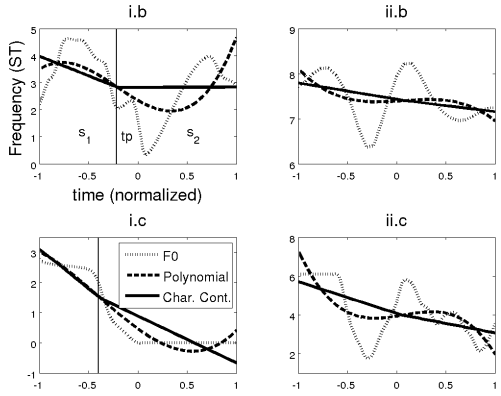


Figure 4: *F0 parameterization: Prototypical examples for each character contour.*

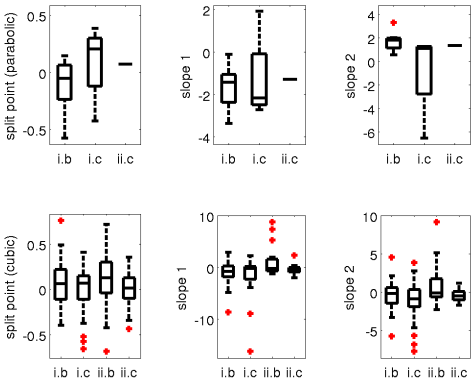


Figure 5: *Parameter distributions for the available character contour classes for parameterizations of parabolic (top row) and cubic shapes (bottom row).*

CCs are represented by a set of three features: sp: the position of the split point within the normalized time interval, s_1: the slope of the regression line form the beginning of the AG to the turning point, and s_2, the regression line from the turning point to the end of the AG. Since in our data only a single F0 shape was identified as linear it was subsumed to the set of parabolic shapes simply treating the line midpoint as the split point. Due to the lack of comparability between the different split point definitions the subsequent examinations and classifications were carried out separately for parabolic and cubic contours. Prototypical character contour stylization examples are given in Figure 4.

3.2. Relation between parameters and character contours

For the parabolic shape parameterizations none of the parameters differed significantly with respect to the character contour. For the cubic shape parameterization in contrast both slopes s_1 and s_2 showed significant differences across the contour classes for the original F0 contours and the residuals resulting from base- and midline subtraction (ANOVAs with s_1 resp. s_2 as dependent and character contour as independent variable. For s_1: $F[3, 113] = 4.08$, $p < 0.01$, and for s_2: $F[3, 113] = 3.51, p < 0.05$). With reference to Figure 1 one would expect more strongly negative slopes for $ii.c$ as opposed to $ii.b$ and for $i.c$ as opposed to $i.b$, but a Tukey-Kramer post-hoc test ($\alpha = 0.05$) only revealed significant differences for both slopes between the major classes i and ii. Nevertheless the expected tendencies can be seen in Figure 5.

4. Automatic Labeling

4.1. Features and preprocessing

We defined two different feature sets: a CC feature set as well as an extended set, which are presented in Table 1. For the extended feature set a sequential feature selection was carried out using the Matlab function *sequentialfs* to find the feature subset which is best in the sense of an optimality criterion. In our case the error to be minimized was calculated from the mean silhouette over all data points. We adopted this cluster evaluation measure to judge the goodness of separability of CC types by the respective feature combination. The mean silhouette $\hat{S}$, whose values range from -1 (bad) to 1 (good separability) was transferred to an error measure e ranging from 0 to 1 by $e = \frac{1-\hat{S}}{2}$. Subsequently the CC features as well as the selected features of the extended set were orthogonalized by means of a principal component analysis.

Table 1: *Feature sets for character contour classification.*

Feature	Description	CC	Extended
sp	position of split point	+	+
s_1	slope of first line	+	+
s_2	slope of second line	+	+
s_d	s_1-s_2	–	+
sp_y	F0 value of split point	–	+
$\text{coeff}_{0..3}$	polynomial coefficients	–	+

4.2. Classifiers

Separately for parabolic and cubic contours we trained classification trees (CART) [19] and support vector machines (SVM) [20] with a linear Kernel function to predict the character contour type from the respective feature set. For the training the Matlab functions *classregtree* and *svmtrain* with their default initializations were used.

4.3. Automatic label correction

In a postprocessing step we tried to improve the classification performance by treating the task of automatic labeling as a noisy channel problem: a correct label sequence (the expert annotation C) cannot be observed directly but only in form of a defective channel output O, which is given by the outputs of our classifiers. Thus label correction is basically the same as revealing the most probable hidden label sequence C that underlies the classifier output O: $\hat{C} = \arg\max_C[P(C|O)]$. We addressed this problem by Hidden Markov modeling (HMM). For the transition probabilities a linear interpolated uni- and bigram model was trained. Counts were smoothed by Good-Turing discounting.

5. Results

5.1. Classification accuracies

The accuracies for both classifiers CART and SVM, for the parabolic and cubic F0 contours and all its residuals, and for the CC and the extended feature set are presented in Tables 2 and 3. The accuracies were measured by means of leave-one-out evaluation. It turned out that:

- parabolic contours can be classified with higher accuracy than cubic contours. This can mainly be explained by the fact that parabolic contours were observed only for three of the four contour classes.
- F0 level normalization has a positive impact on CART performance while for the SVM performance rather range normalization is crucial.
- The best results were obtained by a CART trained on the reduced CC feature set and on midline residuals.

Table 2: *Character contour classification accuracies for classification trees.*

	CC feature set		extended feature set	
	parabolic	cubic	parabolic	cubic
original	69.23	29.91	53.85	23.08
baseline residual	76.92	30.77	69.23	36.75
midline residual	**84.62**	**43.59**	69.23	30.77
range residual	62.50	35.96	62.50	39.47

Table 3: *Character contour classification accuracies for support vector machines.*

	CC feature set		extended feature set	
	parabolic	cubic	parabolic	cubic
original	69.23	34.18	53.85	33.33
baseline residual	69.23	36.75	53.85	30.77
midline residual	69.23	35.04	69.23	32.48
range residual	75.00	41.22	68.75	42.98

5.2. Accuracies after error correction by HMMs

We trained HMMs to map the partly erroneous output of our best classifier (the CART classifying F0 midline residuals based on CC features) to the correct labels of the manual annotation.

In a tenfold cross validation it turned out that the performance could be improved with high significance (Mann-Whitney test, $z = -2.5809, p < 0.005$). The result is presented in Figure 6.

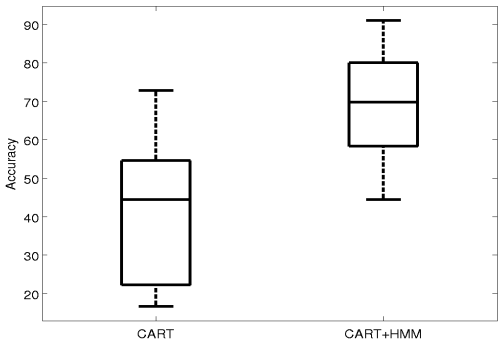


Figure 6: *The classification correction by means of an HMM leads to a significant accuracy improvement.*

6. Discussion and Conclusion

The findings of this study suggest that the application of the character contour framework on spontaneous speech is challenging. First, in our data only four character contour types had been observed which is only a subset of the contour inventory of [1]. This is partly to be explained by the shortcoming that our data does not contain questions, so that end-falling characters are very unlikely. But it also indicates that the other contour types are very unevenly distributed in spontaneous speech (*i.a* and *ii.a* from Figure 1 have not been observed at all). This uneven distribution in spontaneous speech was already observed by [1, p. 52]. In his data 60% of the contours were of type *ii.b*, whereas all *iii*-types taken together only occurred in less than 2% of the accent groups. Second, in spontaneous speech the extracted character contour parameters, slopes and turning point, overlap to a high extent across different contour classes so that it is difficult to relate a realized contour to its underlying prototype.

Principally, the parameterization proposed here is not strongly affected by the first drawback that not all contour types are contained in our data. The parameterization is extendable to the remaining contours, since for the observed contours their prototypes can be described by means of one or two line slopes.

Automatic labeling in contrast is heavily affected by the second drawback of parameter value overlap making it difficult to distinguish and identify character types in spontaneous speech. Also the usage of additional features as polynomial coefficients did not turn out to be gainful. Thus, the automatic labeling of character contours remains a difficult task.

Nevertheless, a significant improvement was achieved by postprocessing the classifier output with HMMs. From this finding we draw two conclusions. First, since postprocessing accounts for the left label context in terms of transition probabilities, one can infer that context plays a role for label assignment. Neither standard CARTs nor SVMs consider context, if it is not explicitly contained in the feature vectors. Context might be introduced by additional features (which requires more training data than available for this study) or by an appropriate normalization of the given features. The latter would be an interesting issue to be addressed in a follow-up study. Alternatively, one could directly use HMMs for intonation label assignment. However, using HMMs as a separate postprocessing classifier also has its benefits which brings us to our second conclusion: the learnability of correct labels from erroneous ones indicates that there are systematic correspondences between the classification output and reference labels. This finding might be of more general use for applications that automatically correct the output of labeling tools.

Finally, comparing the results between different classifiers and F0 normalization methods, one can see that normalization generally increases classification performance, but not in a uniform way for all classifiers. More systematic examinations are needed to get better insight into these complex relations.

7. References

[1] L. Varga, *Intonation and Stress: Evidence from Hungarian*. Hampshire, New York: Palgrave Macmillan, 2002.

[2] ——, "Prozodémák a magyar beszédben és jelölésük az intonációs átiratban," in *Műhelymunkák a nyelvészet és társtudományai köréből III*, MTA Nyelvtudományi Intézet, Budapest, 1987, pp. 91–119.

[3] D. Bolinger, "Intonation: Levels Versus Configurations," *Word*, vol. 7, pp. 199–210, 1951.

[4] M. A. K. Halliday, *Intonation and Grammar in British English*. Den Haag: Mouton, 1967.

[5] J. Pierrehumbert, "The phonology and phonetics of Englisch intonation," Ph.D. dissertation, MIT, Cambridge, MA, 1980.

[6] S. Ananthakrishnan and S. S. Narayanan, "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 1, pp. 216–228, 2008.

[7] S. Rapp, "Automatic labelling of German prosody," in *Proc. ICSLP*, 1998, pp. 1267–1270.

[8] C. Brindöpke, G. Fink, F. Kummert, and G. Sagerer, "A HMM-based recognition system for perceptive relevant pitch movements of spontaneous German speech," in *Proc. ICSLP*, Sydney, 1998, pp. 2895–2898.

[9] I. Bulyko and M. Ostendorf, "Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis," in *Proc. ICASSP*, 2001, pp. 781–784.

[10] A. Schweitzer and B. Möbius, "Experiments in Automatic Prosodic Labeling," in *Proc. Eurospeech*, Brighton, 2009, pp. 2515–2518.

[11] G. Möhler and A. Conkie, "Parametric modeling of intonation using vector quantization," in *Proc. 3rd ESCA Workshop on Speech Synthesis*, 1998, pp. 311–316.

[12] U. Reichel, "Automatisation of intonation modelling and its linguistic anchoring," in *Proc. Speech Prosody*, Shanghai, 2012, pp. 63–66.

[13] H. Mixdorff, "An Integrated Approach to Modeling German Prosody," Ph.D. dissertation, TU Dresden, 2002.

[14] H. Pfitzinger, H. Mixdorff, and J. Schwarz, "Comparison of Fujisaki-model extractors and F0 stylizers," in *Proc. Interspeech*, Brighton, 2009, pp. 2455–2458.

[15] P. Taylor, "Analysis and Synthesis of Intonation using the Tilt Model," *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, 2000.

[16] T. Rietveld and P. Vermillion, "Cues for Perceived Pitch Register," *Phonetica*, vol. 60, pp. 261–272, 2003.

[17] U. Reichel and K. Mády, "Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech," in *Elektronische Sprachsignalverarbeitung*, ser. Studientexte zur Sprachkommunikation, P. Wagner, Ed., vol. 65. Bielefeld: TUDpress, 2013, pp. 223–230.

[18] K. Mády, U. Reichel, and v. Beñuš, "Accentual phrase in languages with fixed word stress: a study on Hungarian and Slovak," in *Workshop Advancing Prosodic Transcription for Spoken Language Science and Technology II, Phonetics and Phonology in Iberia 2013*, Lisbon, 2013.

[19] L. Breiman, J. Friedman, C. Stone, and R. Olshen, *Classification and Regression Trees*. Pacific Grove, CA.: Wadsworth & Brooks, 1984.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, 1995.

Local and global convergence in the temporal domain in Polish task-oriented dialogue

Maciej Karpiński[1], Katarzyna Klessa[1], Agnieszka Czoska[2]

[1]Institute of Linguistics, [2]Institute of Psychology,
Adam Mickiewicz University in Poznań, Poland
maciej.karpinski@amu.edu.pl, klessa@amu.edu.pl, agaczoska@gmail.com

Abstract

Conversational parties tend to mutually adapt their communicative behaviour in a number of dimensions, from the level of physical aspects of speech signal and gesture, utterance properties, up to the level of mental representations. In the present study, an attempt is made to track the process of convergence in the temporal domain both as a global tendency and a local phenomenon. The material under study consists of two sets of task-oriented dialogues recorded with or without eye contact (telephone conversations) between the speakers. All the recordings were segmented into syllables and analysed in terms of speech rate and syllable timing pairwise variability (s*nPVI)* for each speaker as well as for the correlations between the speakers in each pair. Global convergence tendencies were proven to be weak but some influence of dialogue settings and gender was found. The results seem to support the hypotheses that the alignment-related processes remain under the influence of many factors related to the dialogue flow and cannot be modelled as simply incremental.

Index Terms: alignment, convergence, timing, dialogue, recording scenario

1. Background and previous studies

It is generally agreed that participants of dialogue tend to mutually accommodate, align and converge in various domains [1,2] that include the phonetic level of utterances [3,4,5,6,7], lexicon and syntax [8], postures, gestures and facial expressions [9,10,11] and, ultimately, mental models of situations [2,12].

Alignment is sometimes viewed as an almost automatic process based on priming [13] but it may be well driven by conscious communicative strategies [14]. Increased alignment has been shown to predict better communication efficiency, outcomes, and, presumably, better results of common task solving [15,16,17]. On the other hand, convergence has its limits: when mimicry goes too far it may result in the impression of parody. Moreover, in many communicative situations alignment may not work as a fully symmetrical process of increasing convergence between speech parameters of dialogue participants. Due to a higher social position, more assertiveness or some other factors, one of the dialogue parties may become the leader, while the other may decide to take a more subordinate role just not to inhibit the flow of communication [18].

The temporal aspects of dialogue interaction early attracted the attention of researchers, especially on the grounds of conversational analysis, where the precision in the arrangement of turn-taking has been often stressed [19]. More recently, formal approaches to turn-taking modelling have become more influential [20,21,22] and some more light has been shed on turn-boundary phenomena [7]. A number of studies have been devoted to the convergence of phonetic and especially of prosodic parameters of speech between conversational partners [23,4,24,25,26,28]. Nevertheless, many other aspects of the prosodic and temporal dynamics of dialogue remain less explored or, although extensively analysed, they still require more empirical evidence. Speech rate interdependencies in dialogue have been analysed in a limited number of studies, e.g. [3,5]. This type of prosodic alignment may be highly sensitive to various phenomena typical of spontaneous communication (hesitations, repairs, pauses, etc.) as well as the characteristics of the communication situation (external distractors) that can easily distort or disrupt it.

2. The aim of the study

According to the results of the abovementioned studies, the parties of dialogue tend to align locally or globally, and this process may also include convergence in the temporal domains, e.g., in the domain of speech rate and speech rhythm irregularity. The aim of the present study is to explore this type of convergence in the recordings of two types of Polish task-oriented dialogues (mutual visibility vs. lack of mutual visibility). Speech rate changes and speech rhythm irregularities may reflect important aspects of dialogue flow, including various dialogue fluency stages as well as task realisation stages and, in general, the quality of dialogue interaction [29].

3. Data and analysis

3.1. Recordings, segmentation and transcription

The recordings of task-oriented dialogues come from two Polish corpora: DiaGest2 corpus [30] and the Paralingua corpus [31]. Altogether, we analysed data based on the recordings of 40 voices, 20 from each of the corpora: 15 female speakers and 5 male speakers from the DiaGest2 corpus; 12 females and 8 males from the Paralingua corpus. All the recorded sessions were either conversations of two female speakers or a male and a female (no dialogues between two male interlocutors). In the DiaGest2 corpus, the task of the participants resolved itself in re-creating a figure made of paper by one of the participants (Instruction Follower, IF) according to instructions by the other (Instruction Giver, IG).

Only IG could see the original figure. IF was provided with all the materials necessary to re-construct the figure. IG and IF could see each other. All the IFs were females while there were five females and five males among the IGs. The speakers were given a time limit of 5 minutes to solve the task. The recordings from the Paralingua corpus dialogue were obtained based on a dialogue task in which the speakers were asked to find the differences between two pictures of a room. The task was to co-operate with the interlocutor in order to find as many differences as possible in the shortest possible time. The major differences between the tasks in DiaGest2 were the asymmetry (DiaGest2) vs. symmetry (Paralingua) of speakers' roles and mutual visibility (DiaGest2) vs. the lack of mutual visibility (Paralingua).

The material used for the present study was segmented into syllables. It was transcribed orthographically on the word level of segmentation and phonemically on the syllable level. Filled pauses were marked on a separate tier. Two software tools were used for transcription and annotation: Praat [32] (annotation and transcription of the DiaGest2 corpus) and Annotation Pro [33] (annotation and transcription of the Paralingua corpus as well as further annotation-mining and analyses of both corpora).

3.2. Analysis of speech rate changes and syllable duration variability

The data were explored using a "moving time window" approach, where a selected parameter was measured and averaged for each speaker within the window that was moving along the time axis. A similar method was used in [5,34] where additionally weights were used in the formula applied to measurements of accommodation of various acoustic/prosodic features, e.g., pitch, intensity or speech rate (thus using a weighted mean, where the interval durations were the weights). Four different window sizes were tested (5, 10, 30 and 60 seconds) in order to choose the optimum one, while the step was calculated as a proportion of the window size (33% in each case). Extremely narrow ones resulted in a substantial number of empty frames (i.e., not including any speech on one or both sides of the conversation) and thus did not support time-aligned analysis of dialogue turns. Applying very wide time windows helped to avoid this issue but, on the other hand, it might have concealed small-scale phenomena. As a consequence, the time window size was a result of a compromise. For most of the below calculations, the size of 30 seconds was applied as the smallest one that allowed to avoid empty frames.

Two timing-related parameters were taken into account. A "net" measure of speech rate (henceforth *NSR*) was calculated as the number of syllables within a time window (including those partially contained in the window as the respective fractions and excluding all kinds of pauses) divided by their total duration. Additionally, using a similar "moving time window" approach, a syllable-based *nPVI* (henceforth *snPVI*) [35] was calculated for each time window and each speaker in the dialogues under study. In case of the segments partially contained in a given window, their total durations were used in the *nPVI* formula. The *nPVI* was used with syllable-sized segments; cf. [36,37,38,39] as regards the definitions and plausibility of basic units for the analysis of pairwise variability.

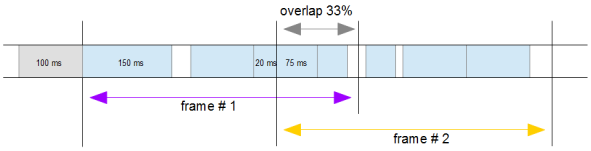


Figure 1. Segment rate moving average scheme

The transcripts prepared in Praat were converted to Annotation Pro native format and processed further using two Annotation Pro plug-ins specifically designed for the abovementioned calculations of *NSR* and *nPVI* (available from http://annotationpro.org/plugins/).

The results were exported as CSV files and analysed using SPSS statistical package (IBM Corp., 2012). Analyses included descriptive statistics calculated for all the speakers and dialogues, and regression analyses as well as some others, inspired by the data themselves.

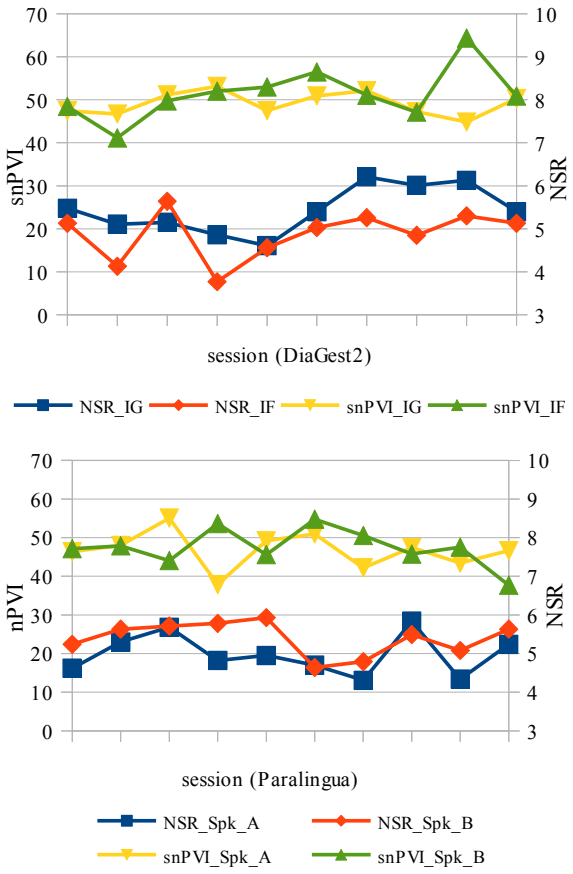


Figure 2. Mean normalised speech rate and mean snPVI for DiaGest2 (upper panel) and Paralingua (lower panel) dialogues

3.2.1. Speech rate analysis

The mean values of *NSR* calculated from 30-second time windows 5.23 (std.dev.=0.93, range=1.2–8.4) for DiaGest2 and 5.13 (std.dev.=1.01, range=3.2–7.4) for Paralingua. The difference between the mean values of the NSR for the two data sets turned out to be non-significant at p=0.05. DiaGest2 data were analysed additionally to test the difference between IG and IF speech rate (using Student's t-test). For the

difference was significant for both the 30-second time windows (IG mean=5.43, std.dev.=0.99; IF mean=4.88, std.dev.=1.19; p>0.0001) and the 60-second time windows data (IG mean=5.51, std.dev.=0.72; IF mean=4.92, std.dev.=0.86; p>0.0001).

Linear regression analysis was carried out for the data based on 10, 30 and 60 second time windows. While for the smallest window size no significant results have been found in the two corpora, with a 30 second window, four significant results of regression analysis were found in the DiaGest2 data and one in Paralingua. For the 60 second window, one significant result of regression was found in each corpus.

Significant results for the 30 second time window were found in dialogues 2 (R=0.314, p=0.024), 6 (R=0.629, p<0.001), 8 (R=0.415, p=0.007) and 9 (R=0.39, p=0.013) from the DiaGest2 data and dialogue 12 (R=0.432, p=0.006) from the Paralingua corpus. The significant results for the 60-second frame were found in different dialogues: dialogue 3 (R=0.617, p=0.021) from the DiaGest2 corpus and dialogue 7 (R=0.983, p=0.008) from the Paralingua corpus.

The results may suggest that alignment, understood as linear correlation between the speech rates of interlocutors, occurres rarely in this kind of task dialogues. On the other hand, when the phenomenon takes place, the correlation is very strong (98% in dialogue 7 from the Paralingua data) or medium (40%–60% in the other dialogues). Moreover, local alignment (30-second time window) occurred more often, than the global one. In this data alignment occurred more often in dialogues which participants had different roles (IG and IF). Although IF's speech rate showed a global tendency to be lower, the difference did not prevent the occurrence of alignment, but might even encourage it.

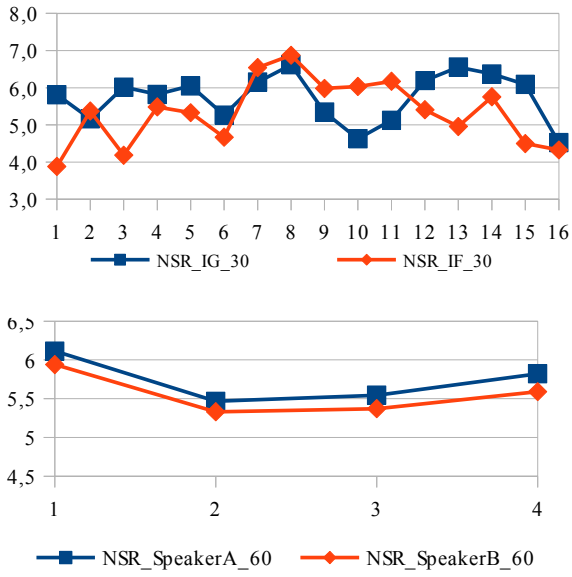


Figure 3. *NSR* changes in dialogue 6 from DiaGest2 corpus (30 second window; upper panel) and in dialogue 7 from Paralingua (60 second window; lower panel)

An additional Chi-squared test was used to test the hypothesis that alignment would be more frequent when interlocutors are of different sex. Significant and strong results were found for the DiaGest2 data alone (Pearson's chi-squared=23.333, p<0.001) in contrast with non-significant results for the Paralingua data (Pearson's chi-squared=1.667, p=0,4). The results were not significant for the 30-second time window data including both DiaGest2 and Paralingua results (Pearson's chi-squared=3.3, p=0.069). In DiaGest2 data three occurrences of alignment were found between participants of different sex (out of five mixed pairs), and only one (dialogue 2) in a dialogue between two women. This may suggest that there is a higher probability of the occurrence of the local alignment between interlocutors of different sex. When it comes to the 60-second time window, the two cases of alignment found occurred in dialogues between two women (due to the small number of alignment occurrence no Chi-squared test was performed). This result may suggest further differences between local and global alignment.

3.2.2. The analysis of snPVI

The mean value of *snPVI* (30-second time window) equalled to 50.6 (std.dev.=8.8, range=20.4–80.30) for DiaGest2 and 48.2 (std.dev,=6.1, range=35.2–70.1) for Paralingua. The difference between the mean values of *snPVI* between the DiaGest2 and Paralingua data turned out to be statistically significant at p<0.01. DiaGest2 data were analysed additionally to test the difference between IG and IF speech rate. For the difference was significant for the 30-second time windows (IG mean=49.22, std.dev.=6.58; IF mean=51.41, std.dev.=11.84; p=0.041) but non-significant for the 60-second time windows data (IG mean=48.27, std.dev.=9.15; IF mean=49.73, std.dev.=12.69; p=0.39).

For the values of the *snPVI* linear regression analysis was carried out for each pair of speakers using 30 and 60 second time windows. Results were found significant (p<0.05) for two pairs for each size of the time window in the DiaGest2 data and for three pairs at 30 second time window in the Paralingua data set.

The significant results for 30-second window were found in dialogues 6 (R=0.306, p=0.026) and 10 (R=0.338, p=0.014) from the DiaGest2 data and dialogues 2 (R=0.502, p=0.022), 14 (R=0.896, p=0.004) and 15 (R=0.816, p=0.005) from the Paralingua data. The significant results for the larger time window were found in dialogues 6 (R=0.86, p<0,001) and 7 (R=0.631, p=0.011) from the DiaGest2 data.

Again alignment occurred rarely, but the correlations were often medium (dialogue 2 from DiaGest2) or strong. Only two weak correlations were found within the local time window data (dialogues 6 and 10 from DiaGest2). Like previously, local alignment occurred more frequently than global and there was more alignment in the DiaGest2 corpus. The difference in the mean *snPVI* between IG and IF participants has disappeared for the 60-seconds time window data but prevailed for the local data.

Significant results (the occurrence of alignment) were found in dialogue 6 from the DiaGest data for both measures for the local time window, and in both windows for *snPVI*. This dialogue seems to be exceptional, since all the other dialogues show alignment only once across all the analyses.

Chi-squared tests were also performed for the *snPVI* alignment results. Significant results were found for the DiaGest2 data (Pearson's chi-squared=25.0, p<0.001 for 30-second frame, Pearson's chi-squared=25.0, p<0.001 for 60-

second frame). The results are the same, since alignment occurred in both cases within two mixed pair dialogues.

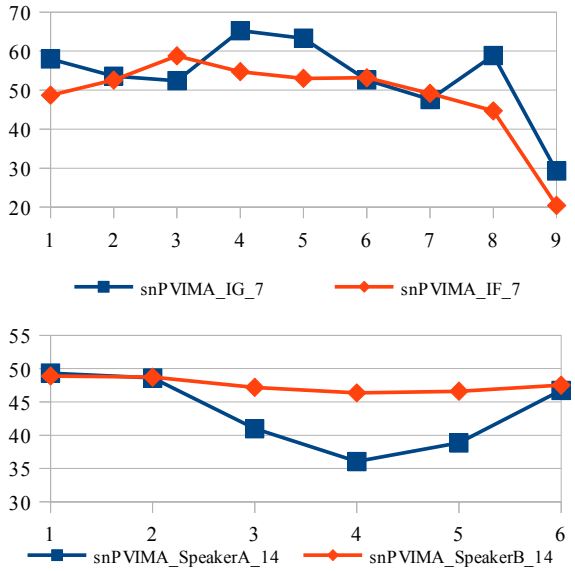


Figure 4. *snPVI* changes in dialogue 7 from DiaGest2 corpus at 60-second time window (upper panel) and in dialogue 14 from Paralingua corpus at 30-second time window (lower panel)

The results were not significant for the data from both corpora analysed together (Pearson's chi-squared=1.818, p=0.479 for 30-second frame, Pearson's chi-squared=2.716, p=0.127 for 60-second frame) or the Paralingua data separately (Pearson's chi-squared=1.27, p=1.0 for 30-second window). Results obtained from the DiaGest2 data may suggest a correlation between interlocutors sex and the probability of alignment. This suggestion may be strengthened by the fact, that two, out of three cases of alignment from the Paralingua data occurred in mixed pairs. However, the test results were non-significant for the whole data.

4. Conclusions and further work

In the present study, an attempt was made to explore and detect inter-speaker convergence in the dimensions of speech rate *(NSR)* and syllable timing pairwise variability *(snPVI)*. According to Kousidis [34], alignment in task-oriented dialogues may be much more difficult to detect. Nevertheless, such attempts have been made, e.g. [28], although results are not always clear-cut.

The overall speech rate means observed for both of the analysed corpora are similar (5.23 vs. 5.13 syll/sec) although the min-max range appeared to be wider in the DiaGest corpus (the overall rates are also in line with the rates for Polish normal reading rate [39]). Malisz [40] reported a higher overall mean around 6.9 syll/sec for Polish dialogues but using only fluent and coherent utterances with no unintelligible parts, false starts or hesitation markers while in the present work all utterances were used with the only exception of the segments labelled as filled pauses that were arbitrarily excluded from the analyses.

The results may suggest that the SR co-variability may be prevalently a local phenomenon that tends to occur most clearly within homogeneous, fluent stretches of conversation and may work very quickly and precisely as in the case of syllable boundary alignment [7]. On the other hand, it does not exclude the possibility of long-term, more permanent speech rate and syllable length variability accommodation even in the participants of task-oriented dialogues [23,34]. Verification of this claim, however, requires studies on larger corpora so that less prominent tendencies can be detected.

The influence of participants' gender on the level of *NSR* and *snPVI* convergence was also investigated. The results for the DiaGest data are coherent with Street's observations [3] and show less convergence in female-female pairs. As no influence was observed in the Paralingua data, one may hypothesize that the results may be somehow related to the mutual visibility condition. However, the factor of mutual visibility may somehow account for worse results in the timing-related convergence independently of gender matching.

In the DiaGest2, more manual activity (re-contructing a figure made of paper) was necessary while in Paralingua task, the participants had to scrutinize images (in order to find differences). It is difficult to judge which of the tasks tended to consume more attention and re-direct it from the partner to an object.

The differences in the overall ranges of *NSR* and *snPVI* might reflect the discrepancies in timing strategies applied depending on the presence or absence of "symmetry" in the task-solving roles. These observations are aimed to be further investigated also using other methods of syllable duration analysis, e.g. those accounting for deceleration/acceleration patterns [41,42] as well as more detailed analysis of pausing schemes. Note that although filled pauses were excluded from the present computations, it was observed that in DiaGest the number of pauses was considerably higher for IGs than for IFs while in Paralingua the number of pauses produced by the interlocutors was similar even though the total number of pauses was similar in both corpora. This, however, also requires further analyses as the proportion of the total speaking time between the dialogue participants should be taken into account.

Further research will include the analysis of alignment of timing-related phenomena in shorter stretches of speech between major changes of the conversation stages or topic shifts, fluency breakdowns and other phenomena that may significantly influence (hinder) the process of alignment between speakers.

Acknowledgements

Recording of the Paralingua database was supported from the financial resources for science in the years 2010–2012 as a development project (O R00 0170 12). DiaGest2 corpus comes from DiaGest2 project (N N104 010337, 2009-2010).

5. References

[1] Giles, H., Taylor, D. M. and Bourhis, R. "Towards a theory of interpersonal accommodation through language: some canadian data", Language in Society, pp. 177–192, 2010.

[2] Garrod, S., & Pickering, M., Why is conversation so easy? Trends in Cognitive Sciences, 8, pp. 8 – 11, 2004.

[3] Street, R. L. “Speech convergence and evaluation in fact-finding interviews. Human Communication Research”, Vol. 11, No. 2, pp. 139–169, 1984.
[4] Pardo, J. “On phonetic convergence during conversational interaction”, Journal of the Acoustical Society of America, 119, (2382-2393), 2006.
[5] Kousidis, S., “A Study of Accomodation of Prosodic and Temporal Features in Spoken Dialogues in View of Speech Technology Applications”, Doctoral Thesis. Dublin Institute of Technology, 2010.
[6] Kim, M., Horton, W. S. & Bradlow, A. R., “Phonetic convergence in spontaneous conversations as a function of interlocutor language distance”, Laboratory Phonology, 2, 125–156, 2011.
[7] Włodarczak, M., Simko, J., Wagner, P., “Syllable boundary effect: Temporal entrainment in overlapped speech”, Proceedings of Speech Prosody 2012.
[8] Branigan, H., Pickering, M., & Cleland, A., Syntactic co-ordination in dialogue. Cognition, 75, pp.13 – 25, 2000.
[9] Parrill, F., & Kimbara, I., “Seeing and hearing double: The influence of mimicry in speech and gesture on observers”, Journal of Nonverbal Behavior, 30, pp. 157–166, 2006.
[10] Mol, L., Krahmer, E., Maes, A., & Swerts, M. “Adaptation in gesture: Converging hands or converging minds?”, Journal of Memory and Language, 66, pp. 249–264, 2012.
[11] Bergmann K, Kopp, S., “Gestural Alignment in Natural Dialogue”, Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012), Cooper RP, Peebles D, Miyake N (Eds); Austin, TX: Cognitive Science Society: pp. 1326–1331, 2012.
[12] Zwaan, R. A. & Radvansky, G. A., “Situation models in language comprehension and memory”, Psychological Bulletin, 123, 162–185, 1998.
[13] Pickering, M. J. & Garrod, S., “Toward a mechanistic psychology of dialogue”, Behavioral and Brain Sciences, vol. 27, pp. 169–226, 2004.
[14] Krauss, R. M., Pardo, J. S., “Is alignment always the result of automatic priming?”, Behavioral and Brain Sciences, 27(02), pp. 203–204, 2004.
[15] Porzel, R., Scheffler, A. & Malaka, R. “How entrainment increases dialogical efficiency”, Proceedings of Workshop on Effective Multimodal Dialogue Interfaces, Sydney, 2006.
[16] Pickering, M. J. & Garrod, S., “Alignment as the Basis for Successful Communication”, Research on Language and Computation, Volume 4, Issue 2–3, pp. 203 – 228, 2006.
[17] Reitter, D. & Moore, J. D., “Predicting success in dialogue”, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 808–815, 2007.
[18] Karpiński, M., “New Challenges in Psycholinguistics: Interactivity and Alignment in Interpersonal Communication”, Festschrift for Prof. Piotra Łobacz (in print).
[19] Sacks, H., Schegloff, E., Jefferson, G., A simplest systematics for the organization of turn-taking for conversation, Language, vol. 50, 4, 1974.
[20] Wilson, M. & Wilson, T. P. “An oscillator model of the timing of turn taking”, Psychonomic Bulletin and Review, vol. 12, no. 6, pp. 957–968, 2005.
[21] O’Dell, M., Nieminen,T., “Coupled Oscillator Model for Speech Timing Overview and Examples”, Nordic Prosody. Proceedings of the Xth Conference, 2009.
[22] O’Dell, M., Lennes, M., Werner, S., Nieminen, T., “Looking for Rhythms in Conversational Speech”, Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken, pp. 1201—1204, 2007.
[23] Buder, E. H. & Eriksson, A., “Time-series analysis of conversational prosody for the identification of rhythmic units”, Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco, 1–7. August, 1999.
[24] Edlund, J., Heldner, M. & J. Hirschberg, “Pause and gap length in face-to-face interaction”, Proceedings of Interspeech, pp. 2779–2782, 2009.
[25] Vaughan, B., “Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement”, Proceedings of Interspeech, pp. 1865–1867, 2011.
[26] De Looze, C. and Rauzy, S., “Measuring speakers’ similarity in speech by means of prosodic cues: methods and potential”, Proceedings of Interspeech, pp. 1393–1396, 2011.
[27] Levitan, R., Hirschberg, J., “Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions”, In: Proceedings of Interspeech, Florence, Italy, pp. 3081–3084, 2011.
[28] Truong, K. P. & Heylen, D., “Measuring prosodic alignment in cooperative task-based conversations”, Proceedings of Interspeech, 9-13 September 2012.
[29] Manson, J.H., Bryant, G.A., Gervais, M.M., Kline, M. A., Convergence of speech rate in conversation predicts cooperation. Evolution and Human Behavior, 34, 6, pp. 419–426, 2013.
[30] Karpiński, M. & Jarmołowicz-Nowikow, E. “Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances”, Proc. Speech Prosody, Chicago, 2010.
[31] Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M., “Paralingua – a new speech corpus for the studies of paralinguistic features”, Procedia – Social and Behavioral Science, vol. 95, pp. 48–58, 2013.
[32] Boersma, P. & Weenink, D., Praat: doing phonetics by computer (Version 5.1.05) [Computer program]. Available from http://www.praat.org/, accessed in November 2013.
[33] Klessa, K., Karpiński, M., Wagner, A., Annotation Pro – a new software tool for annotation of linguistic and paralinguistic features. Proceeding of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013.
[34] Kousidis, S., Dorran, D., Wang, Y., B. Vaughan, Cullen, C., Campbell, D., McDonnell, C. and E. Coyle, “Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues”, Proceedings of Interspeech, pp. 1692–1695, 2008.
[35] Grabe, E. & Low, E. L. “Durational variability in speech and the rhythm class hypothesis”, In C. Gussenhoven & N. Warner (Eds.), Laboratory Phonology 7, pp. 515–546, 2002.
[36] Asu, E.L., Nolan, F. "Estonian and English rhythm: a two-dimensional quantification based on syllables and feet." Proc. Speech Prosody, Dresden, Germany, 2006.
[37] Barry, W.J.; Andreeva, B.; Russo, M.; Dimitrova, S.; Kostadinova, T., "Do rhythm measures tell us anything about language type?" Proc. of the 15th ICPhS. Barcelona, 2693-2696, 2003.
[38] Deterding, D. The measurement of rhythm: a comparison of singapore and british english. Journal of Phonetics, 29:217–230, 2001.
[39] Yu, J., Gibbon, D., Klessa, K. “Computational annotation-mining of syllable durations in speech varieties”, to appear in Proc. Speech Prosody, Dublin, Ireland, 20-23 May 2014.
[40] Malisz, Z., “Speech rhythm variability in Polish and English: A study of interaction between rhythmic levels”, PhD Thesis, Faculty of English, Adam Mickiewicz University, 2013.
[41] Gibbon, D., “TGA: a web tool for Time Group Analysis”. Proc.Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix en Provence, 2013.
[42] Klessa, K. and Gibbon, D. “Annotation Pro + TGA: automation of speech timing analysis”. To appear in Proceedings of Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 26-31 May 2014.

Speech and song synchronization: A comparative study

Beatriz Raposo de Medeiros[1], *Fred Cummins*[2]

[1]Universidade de São Paulo
[2]UCD School of Computer Science and Informatics, University College Dublin
biarm@usp.br fred.cummins@ucd.ie

Abstract

Does synchronization among speakers or singers require the presence of a beat? Is an implied underlying pulse or meter relevant? We set out to explore synchronization among speakers and singers as they speak or sing a variety of texts. We compare metrically strong nursery rhymes with non-metered prose. We compare singing in genres with two very different types of rhythm (samba and rock), and we compare sung and spoken versions of texts. In each case, we ask whether the rhythmic qualities of the texts facilitate synchronization. The metrical structure of the nursery rhyme does not facilitate synchronization compared to prose, while the simple beat of rock music does help. Further comparisons are provided in the text.

Index Terms: synchronous speech, song, syncopation, stress timing

1. Introduction

The ability of speakers to synchronize when reciting a common text is clearly seen in the ubiquitous practices of protest and prayer worldwide. Many such texts are over-practiced, as in the mantra-like repetition found in the Catholic rosary, or the texts are very short and are repeated rhythmically, as in most protest chants. In the former case, over-practice may help to support synchronization as the temporal patterns, while somewhat irregular, become predictable by virtue of familiarity. In the second, the speech borders on the musical through the use of emphatic beats, and perhaps also simply through repetition, as in the speech-to-song illusion [1].

Studies of synchronization among speakers have hitherto used the kind of pragmatically vacuous text so familiar from such classic corpora as the TIMIT or CSLU speaker recognition corpora [2, 3, 4]. While these studies have revealed much about the formal characteristics of synchronized speaking, much remains to be explored in assessing the influence of one or other text type, and the influence of music-like regularities in the process of synchronization. The role of periodicity and meter in particular warrant attention, as most theories of the temporal control of action would suggest that periodicity is beneficial, or even necessary, for synchronization [5], and yet prior studies have verified that synchronization of speaking is possible in the absence of any demonstrable periodic structure [6, 4].

We adopt a stance with respect to speech that sees it as temporally structured, coordinated movement. In this, our explorations lie within a long tradition stretching back to Stetson who famously characterised speech as "movement made audible" [7]. We also adopt a dynamical perspective, viewing synchronization among speakers as a form of entrainment [8, 4]. If we use a relatively strict definition of synchronization as "doing the same thing at the same time", there are relatively few truly synchronized behaviours. These include military marching, some sports such as swimming, rowing, trampolining and diving, some forms of dancing, and music making in unison. All of these activities have at least one of the following two features, and some have both: There may be a clearly perceptible beat or isochrony to the behaviour, and/or the temporal evolution of the activity is strongly scaffolded by inertial, elastic, or gravitational constraints. Synchronous speaking, we note, not only frequently lacks overt periodicity, but it is also achieved in the absence of any such strong physical scaffolding [9].

While the use of a dynamical systems vocabulary provides us with a rich set of tools for approaching such coordinated behaviour within and between speakers, it leaves us with something of a conundrum, as synchronization among speakers can take place in the absence of any clear periodic structure [4], although, as noted, group chants frequently tend towards music-like repetition. We are therefore motivated to explore synchronization phenomena in which we vary the underlying temporal anchoring of the utterances, including utterances that are clearly periodic in underlying form, and those that are clearly non-periodic. To this end, we here explore both sung and spoken texts, with varying amounts of simple periodicity. A basic question we pose is whether periodicity actually facilitates synchronization, as a common-sense intuition suggests it ought.

The work presented here examines synchronization among speakers as we vary the speech material being spoken. In a first comparison, we look at synchronization of metrically regular versus metrically irregular speech, using a prose text and a nursery rhyme as central examples. In the second comparison, we use songs, shorn of accompanying music. We compare the synchronization in samba with synchronization found in rock. We also obtain spoken versions of the sung texts to allow a direct comparison of speech and song.

In surveying the large and poorly mapped lands between speech and song, this study is necessarily exploratory in spirit. The essential questions to be addressed are these:

- Will synchronization be facilitated by the presence of strong metrical structure in spoken texts (Nursery rhyme versus prose)
- Will rhythmic complexity affect the degree of synchronization observed (rock versus samba)
- Will synchronization be facilitated by the presence of an underlying, implied, musical beat (singing versus speaking)

2. Different Types of Speech

Four types of source text in Brazilian Portuguese provide the material of this study: one prose text, a nursery rhyme, a samba song and a rock song. The text in prose was extracted from a

novel (*Um Sopro de Vida (A Breath of Life)*, by Clarice Lispector) and was chosen for possessing short and long sentences, as well as short and long words and topicalization, aspects commonly founded in oral speech. Prose's principal feature is to offer an irregular sequence of accents without any metrical structure. We use a nursery rhyme as a form of poetry in which metrical structure is more regular and in which beat expectations are maximally strong. Nursery rhyme recitation in group is a very common situation in infant-caretaker play in many cultures.

As regards the songs, we chose two different rhythmical types: a rock song (*Aluga-se*) and a samba song (*Preciso me encontrar*)[1]. For each song we obtained both a sung version and a spoken version, in order to see whether the musical meter would facilitate synchronization. We perceive, intuitively, that the rhythm of samba differs from that of rock, marking them as quite distinct. We refer to the rhythm of samba as syncopated, which positions the genre within a family of Latin American rhythms (e.g. salsa, habanera). In the case of rock, syncopation is not obligatory, and even, in some cases, this is not desirable (e.g. heavy metal).

Singing is understood in the present study, as well as in a previous study [10], as a specific variant of speech we can call sung speech. Despite being relatively rare among linguistic studies, comparisons between singing and speaking can raise interesting questions about how apparently very different systems, such as language and music, fuse so well in song.

3. Singing as a Type of Speech

Singing is a ubiquitous phenomenon, though its function can greatly vary among human groups. Popular song is widespread in Western culture and, broadly speaking, is consumed as entertainment. In the particular case of Brazilian song, there are many types of songs that could be categorized by theme or rhythm, however the present study will focus on only two types: samba and rock. Popular song is the most prevalent genre in Brazil and is profoundly integrated into everyday social life. It is no overstatement to say that the song is Brazil's music. A typical feature of several types of Brazilian song is singing together: From the very popular rural work songs to the urban *rodas de samba*[2]. The first samba composers, those from the early 20th century, did not have any formal musical education, and so did not read or write scores. Samba composition was oral, writing down only the lyrics and repeating the melody many times in order to memorize it. Some Brazilian song scholars have suggested that the melody of samba bears the hallmarks of the intonation patterns of Brazilian Portuguese, though this remains largely untested. It is a point we revisit in the discussion.

Singing is a kind of hybrid phenomena in which speech and music meet. The songs chosen for use in this study are from two different rhythmic genres: samba and rock. Despite their differences, both genres allow an easy fusion of text and music. The principal difference lies in how the rhythmic phrases in these songs are related to the underlying stream of pulses, as discussed in the next section. Comparisons between speech and singing may shed a light on both segmental and prosodic characteristics of speech, showing how musical constraints adapt to language constraints and vice-versa [11, 12, 10, 13]. In this sense, it appears entirely plausible to consider the song a kind of speech.

[1]The rock song *Aluga-se*, by Raul Seixas and the samba song *Preciso me encontrar*, by Candeia are very popular in Brazil.

[2]*Roda de Samba*: name given to a *samba* session, where people sit in circle or around tables, playing instruments and singing together.

Fig. 1 illustrates some of the radical temporal differences found between the sung and spoken manifestations of the same text. The extreme prolongation of the final word only makes sense within a musical context.

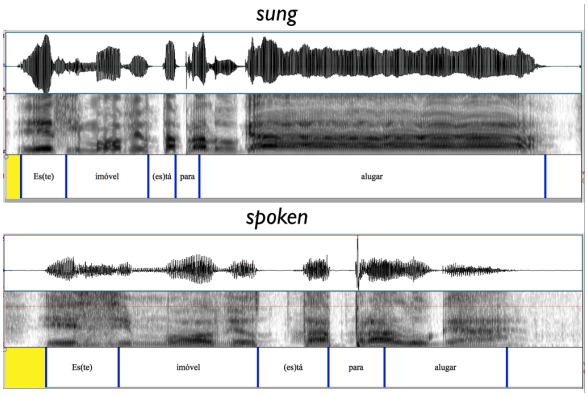


Figure 1: Samples of sung (top) and spoken (bottom) utterances from the rock song. Timescales differ between panels.

3.1. Different Song Rhythms

In our exploratory study here, we consider the relation between speech timing and metrical structure in both a poetic and musical context. Furthermore, we enrich the exploration of the relation between speech and musical rhythm by considering two types of musical rhythm: a simple 4/4 rock rhythm and a more complex samba rhythm.

A samba meter typically uses two beats per measure, and is thus a binary rhythm. When samba is transcribed, the conventional time signature is 2/4. Around these two beats, an off-beat system is built that is traditionally and commonly known as the samba syncopated rhythm. Specifically, there is a recurrent accent displacement that prevents a note from aligning with the second beat of the bar. This shift is caused by an sixteenth-note (semi quaver) of short duration that occurs just before the second beat, extending into and beyond the second beat. Another possible (and very frequent) accent shift is created by a bar-final sixteenth-note elongated and extending into the beginning of the following bar. Both accent shifts may be seen in Bar 3 of the score in Fig. 2.

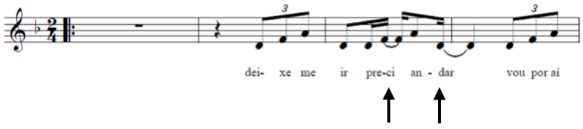


Figure 2: First four bars of Preciso me encontrar (*I need to find myself*). Brazilian *samba* by Candeia. Bar 3 depicts the typical rhythmic gestalt of samba and two common kinds of accent shifts.

As with other syncopated rhythm structures found in music, samba off-beats, as shown here, constitute the scaffold of a complex rhythm, as it can be characterized as shifting the initially proposed accent in a very specific way. There is some controversy as to whether the accent shifts found in Brazilian music stem from an African origin, or whether they are better

understood as a Westernization of the complex polyrhythms of the African sources [14].

It would be overly simplistic to describe rock rhythms as non-syncopated. Syncopation is not unusual in rock. It is a well-known and salient characteristic of related genres such as ska and reggae, and is found to some degree in many mainstream examples of rock music [15]. However, a great deal of rock music displays a fixed 4-beats-to-the-bar meter, in which musical notes are aligned with the beats, strong accents occur in metrically strong positions, and the position of the beats is clearly signalled by the drum track (See Fig. 3). Where samba has an obligatory syncopation, it remains a stylistic option in rock.

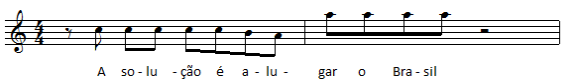


Figure 3: First two bars of Aluga-se (*For rent*). Brazilian rock by Raul Seixas.

4. Methods

Ten dyads were recorded, of which five were all female, three were male-male, and two were mixed sex. Subjects were aged between 25 and 45. All subjects reported no known problems with speaking or hearing. All were self-professed competent singers, and all were native speakers of Brazilian Portuguese.

Six texts provided six different experimental conditions. Texts are described in Section 2 above, and include a nursery rhyme, a prose excerpt, a sung samba song, a sung rock song, and spoken versions of the samba and rock lyrics. For each text, subjects were recorded in pairs, and were asked to remain in synchrony with one another. Choice of key and tempo was left to the subjects themselves. All singing/speaking was unaccompanied by music or an overt beat of any sort. Prior to recording, subjects signed an informed consent and read a text in prose (not recorded) in order to practice. Recordings were made using head mounted microphones (Shure SM10A) connected to a Marantz (PMD 661). Subjects stood facing each other inside the booth (Whisperroom 4872S) 1.3 meters from each other and began speaking/singing after the words "One, two, three. Ok" said by the researcher.

All recordings were segmented into sentence-length units: 7 for prose, 6 for the nursery rhyme, 13 for samba (sung and spoken) and 18 for rock (ditto). All 750 dyadic recordings were used in the subsequent analysis. The longest sentence belonged to prose, containing 36 syllables, and the shortest sentence was a rock verse, having 4 syllables. None of the sentences exceeded 11 seconds in duration.

A quantitative estimate of asynchrony was computed for each sentence using the method introduced in [4]. Two time aligned utterances are compared. Each is first represented as a sequence of Mel-Frequency Cepstral Coefficients. The sequences of MFCC vectors are then subjected to time warping, and the amount of warping necessary to map one utterance onto the other provides a quantitative estimate of the asynchrony between them. The measure is normalised by the number of time windows in the sequences, so that asynchrony values for shorter and longer sentences are comparable. The algorithm has been found to be most reliable when the calculation of the amount of warping, as indexed by the area under the warping curve, is restricted to voiced portions of the speech [4].

5. Results

Asynchrony values are not distributed normally, but are skewed right, due, in part, to some relatively large outliers. All asynchrony values were therefore first log transformed. Fig. 4 shows the distribution of asynchrony scores for the six conditions. The units of the asynchrony calculation are derived from the area under the warping function.

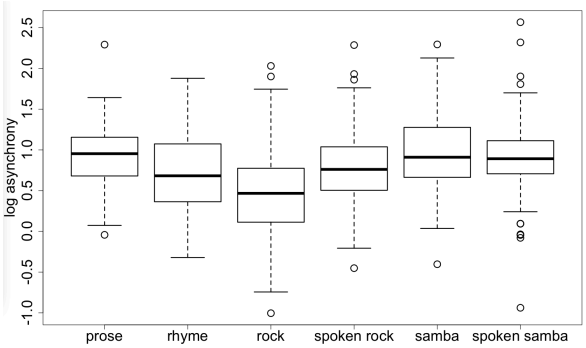


Figure 4: Boxplot of the distribution of log-transformed asynchrony scores for six different texts.

We are interested in specific planned comparisons that are of theoretical interest. The choice of comparisons is driven by our curiosity about the role of an underlying beat which may be pronounced and unambiguous (rock, nursery rhyme), present but subtle (samba) or absent (prose). Furthermore, we have available to us a comparison between sung and spoken lyrics within each genre. For each comparison, we conduct a simple t-test, and all t-tests are subject to a conservative Bonferroni correction to protect family-wise error rates.

The first comparison of interest is between prose and the nursery rhyme. Unexpectedly, the small difference in asynchrony observed between these two conditions was not significant (t(116)=2.4, n.s.), so there was no clear benefit of the regular meter in the nursery rhyme compared to the prose.

The next comparison examines synchronization in the sung versions of the rock song and the samba. Here, there is a marked difference. t(294)=8.1, $p < .001$. As expected, synchronization is greater in the rock condition than the samba, though we withhold interpretation of these results until the discussion.

The rock song was both sung, and spoken. The difference between these two is also highly significant, and as expected, synchrony is greater in the sung version. t(345)=5.8, $p < .001$. A similar comparison for the samba did not yield any difference between the sung and spoken conditions, however. t(255)=1.0, n.s.

A final comparison was done to look at the two conditions that contain the clearest metrical structure: the sung rock song and the spoken nursery rhyme. Here, a small difference was significant after correction of the p-value: t(118)=3.4, $p < .05$. Synchrony was greater for the rock song than the nursery rhyme.

6. Discussion

A naive assumption that we set out to test is that periodicity, overt or implied, would facilitate synchronization among speak-

ers/singers. Although this hypothesis seems to be entirely in accord with common sense, there is room for doubt. The remarkable ability of speakers to synchronize in the absence of any overt or implied periodicity has now been well documented [6, 4]. Most strongly synchronized activities that humans engage in make use of external constraints to facilitate synchronization, and these take the form of a regular pulse and/or the structuring of the behaviour through the presence of strong inertial or gravitational constraints [9]. Synchronous speech exhibits neither property, while singing without musical accompaniment, as here, makes use of an implied beat.

To start with the spoken domain proper, we were somewhat surprised to find that the regular meter of the nursery rhyme did not seem to facilitate synchronization compared with prose reading. This result is consistent with past findings that speech synchronization does not require periodicity, but is at odds with the simple hypothesis that periodicity will necessarily facilitate staying in time with one another.

The picture changes when we compare rock to samba. Now the strong and clear periodicity underlying the rock produces a clear advantage compared with the more fluid and complex rhythms of samba. The underlying musical pulse also ensures that the sung rock is considerably more synchronous than the spoken version. The same comparison for the samba yields no advantage at all for the musical, sung version.

Song lyrics are conventionally sung, and embedded within the temporal structure of the song. It is possible that asking singers to speak lyrics instead of singing them may generate some confusion as to just how much of the musical structure to reproduce. For example, Fig. 1 shows a sung phrase in which the final word is massively prolonged. Subjects may have been uncertain about whether to reproduce such temporal effects in speaking. One might argue that any such uncertainty due to the instructions ought to influence rock and samba productions alike, and this is clearly not the case. However an alternative account might argue that there is a more intimate link between the rhythms of samba and those of speech, so that the kind of gross temporal exaggeration found in rock (see Fig. 1) is less likely to occur. While it does not settle the matter, the present investigation opens a potential empirical route of approach to such a discussion.

Both the sung rock text, and the spoken nursery rhyme text are metrically structured, yet the presence of an underlying meter does not facilitate synchronization in the same way for the two genres. Further investigation of the differing manifestations of temporal structure in speech and singing will have to be sensitive, not only to meter, but also to the conventions of the genre. Nursery rhymes, it appears, belong squarely in the speech camp where temporal expectations are very different from the musical domain.

The sampling of texts presented in this study is neither comprehensive, nor even representative of the many differentiations one could make in the grey area between speech and music. However, they are sufficiently diverse to illustrate some important characteristics of the relation between overt temporal form and underlying metricality and structure. Perhaps the clearest message to be gleaned is that periodicity by itself is neither essential to, nor required for, highly synchronized coordinated movement among simultaneous speakers. Much of the legacy treatment of rhythm in the study of human behaviour has tended to conflate the distinct concepts of periodicity and rhythmicity, to the extent that the mere presence of periodicity in an observed phenomenon is oftentimes sufficient for it to be labelled "rhythmic". But the rather ill-defined concept of "rhythm" is called upon to do duty in many contexts, from the mysterious oscillations found deep within brains to the aesthetic gyrations of pairs of dancers. From what we have seen here (and elsewhere), it is apparent that the term rhythm picks out quite different things in different domains. The rhythmicity of a nursery rhyme is not the same as that of a rock song, or a samba song, which are, again, mutually distinct. The coordinative relations observed in the temporal signatures of these diverse behaviours must be understood with respect to the skills shared by performers, the conventions of specific genres, and perhaps also the nature of the bond between practitioners.

7. References

[1] D. Deutsch, T. Henthorn, and R. Lapidis, "Illusory transformation from speech to song," *The Journal of the Acoustical Society of America*, vol. 129, p. 2245, 2011.

[2] J. S. Garofolo, *TIMIT: Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, 1993.

[3] R. A. Cole, M. Noel, and V. Noel, "The CSLU speaker recognition corpus." in *Proc. ICSLP*, vol. 98, 1998, pp. 3167–3170.

[4] F. Cummins, "Rhythm as entrainment: The case of synchronous speech," *Journal of Phonetics*, vol. 37(1), pp. 16–28, 2009.

[5] A. Cutler and J. Mehler, "The periodicity bias," *Journal of Phonetics*, vol. 21, no. 1/2, pp. 103–8, 1993.

[6] F. Cummins, "Practice and performance in speech produced synchronously," *Journal of Phonetics*, vol. 31, no. 2, pp. 139–148, 2003.

[7] R. H. Stetson, *Motor Phonetics*, 2nd ed. Amsterdam: North-Holland, 1951.

[8] R. Port and T. van Gelder, Eds., *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: Bradford Books/MIT Press, 1995.

[9] F. Cummins, "Joint speech: The missing link between speech and music?" *Percepta—Revista de Cognição Musical*, 2013, in press.

[10] B. Raposo De Medeiros, "Descrição comparativa de aspectos fonetico-acusticos selecionados da fala e do canto em portugues brasileiro," Ph.D. dissertation, University of Campinas, Campinas, Brazil, 2002.

[11] J. Sundberg *et al.*, *The science of the singing voice*. DeKalb, Il.: Northern Illinois University Press, 1987.

[12] J. Ross and I. Lehiste, *The temporal structure of Estonian runic songs*. Walter de Gruyter, 2001, vol. 1.

[13] R. Kolinsky, P. Lidji, I. Peretz, M. Besson, and J. Morais, "Processing interactions between phonology and melody: Vowels sing but consonants speak," *Cognition*, vol. 112, no. 1, pp. 1–20, 2009.

[14] C. Sandroni, *Feitiço decente: Transformações do samba no Rio de Janeiro, 1917-1933*. Jorge Zahar Editor, 2001.

[15] D. Temperley, "Syncopation in rock: a perceptual perspective," *Popular Music*, vol. 18, no. 01, pp. 19–40, 1999.

Accentual phrases in Slovak and Hungarian

Katalin Mády[1], *Uwe D. Reichel*[2], *and Štefan Beňuš*[3,4]

[1]Institute for Linguistics, Hungarian Academy of Sciences, Budapest
[2]Institute for Phonetics and Speech Processing, University of Munich
[3]Constantine the Philosopher University, Nitra
[4] Institute of Informatics, Slovak Academy of Sciences, Bratislava
mady@nytud.hu, reichelu@phonetik.uni-muenchen.de, sbenus@ukf.sk

Abstract

Languages with primarily delimitative function of word stress commonly make use of accentual phrases (APs) in their intonational phonology (e.g. Tamil or French). Slovak and Hungarian are genetically unrelated but geographically close languages with word-initial lexical stress. In this paper we compared the stylised f0 of single accent groups (AGs) with the f0 level pattern of the entire intonational phrase (IP) to test if AGs are relevant for the intonational phonology of Slovak and Hungarian. Steep f0 slopes with a recurring pattern (rising or falling) and large deviations from IP level patterns were interpreted as evidence for the autonomy of the AG in the given language. The results suggest that in Hungarian, accent groups form an independent unit with a falling pitch contour. Such evidence was not found for Slovak.

Index Terms: prosodic phrasing, accentual phrase, intonation modelling, Hungarian, Slovak.

1. Introduction

Most prosodic models [1, 2, 3] assume that the highest prosodic unit is the intonational phrase (IP) that is marked by final and potentially initial boundary tones (or edge tones), phrase-final lengthening and an optional pause following the phrase. Other models subordinate the IP to a higher-level unit, such as the utterance, e.g. [4]. The smallest prosodic phrase is the prosodic word (PW). Two other units, the intermediate phrase (ip) and the accentual phrase (AP) are located between the intonational phrase and the prosodic word [5]. Whereas IPs and PWs are present in all languages, ips and APs are optional. (See [6, p. 444] for an overview of the presence or absence of certain prosodic units for a set of languages.)

Both Hungarian and Slovak have lexical stress on the left-most syllable of a prosodic word (apart from Eastern Slovak dialects with penultimate stress). In Hungarian, lexical stress is fixed to the first syllable, i.e. *INtonáció* 'intonation' (the "accent" sign marks long vowels and not stress). In Slovak, stress can be optionally moved from the word-initial syllable if the word is preceded by a preposition, e.g. *HOry* 'hills', *DO hory* 'to the woods'. Hence, Slovak lexical words may combine with preceding monosyllabic prepositions and possible clitics following this word into a prosodic word with left-most stress.

In Hungarian, pitch accents are typically realised within the initial syllable of the accented word. In Slovak, pitch targets might be aligned within or after the accented syllable, but it is not clear if this is due to phonological differences or phonetic implementation.

In languages with fixed stress towards the left or right edge of the word, stress is often used as the onset or offset of a prosodic phrase [5], while the other edge of the phrase is marked by a boundary tone. In languages in which sequences between two accents form an accentual phrase (AP), the edge tones often show a regular pattern. Tamil for example has phrase-initial L* tones and AP-final H tones (Ha).

[5] describe accentual phrases in the following way: (1) an AP can contain only one pitch accent, (2) there is a boundary tone at the other edge of the prosodic unit, (3) the edge tone is insensitive of stress, i.e. its realisation is independent of stress location, (4) APs are characterised by a recurrent rising or falling pattern depending on the language.

Given the observed correlation between fixed lexical stress position and relevance of AGs for the intonational phonology of a particular language, we investigate if left-headed prosodic words in Slovak and Hungarian also constitute a prosodic unit that is marked by tonal (f0) means. While initial efforts in building a ToBI system for Slovak prosody do not propose units below intermediate phrases [7], previous work on Hungarian suggests that pitch accents indeed initiate APs within intonational phrases [8, 9].

In left-headed accentual phrases, pitch accents are always preceded by a prosodic boundary and thus automatically function as an edge marker. This would explain the relatively frequent occurrence of pauses before accented words in unexpected positions in Hungarian, e.g. between a definite article and an emphasised noun [10]. In this case, the pause can be interpreted as an additional cue marking the phrase boundary and thus strengthening the emphasis on the pitch-accented word. Apart from our theoretical interest in the phrasal organisation of Slovak and Hungarian prosody, the function of the phrase types can help to gain a better understanding of underlying communicative intentions and also to enhance the effectiveness of speech recognition and the naturalness of speech synthesis.

The analysis presented here is based on the assumption that an accentual phrase has its own f0 slope in form of a rising or falling pitch contour, and that this pattern is independent of the overall pitch contour of the intonational phrase. Second, since APs are characterised by the frequent occurrence of the same pattern (e.g. rising pattern for French and Tamil), the frequent occurrence of a given pattern is interpreted as a further evidence for the presence of APs in the language.

2. Material and methods

50 Slovak and Hungarian spontaneous utterances forming a single IP were selected from collaborative dialogues (5 utterances

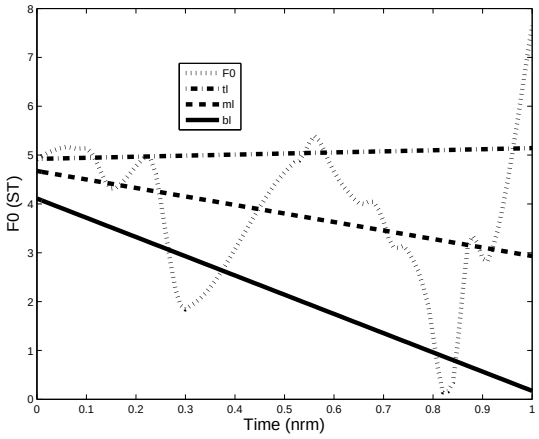


Figure 1: *Stylised top-, mid- and baseline.*

of 10 Slovak and 10 Hungarian speakers, respectively). All IPs had a low phrase-final boundary tone. Only IPs with at least two pitch accents were used for the analysis. Both IPs and pitch accents were identified manually by a phonetically trained native speaker (1st and 3rd author of this paper). Analysis was based on the f0 level and the range pattern throughout the IP and the accent group (AG), ranging from one accented syllable till the last unaccented syllable before the next pitch accent or the IP boundary. The Slovak material consisted of 157 accent groups and the Hungarian samples of 130 accent groups.

2.1. F0 extraction and preprocessing

F0 was extracted by autocorrelation (Praat 5.3, sample rate 100 Hz). Data were further processed in Matlab. Voiceless utterance parts and f0 outliers were interpolated by piecewise cubic splines [11]. The contour was then smoothed by Savitzky-Golay filtering using third order polynomials in 5 sample windows and transformed to semitones relative to a base value [12, 13]. This base value was set to the f0 median below the 5th percentile of an utterance and served to normalise f0 with respect to its overall level.

2.2. Stylisation

To capture the f0 register in terms of its level and range [14] we fitted a base-, a mid-, and a topline separately for the IP and all AGs within this IP (Fig. 1). The midline represents the f0 level, whereas the base- and topline provide the f0 range information. The robust fitting procedure that is motivated and explained in more detail in [15] consists of the following steps:

- A window of 200 ms length is shifted along the f0 vector with the step size of 10 ms.
- Within each window the f0 median is calculated
  - of the values below the 10th percentile for the baseline,
  - of the values above the 90th percentile for the topline, and
  - of all values for the midline.

  This gives 3 sequences of medians, one for the base-, the mid-, and the topline, respectively.
- Within each median sequence outliers are replaced by linear interpolation.
- Finally, for all three median sequences linear polynomials are fitted.

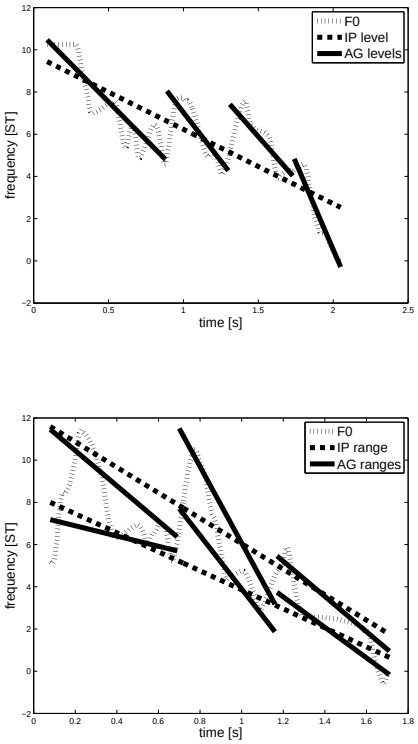


Figure 2: *IP and AG level (top) and range (bottom) patterns for Hungarian.*

2.3. Features

If accentual phrases play a role in the intonational phonology of a language, they should show a significant difference from the corresponding part of the intonational phrase in terms of local f0 level deviations and f0 ranges. In other words, the slope of the f0 midline of a given AP is supposed to be steeper than that of the IP, and the range larger than can be expected on the basis of the overall top- and baseline differences in the IP.

We extracted features to capture first the general AG register and second its deviation from the IP.

Level

For the f0 level, the AG pattern is represented by the slope of the fitted midline (*mlslope*), as shown in the left graph in Figs. 2 and 3. AG deviation was measured (1) with respect to shape in terms of the absolute slope difference of the AG and the IP midlines (*mlSlopeDiff*), (2) with respect to overall distance given by the root mean squared deviation of the AG line from the corresponding section of the IP line (*mlRms*, and (3) with respect to local differences by subtracting the corresponding values of the

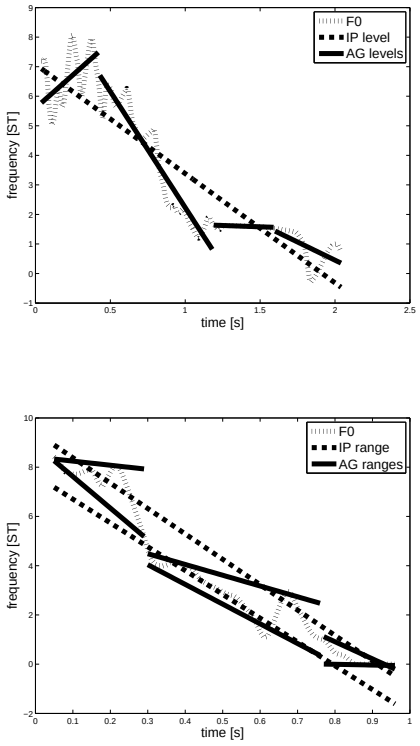


Figure 3: *IP and AG level (top) and range (bottom) patterns for Slovak.*

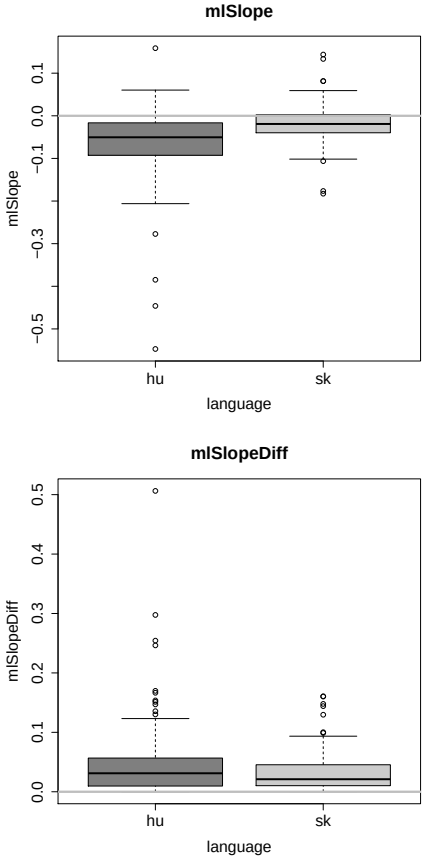


Figure 4: *Slope (mlSlope) and absolute slope difference (mlSlopeDiff) for Hungarian and Slovak. The grey line indicates 0.*

IP line from the initial and the final f0 values of the AG line (*mlInitYDiff, mlFinYDiff*).

Range

Next to level deviations, AGs can be set apart from their underlying IPs by the AG range, see the bottom graphs in Figs. 2 and 3. We thus measured the range of each AG in terms of the root mean squared deviation between its base- and topline (*rangeRms*).

The expected acoustic correlates for the presence of accentual phrases are prominent f0 movements reflected in high AG range values (*rangeRms*) as well as considerable local level deviations between the AG and the IP expressed in high values for the features *mlSlopeDiff, mlRms, mlInitYDiff*, and *mlFinYDiff*. Results were compared by means of t-tests. First, a one-sample t-test was carried out for the features for both languages, i.e. results were compared to a sample with $mean = 0$. Subsequently, Hungarian and Slovak samples were compared to each other. If the data were not normally distributed, the Mann-Whitney test was carried out instead. The equality of variances was tested by the Levene test that is also applicable to non-normally distributed data. Significance level was set to $p = 0.05$.

Languages with falling or rising APs are expected to have mostly negative or mostly positive slopes that differ considerably from the IP slope. At the same time, the onset and/or the offset of the AG slope is supposed to differ from the corresponding section of the IP. Additionally, larger overall distances (root mean square distances) and larger ranges demonstrate a greater autonomy of the AG as a prosodic phrase.

3. Results

3.1. Slope

The majority of slopes (*mlSlope*) was negative in both languages. Absolute slope differences (*mlSlopeDiff*) were larger for Hungarian (Fig. 4). Mean slopes and mean slope differences differed significantly from 0 in both languages, and also between the two languages. All differences were highly significant ($p < 0.0001$).

The overall distance (*mlRms*) between the AG midline and the corresponding IP section (Fig. 5) differed significantly from 0, but not between the two languages ($p = 0.24$).

3.2. Local initial and final AG–IP difference

The distance between the AG and the IP midlines was calculated both for the onset (*mlInitYDiff*) and the offset (*mlFinYDiff*) of the AG. Larger distances refer to a larger deviation and thus a larger independence of the AG midline. The distance was expressed in absolute values.

Both AG-initial and AG-final distances were significantly larger than 0 in both languages ($p < 0.01$) (Fig. 6). Onset

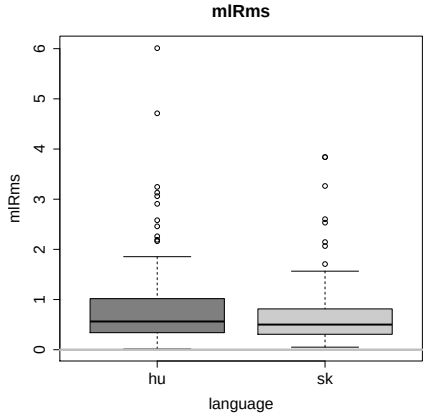


Figure 5: *Root mean square between the AG and IP midline (mlRms for Hungarian and Slovak. The grey line indicates 0.*

distances were significantly larger in Hungarian than in Slovak, whereas offset distances did not differ between the two languages.

3.3. Range

Finally, the range between the base- and the topline was compared for the AG. It was assumed that a larger range within the AG shows the autonomy of the AG as a separate prosodic unit.

Ranges differed significantly from 0, but not between languages ($p = 0.67$).

4. Conclusions

AGs show a falling pattern in both languages, i.e. their f0 slope is steeper than the IP slope. This tendency is more clear-cut in Hungarian than in Slovak. The onset and the offset of the AG slope deviate from the IP slope for both languages, while the distance of the onset is greater in Hungarian than in Slovak. F0 slopes in Slovak are rather flat, and the tendency for uniform patterns is not present. In other words, the findings do support the hypothesis that Hungarian prosody involves accentual phrases, but the evidence is less clear for Slovak.

One possible reason for the flatness of Slovak f0 slopes is that the AGs might have a non-linear pattern. As was said before, the f0 maxima of pitch accents are often delayed in Slovak which might cause a rising-falling pattern or be triggered by it. This option needs further testing by non-linear stylisation.

A future large-scale study will involve more languages without fixed lexical stress. This would allow to specify the interrelation of f0 slopes and prosodic phrases in a more detailed way.

5. Acknowledgements

This work was supported by the Hungarian Scientific Research Fund (PD 101050) and the Momentum program of the Hungarian Academy of Sciences (first author) and by the ERDF Research & Development Operational Programme *Research and development of new information technologies for forecasting and mitigation of crisis situations and safety*, ITMS 26240220060 (third author).

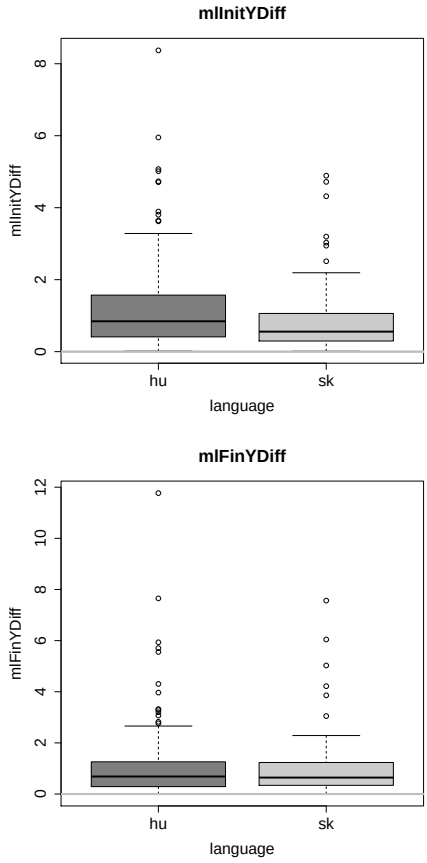


Figure 6: *Local initial (mlInitYDiff) and final (mlFinYDiff) AG–IP differences for Hungarian and Slovak (absolute values). The grey line indicates 0.*

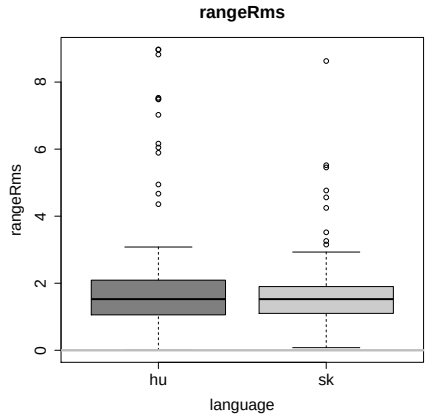


Figure 7: *Range between base- and topline (rangeRms) for Hungarian and Slovak. The grey line indicates 0.*

6. References

[1] A. Cruttenden, *Intonation, 2nd ed.* Cambridge: University Press, 1997.

[2] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology*, pp. 255–309, 1986.

[3] M. Vogel and I. Nespor, *Prosodic phonology*. Dordrecht: Foris Publications, 1986.

[4] C. Gussenhoven, "Transcription of Dutch intonation," in *Prosodic typology*. Oxford: University Press, 2005, pp. 118–145.

[5] S.-A. Jun and J. Fletcher, "Methodology of studying intonation: From data collection to data analysis," in *Prosodic Typology II: the new development in the phonology of intonation and phrasing*. Oxford: University Press, 2014.

[6] S.-A. Jun, Ed., *Prosodic typology*. Oxford: Oxford University Press, 2005.

[7] M. Rusko, R. Sabo, and M. Dzúr, "Sk-tobi scheme for phonological prosody annotation in Slovak," in *Text, speech and dialogue*, ser. Lecture notes in computer science, V. Matoušek and P. Mautner, Eds. Berlin & Heidelberg: Springer, 2007, vol. 4629, pp. 334–341.

[8] L. Hunyadi, *Hungarian sentence prosody and universal grammar: on the phonology–syntax interface*. Frankfurt/Main: Lang, 2002.

[9] K. Mády, A. Szalontai, A. Deme, and B. Surányi, "On the interdepencende of prosodic phrasing and prosodic prominence in Hungarian," in *Proc. 11th International Conference on the Structure of Hungarian*, Piliscsaba, Hungary, 2013.

[10] K. Mády and F. Kleber, "Variation of pitch accent patterns in Hungarian," in *Proc. 5th Speech Prosody Conference, Chicago*, 2010, pp. 100 924:1–4.

[11] C. de Boor, *A Practical Guide to Splines*, ser. Applied Mathematical Sciences. Springer, 1978, no. 27.

[12] A. Savitzky and M. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[13] P.-O. Persson and G. Strang, "Smoothing by Savitzky-Golay and Legendre Filters," in *Mathematical systems theory in biology, communications, computation, and finance*, D. Gilliam, Ed. Springer, 2003, pp. 301–315.

[14] T. Rietveld and P. Vermillion, "Cues for Perceived Pitch Register," *Phonetica*, vol. 60, pp. 261–272, 2003.

[15] U. D. Reichel and K. Mády, "Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech," in *Elektronische Sprachsignalverarbeitung*, ser. Studientexte zur Sprachkommunikation, P. Wagner, Ed., vol. 65. Bielefeld: TUDpress, 2013, pp. 223–230.

Final devoicing of /l/ in Reykjavík Icelandic

Nicole Dehé

Fachbereich Sprachwissenschaft, Universität Konstanz, Germany

nicole.dehe@uni-konstanz.de

Abstract

Icelandic has a phonological process which devoices sonorants after voiced segments in domain-final position, but to date the category of the relevant domain and potential further factors affecting it have not been identified. The present paper studies final devoicing of /l/, by which /l/ is realized as the voiceless lateral fricative [ɬ] in domain-final position. It reports on the results of an experimental reading study designed to test the exact environments of this process and the implications for a prosodic hierarchy for Icelandic. The results suggest that devoicing of /l/ is bound by the prosodic utterance. All instances of /l/ were devoiced in utterance final position. Within the utterance, final devoicing is optional, but the frequency of its application reflects the syntactic and prosodic hierarchy such that it is most frequent at a clause/an IP-boundary, significantly less frequent at a syntactic XP-edge and it almost never occurs within a syntactic XP.

Index Terms: final devoicing, sonorant devoicing, Icelandic, prosodic hierarchy, /l/

1. Introduction

Icelandic has a phonological process which devoices sonorants "after voiced segments in phrase-final position" [1]. According to [2], this process is very common in modern Icelandic speech, indicates a break in the utterance and the end of a phonological phrase or utterance, and is likely to co-occur with a boundary tone (T%). According to [3], "phrase final devoicing in consonants is […] practically obligatory". The devoiced /l/ is phonetically realized as voiceless lateral fricative (rather than, for example, devoiced lateral approximant, e.g. [3]). Examples are given in (1) (from [4]).

(1) a. Jón er á bíl [ˈpiːl] í dag.
Jón is on car today (' John is driving today.')

b. Jón er á bíl. [ˈpiːɬ]

While this process has been described, for example, in the works mentioned above, its distribution has never been studied systematically and formulations such as "the end of (some sort of) a phonological phrase or utterance" [2] are very vague. At the same time, a prosodic hierarchy has not yet been established for Icelandic. For the level of the Intonational Phrase (IP), it has been observed that the tonal inventory of Icelandic has two boundary tones terminating the IP (L% and H%) and that the IP is the domain for declination ([5], [6]). As for a level between the prosodic word (PWd) and the IP (e.g. phonological phrase, intermediate phrase; see [7] for an overview of hierarchies suggested in the literature), conclusive evidence has not yet been provided. To date there is only preliminary evidence from the environment of another phonological process (word-final vowel deletion, see [8]) as well as preliminary evidence for phrase accents L- and T-, i.e. the edge tone of the intermediate phrase (see [6]), but more systematic research is necessary, especially regarding tonal events. The blocking and/or application of phonological processes have long been taken as evidence for the existence of categories in the prosodic hierarchy; e.g. Visarga in Sanskrit for Utterance level (see [9]), Italian Gorgia Toscana for IP level and Raddoppiamento Sintattico for ip level (see [10] for the latter two) to name but a few. The present paper reports on an experiment designed to test the environment of final devoicing of /l/ in Icelandic and the implications for a prosodic hierarchy in Icelandic.

2. The experiment

A reading task was designed to produce data on final /l/ in different positions. Four contexts were considered: utterance-final, clause-final, XP-final and XP-internal. The clause-final context was chosen because according to current prosodic theory, a clause in the syntactic structure is predicted to coincide with an IP in the prosodic structure (see [11], [12]). Similarly, the XP-final context was chosen because an XP in the syntactic structure is predicted to coincide with an ip in the prosodic structure ([11], [12]). Neither IP- nor ip-boundary are predicted to occur XP-internally. Since a prosodic hierarchy has not yet been established for Icelandic on the basis of tonal and/or other kinds of phonological evidence, the target contexts were chosen according to these predictions.

All target words ended in /l/ after a vowel. As a rule, Icelandic word stress falls on the initial syllable. To control for the potential effects of lexical stress, target words were monosyllabic (stress condition) and disyllabic (non-stress condition). To control for the potential effect of the following segment and voicing assimilation, the target words were followed by vowels (voiced condition) or by voiceless fricatives (/s, f, θ/) and in two cases voiceless plosives (/p, t/) (voiceless condition). Example target words and some following units are given in Tables 1 and 2.

	Example target words
mono-syllabic	kál (/ˈkʰauːl/ 'cabbage', ACC-Sg)
	stól (/ˈstouːl/ 'chair', ACC-Sg)
di-syllabic	blómkál (/ˈplouːm. kʰauːl/ 'cauliflower', ACC-Sg)
	viðtal (/ˈvɪð.tʰal/ 'interview', ACC-Sg)

Table 1. *Example target words.*

	Example words following target
voiced	en ('but'), í ('in')
	ömmu ('grandmother', GEN-Sg)
voiceless	sem (Rel-Prn), frá ('from')
	systur ('sister', GEN-Sg)

Table 2. *Examples of words following target words.*

Given that final devoicing has been described as indicating a clear break in the utterance and that a boundary tone has been considered likely to co-occur with final devoicing (see [2]), the predictions with respect to position were as follows:

- Final devoicing occurs, and is likely to be obligatory, at the end of an utterance.
- Final devoicing occurs, and is likely to be obligatory, at the end of an IP, i.e. at the end of a clause in the syntax.
- Final devoicing may occur, if bounded by a level between PWd and IP, at the end of a syntactic XP, specifically between object and adjunct, a position likely to coincide with a phonological phrase boundary in prosodic structure.
- Final devoicing does not occur XP-internally, a position likely to correspond to a PWd boundary in the prosodic structure and not a position where a break in the utterance would be expected.

2.1. Materials

The experiment considered three factors: position (4 levels: utterance-final, clause-final, XP-final, XP-internal), stress on the target syllable (2 levels: stressed, unstressed) and voicing of the following sound (2 levels: voiced, voiceless). Overall, 56 sentences were created, fourteen of which are given in (2) through (5) as examples. Eight sentences had the target word in utterance-final position; of these, four target words were mono-syllabic (see (2)a), and four disyllabic (see (2)b). Sixteen sentences had the target words in clause-final position; eight (four monosyllabic, four disyllabic) were followed by voiced segments, eight (four monosyllabic, four disyllabic) by voiceless segments; see (3) for examples. Sixteen sentences had the target words in XP-final position, with the same distribution regarding stress and following segments as in the clause-final condition (see (4)), and sixteen sentences had the target word in XP-internal position, specifically in NP-internal position (see (5)). The sentences were pseudorandomised according to the usual restrictions. Note that the voicing condition only applies to three positions; in utterance-final position, no segment follows the target word ending in /l/.

(2) Target word utterance(U)-final

a. Í gær borðuðum við svínakjöt, kartöflur og kál.
Yesterday we at pork, potatoes and cabbage.
b. Einu sinni eldaði mamma mín oft blómkál.
In the past my mum cooked cauliflower often.

(3) Target word clause(Cl)-final

a. [Í gær borðuðum við svínakjöt, kartöflur og kál,]$_{clause}$
[en í dag borðum við fiskisúpu]$_{clause}$
Yesterday we ate pork, potatoes and cabbage,
and today we ate fish soup.
b. [Í gær borðuðum við svínakjöt, kartöflur og kál,]$_{clause}$
[sem bróðir minn og kærasta hans elduðu fyrir okkur]$_{clause}$
Yesterday we ate pork, potatoes and cabbage,
which my brother and his fiancé cooked for us.
c. [Einu sinni eldaði mamma mín oft blómkál,]$_{clause}$
[en nú á dögum eldar hún ekki neitt]$_{clause}$
In the past my mum often cooked cauliflower,
and today she does not cook anything.
d. [Einu sinni eldaði mamma mín oft blómkál,]$_{clause}$
[sem hún keypti á markaðnum í hverri viku]$_{clause}$
In the past my mum often cooked cauliflower,
which she bought in the market every week.

(4) Target word XP-final

a. Einu sinni eldaðum við [kartöflur og kál]$_{NP}$ [á hverjum degi]$_{PP}$
In the past we ate pork, potatoes and cabbage every day.
b. Á morgun þarf ég að kaupa [kál]$_{NP}$ [fyrir kjötsúpuna]$_{PP}$.
Tomorrow I have to buy cabbage for the meat soup.
c. Í morgun keypti sambýlismaður minn [blómkál]$_{NP}$ [í Bónus]$_{PP}$
This morning my flat mate bought cauliflower in Bónus.
d. Í Bónus keypti sambýliskona mín [blómkál]$_{NP}$ [frá Spáni]$_{PP}$
In Bónus my flat mate bought cauliflower from Spain.

(5) Target word XP-internal

a. Mér líkar [kál ömmu minnar]$_{NP}$ sem hún eldar.
I like my grandmother's cabbage which she cooks.
b. Mér líkar [kál frænku minnar]$_{NP}$ sem hún ræktar.
I like my cousin's cabbage which she grows.
c. Okkur þykir [súrkál ömmu okkar]$_{NP}$ frábært.
We think that our grandmother's sauerkraut is great.
d. Okkur þykir [súrkál systur okkar]$_{NP}$ rosalega gott.
We think that our sister's sauerkraut is very good.

2.2. Participants, apparatus and procedure

The recordings took place in November and December 2013 in a quiet closed room at the University of Iceland in Reykjavík. The results reported here are based on the recordings of twelve female native speakers of Icelandic. All speakers were from Reykjavík or the greater capital area or had lived there most of their lives. They were between 19 and 35 years of age and volunteered their participation. The participants were seated in front of a laptop computer. The sentences were presented one at a time on the computer screen using Microsoft PowerPoint. Sentences were presented on one line, except for the clause-final condition, where a comma was placed after the target word and a new line was started after the comma to help elicit an IP-boundary. The participants read the sentences at a normal speech rate. All utterances were recorded at a sampling rate of 44100 Hz using a Microtrack II (M-Audio) recorder and Rode NT-5 condenser microphone. The recordings were then edited into individual sound files.

2.3. Data treatment

Overall, the 12 participants produced 672 target sentences (96 utterance-final, 192 clause-final, 192 XP-final, 192 XP-internal). All targets were annotated in Praat ([13]) based on careful inspection of waveform, spectrogram and F0 contour and on perception. In the U-final, XP-final and XP-internal conditions, all utterances were annotated on at least a segmental and a text tier; in the clause-final condition all utterances were additionally annotated on a tone tier to identify boundary tones (T%) at IP-edges in target position (e.g. Figure 3). Boundary tones (L% or H%) occurred in target position in all 192 target sentences produced in the clause-final condition.

As for /l/, three realizations occurred in the data: [l], [ɬ], and [l ɬ], the latter being infrequent (N=14), and mostly produced at a clause boundary (N=11). Based on native speaker perception and articulatory differences between [ɬ] on its own and [ɬ] in an [l ɬ] sequence, [l ɬ] was counted as [l]. The realizations of /l/ as [l] and [ɬ] are illustrated in Figures 1 and 2, respectively, zooming in to the target area.

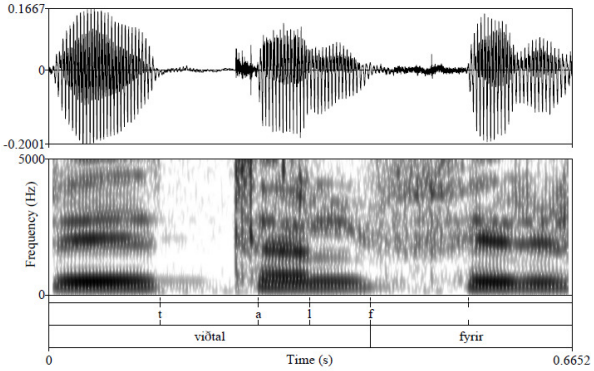


Figure 1: *No devoicing: /l/; Sentence: Blaðamaðurinn tók áhugavert viðtal fyrir Fréttablaðið ('The journalist took an interesting interview for Fréttablaðið.').*

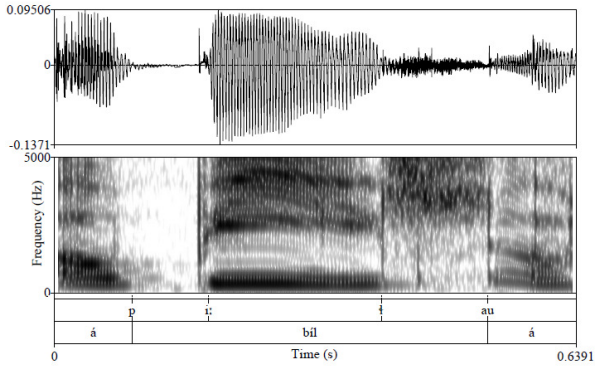


Figure 2: *Devoicing applied: /l̥/; Sentence: Venjulega eru Jón og Brynja á bíl á sunnudögum. ('Normally Jón and Brynja take the car on Sundays.')*

2.4. Results

The results by position are as follows. First, all 96 final /l/ in utterance-final position (see (2)) were devoiced, i.e. realized as [l̥]. Second, 126 (66%) final /l/ in clause-final position (see (3)) were devoiced, i.e. realized as [l̥], 62 (32%) were voiced [l], and four (2%) were unclear. Third, 164 (85%) final /l/ in XP-final position were realized as voiced [l], 22 (12%) were realized as [l̥], and 6 (3%) were unclear. Finally, XP-internally, 182 (95%) final /l/ were realized as [l], only 6 (3%) were realized as [l̥] and 4 (2%) were unclear or not realized at all. These results are summarized in Figure 3 and Table 3 below, excluding the unclear cases.

In order to test whether these differences between positions were significant and whether stress on the syllable ending in /l/ and voicing of the following segment also affected devoicing of /l/, the data were analysed statistically. They were coded by position (as in (2)-(5)), stress on the target syllable (stressed, unstressed) and voicing of the following segment (voiced, voiceless). All unclear cases (N = 14) were discarded from the analysis. The data were aggregated by participants and analyzed using a binomial logistic regression model with devoicing ([l] vs. [l̥]) as the dependent variable and the above-mentioned factors as fixed factors. The analysis showed no effects of stress ($p > 0.2$, see Table 4) and no effect of voicing of the following segment ($p > 0.12$, see Table 5), but a main effect for position. Specifically, there were significantly more [l̥]-realizations at the end of the utterance compared to the end of the clause ($ß = 0.35$, $SE = 0.15$, $p < 0.05$), which again had significantly more [l̥]-realizations than target words produced at the end of XPs ($ß = -1.90$, $SE = 0.27$, $p < 0.0001$), which again had significantly more [l̥]-realizations than target words that occurred within an XP ($ß = -0.94$, $SE = 0.46$, $p < 0.05$). There were no significant interactions between factors.

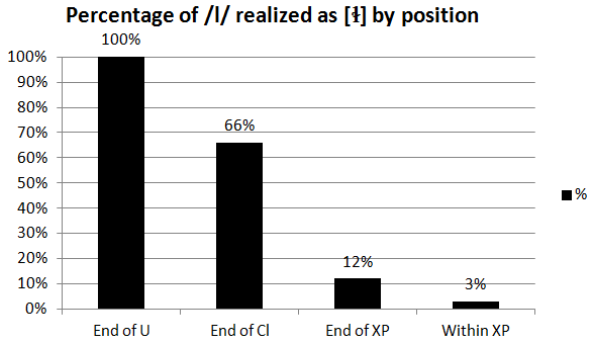


Figure 3: *Results by position*

position	[l]	[l̥]
1 (U-final)	0	96
2 (Clause-final)	62	126
3 (XP-final)	164	22
4 (XP-internal)	182	6

Table 3. *Cross-tabulation of /l/ according to position and (de)voicing*

	stressed (mono-syllabic)		unstressed (disyllabic)	
position	[l]	[l̥]	[l]	[l̥]
1 (U-final)	0	48	0	48
2 (Clause-final)	32	62	30	64
3 (XP-final)	77	17	87	5
4 (XP-internal)	88	6	94	0

Table 4. *Cross-tabulation of /l/ according to position, stress and (de)voicing*

	voiced following segment		voiceless following segment	
position	[l]	[l̥]	[l]	[l̥]
1 (U-final)	0	0	0	0
2 (Clause-final)	19	77	43	49
3 (XP-final)	84	10	80	12
4 (XP-internal)	89	6	92	1

Table 5. *Cross-tabulation of /l/ according to position, voicing of the following segment and (de)voicing*

3. Discussion

Final devoicing of /l/ in Reykjavík Icelandic obligatorily marks the end of the utterance (U). It is thus a U limit rule in Nespor & Vogel's [10] sense and seems to resemble in its

distribution other phonological rules which mark the end of a phonological utterance, such as final devoicing in Spanish (see [10]) and in the Chadic language Angas (see [14]), as well as Visarga in Sanskrit (see [9]), although not all of these have been experimentally studied.

The process is frequent but not obligatory at the end of the IP, thus it does not necessarily co-occur with other IP-boundary markers such as boundary tones. To illustrate this, Figure 4 shows an example of [l] produced at the end of IP, clearly co-occurring with H%, a short break of roughly 50 ms, and followed by pitch reset in the following IP, serving as additional evidence for a new IP, given that the IP has been identified as the domain for declination in Icelandic ([5], [6]).

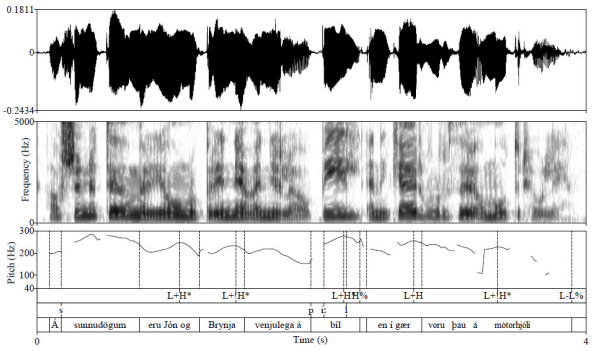


Figure 4: *[l] at IP-boundary; Sentence: Á sunnudögum eru Jón og Brynja venjulega á bíl, en í gær voru þau á mótorhjóli. ('On Sundays Jón and Brynja normally take the car, but today they were by motorbike.')*

Given the statistically significant difference between end of utterance (obligatory application of devoicing) and end of IP (optional application of devoicing, no obligatory co-occurrence with other IP-boundary markers), final devoicing in Icelandic may be taken as evidence for the level of the prosodic utterance in the prosodic hierarchy. Without this level, the difference between the two positions found in the present study cannot be accounted for.

Devoicing of /l/ is clearly disfavored, i.e. does not normally apply, within IP. If there is a level between PWd and IP in the prosodic hierarchy of Icelandic, it does not seem to affect devoicing of /l/. Note that of the 22 /l/ realized as [ɬ] at an XP-boundary, 10 are in the sequence <*bíl til*> ('car to') in the sentence <*Á síðasta ári förum við í bíl til Akureyrar*> ('Last year we drove by car to Akureyri'). Given the sequence /l t/, it is possible that even across a syntactic XP-boundary, it behaves in the same way as the same sequence of sounds word-internally, e.g. in adjectives marked for neuter on a stem ending in –l, cf. *kalt* ('cold', neuter of *kaldur*, stem *kal-*) and *gult* ('yellow', neuter of *gulur*, stem *gul-*), pronounced with devoiced /l/ before /t/ (e.g. [15]). Within XP, devoicing applies even less frequently. However, if the item involving the sequence [*bíl*]$_{NP}$ [*til Akureyrar*]$_{PP}$ is removed from the statistical analysis, the difference between position 3 (XP-final) and position 4 (XP-internal) is not significant anymore ($p > 0.9$). The few remaining instances of [ɬ] at an XP-boundary may be put down to careful articulation or promotion of the prosodic boundary at an XP-edge to IP-level.

The most interesting result is clearly the significant difference between positions 1 (U-final) and 2 (clause/IP-final), given that within IP devoicing is clearly disfavored, and given that the predictions were identical for positions 1 and 2. A factor not included in the experimental design is information structure. All utterances were produced with wide focus. Icelandic has right-most prominence at IP-level (e.g. [5], [16]), thus target words bear nuclear prominence in positions 1 and 2, and prenuclear prominence in positions 3 (XP-final) and 4 (XP-internal), but are never located in post-nuclear position. Given that stress on the target syllable turned out not to be significant, I would not expect the difference between nuclear and prenuclear prominence to be responsible for the clear position effect, and at any rate, it would not explain the significant difference between positions 1 and 2, which are both nuclear. Perhaps the size of the prosodic constituent including the target word may be another factor. However, again, this would not explain the difference between positions 1 and 2, because the size of the target constituents (U, IP) was identical; compare (2) and (3). For the time being, I thus conclude that it is the category of the prosodic constituent (U vs. IP) which is responsible for the obtained effect.

4. Concluding remarks

This paper has shown that final devoicing of /l/ (i.e. realization of final /l/ as [ɬ]) obligatorily marks the end of an utterance in Icelandic and provides evidence for utterance level in the prosodic hierarchy. It is still frequent but not obligatory at the end of an IP. If the prosodic hierarchy of Icelandic has a level between IP and the prosodic word, final devoicing of /l/ is not the process to establish this level, because its application is very infrequent within IP at potential lower level boundaries; the difference between XP-final /l/ and XP-internal /l/ is only marginally significant and disappears when one particular item is removed which may lead to devoicing of /l/ for independent reasons. Future research will have to show whether other phonological processes may help to establish more levels in a prosodic hierarchy for Icelandic. Moreover, it will show whether the results reported here extend to other sonorants which undergo devoicing in word-final position in Icelandic, e.g. /ð, r, ɣ/, and whether utterances in context (e.g. in turn-final position) will behave in the same way as utterances in isolation. Furthermore, only the results for Reykjavík Icelandic have been reported on here, thus future results may reveal regional variation in final sonorant devoicing. Given that Icelandic has phonological dialects (see [17]) and that the voiced vs. voiceless pronunciation of word-internal sonorants (e.g., /l, m, n/ before stops /p, t, k/) is one dialectal feature, it is at least conceivable that dialectal variation extends to final devoicing, too.

5. Acknowledgements

I am grateful to Ari Páll Kristinsson, Aðalsteinn Hákonarson, Haukur Þorgeirsson, Kristján Árnason and Élisabeth Delais-Roussarie for valuable comments and discussion, to Nanna Kristjánsdóttir and Þorbjörg Þorvaldsdóttir for help with the experimental data, and to Bettina Braun for discussion and help with the statistical analysis. This piece of research was supported by a Snorri Sturluson Fellowship from the Árni Magnússon Institute for Icelandic Studies, University of Iceland, to the author.

6. References

[1] Thráinsson, H., "Icelandic", in E. König and J. van der Auwera [Eds], The Germanic Languages, 142-189, Routledge, 1994.

[2] Árnason, K., "Phonological domains in Modern Icelandic", in J. Grijzenhout and B. Kabak [Eds], Phonological Domains: Universals and Deviations, 283-313, Mouton de Gruyter, 1999.

[3] Helgason, P., On Coarticulation and Connected Speech Processes in Icelandic. Málvísindastofnun Háskóla Íslands, 1993.

[4] Árnason, K., The Phonology of Icelandic and Faroese. Oxford University Press, 2011.

[5] Árnason, K., "Toward an analysis of Icelandic intonation", in S. Werner [Ed], Nordic Prosody. Proceedings of the VIIth Conference, Joensuu 1996, 49-62, Peter Lang, 1998.

[6] Dehé, N., "An intonational grammar for Icelandic", Nordic Journal of Linguistics 32(1): 5-34, 2009.

[7] Frota, S., "Prosodic structure, constituents and their implementation", in A.C. Cohn et al [Eds], The Oxford Handbook of Laboratory Phonology, 254-265, Oxford University Press, 2012.

[8] Dehé, N., "To delete or not to delete: The contexts of Icelandic Final Vowel Deletion", Lingua 118(5): 732-753, 2008.

[9] Selkirk, E.O., "Prosodic domains in phonology: Sanskrit revisited", in M. Aronoff and M.-L. Kean [Eds], Juncture: A collection of original papers, 107-129, Anma Libri, 1980.

[10] Nespor, M. and Vogel, I., Prosodic Phonology. Foris, 1986.

[11] Selkirk, E., "On clause and intonational phrase in Japanese: The syntactic grounding of prosodic constituent structure", Gengo Kenkyu 136: 35-73, 2009.

[12] Selkirk, E., "The syntax-phonology interface", in J. Goldsmith, J. Riggle and A.C.L. Yu (Eds), The Handbook of Phonological Theory, 2nd edn, 435-484, Wiley-Blackwell, 2011.

[13] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]: http://www.praat.org/, last access 4 Dec 2013.

[14] Burquest, D.A., Phonological Analysis: A Functional Approach, The Summer Institute of Linguistics, 1998.

[15] Árnason, K., "The standard languages and their systems in the 20th century I: Icelandic", in O. Bandle et al [Eds], The Nordic Languages: An International Handbook of the History of the North Germanic Languages, Vol. 2, 1560-1573, Walter de Gruyter, 2005.

[16] Árnason, K., "Icelandic word stress and metrical phonology", Studia Linguistica 39: 93-129, 1985.

[17] Thráinsson, K. & Árnason, K., "Phonological variation in 20th century Icelandic", Íslenskt mál 14: 89-128, 1992.

The Realization of French Rising Intonation by Speakers of American English

Scott Lee

[1] Program in Linguistics, University of Georgia, Athens, Georgia, United States

gte577z@uga.edu

Abstract

This study examines the realization of French intonational rises by adult native speakers of American English. Production data were gathered using a discourse completion task and a storytelling task from eight American college students beginning a semester-long study abroad program in Southern France. Results suggest that speakers struggled with two particular aspects of French intonation: the grouping of words into Accentual Phrases, and the phonetic realization of phrase-final rises. In particular, the probability distribution for the alignment of the late L elbow was bimodal for L2 speakers but unimodal for L1 speakers, suggesting the use in the learner speech of two distinct tonal patterns instead of the single French LH*. Mean values for overall pitch range and the scaling of continuative rises were significantly lower and less variable than French L1 values as well.

Index Terms: intonation, L2 prosody, French

1. Introduction

Most studies on second language acquisition have focused on the production and perception of segmental material [see e.g. 1, 2, 3]. To gain a better understanding of the phonetic characteristics of foreign accent, though, it is necessary to explore prosodic as well as segmental characteristics of learner speech. A number of recent studies have addressed this issue, exploring features like tonal alignment [4], tonal phrasing [5], and pitch scaling [6]. This study continues this trend by exploring the intonational characteristics of L2 French spoken by native speakers of American English. In prosodic terms, the two languages have a number of important differences, notably that stress is distinctive in English but not in French, and that intonation phrases (IPs) are built from accentual phrases (APs) with a default tonal pattern in French, instead of from a series of pitch accents as they are in English. The main goal of this study is to examine how L1 English speakers navigate these differences in their spoken L2 French, which will hopefully lend support to previous research findings, as well as shed light on interesting directions for future research on the acquisition of L2 prosody.

1.1. Intonation models

Although the scope of this paper is primarily phonetic, it is informed by the intonational phonologies of the languages under investigation. The models of intonation assumed by this paper are the Autosegmental Metrical (AM) systems developed for English by Pierrehumbert and Beckman [7] and for French by Jun and Fougeron [8, 9]. The two systems have a few structural differences that are relevant to this study. First, IPs in French are built from APs, which have a default /LHiLH*/ tonal sequence. The phonetic realization of this sequence depends on a number of factors, like speech rate, speech style, and how many syllables are in the phrase. By contrast, IPs in English are built from a sequence of pitch accents on stressed syllables, with nuclear or phrase-final accents being followed by a phrase accent and boundary tone. A second important difference is that English has two bitonal pitch accents, while French only has one. A second rise has been proposed for French [10], but the accent spans two syllables, with the L target appearing on the penultimate syllable and the H target on the final syllable. In English, however, the targets for both pitch accents may associate with a single stressed syllable, producing a contour that is sometimes similar to the French contour (in the case of L+H*) but sometimes not (in the case of L*+H). For the latter, the L target aligns with the stressed syllable, and the H target is realized somewhere near the following syllable boundary, producing a "scooped" shape with a relatively pronounced dip in f_0.

1.1.1. Alignment of the late rise in French

In her study of French tonal structure, Welby [11] reported values for the alignment of tonal targets in the early and late rises. In general, late H* was aligned with the last full syllable of the phrase, and the late L was realized close to the preceding syllable boundary. The position of the late L was not correlated with the duration of either syllable (penultimate or final), but it appeared in the final syllable 82% of the time. Her conclusions support an AM analysis of French intonation, and they support earlier analyses [see e.g. 12] of the late rise as a bitonal pitch accent.

1.2. Previous research

1.2.1. Korean and English

Jun and Oh [5] looked at how native speakers of American English acquired the phonology and phonetics of Korean intonation, which shares important phonological characteristics with French intonation. As in French, Korean IPs are built from APs, and the APs are built from rising sequences of L and H tones, with the exact tonal specification depending on the number of syllables in the phrase. To explore the relationship between the speakers' level of experience with Korean and their spoken proficiency, Jun and Oh designed a set of 40 sentences to test two main intonational features: the grouping of words into APs, and the phonetic realization of tone sequences. Interestingly, their results show that although advanced speakers produced more correct phrasings than intermediate and beginning speakers, they were generally not more successful at realizing the underlying tonal sequences phonetically. Specifically, they found that pitch range in AP-initial HL sequences was significantly smaller for learners than for native speakers, and that AP-final H tones were easier for learners to accurately produce than those elsewhere in the phrase, presumably because they are produced more regularly in L1 Korean and are thus more perceptually salient.

1.2.2. Dutch and Greek

Mennen [1] examined bi-directional intonational transfer in Dutch non-native speakers of Greek. Prenuclear or non-final rises have the same phonological structure in both languages, but they differ in their phonetic realization, with peak alignment not only occurring earlier but also being affected by vowel length in Dutch. The study consisted of two experiments, one looking for transfer from the L1 to the L2, and one looking for transfer from the L2 to the L1. Results for the first experiment showed that only one of the speakers was able to produce native-like L2 rises; the remaining four speakers produced L2 rises with alignment patterns from the L1. Results for the second experiment showed that the same four speakers produced L1 alignment patterns that were significantly different from those produced by monolingual L1 speakers, indicating an effect of exposure to the L2. Together, the experiments support the claim that intonational transfer can go both ways, i.e. from the L1 to the L2 and from the L2 to the L1. Crucially, they also support Jun and Oh's findings that the phonetics of L2 intonation are difficult for speakers to acquire.

1.3. Research questions

Using AP-final rises in French as a test case, the production experiment was designed to answer a number of questions raised by previous studies about the acquisition of L2 intonation. First, do learners acquire some aspects of L2 intonation (e.g. phrasing) more easily than others, and if so, which ones? If Jun and Oh's results hold true for the speakers in this study, then we would expect them to have more success grouping words into APs than realizing tonal sequences with native-like phonetics. Second, are phonologically similar tonal sequences realized differently in the learners' L2 than they are in their L1, and if so, how? French provides a good platform for testing both of these questions, since it is phonologically and phonetically different from English in terms of its intonation.

2. Methods

2.1. Participants

Production data were gathered from eight adult native speakers of American English participating in a study abroad program in Montpellier, France. This paper reports the results of the pre-departure pilot study that was conducted immediately before the participants left. Although their proficiency with spoken French varied, all speakers had completed the equivalent of four semesters' (i.e. two years') worth of undergraduate French. Their language backgrounds also varied, with two speakers having completed the same level of coursework or higher in Spanish, and one having working knowledge of Spanish, Italian, and Portuguese; the remaining five had only French as their L2. Production data were then gathered from four of the participants' native French host families for comparison.

2.2. Materials

All participants were asked to complete two tasks: a storytelling task, and a discourse completion task (DCT) as outlined by Prieto [13]. The DCT consisted of 31 scripted sentences elicited from the participants, which they read in response to situational prompts described to them by the interviewer. A sample item is given in (1), where the response is italicized and bolded.

(1) Tu as acheté de la glace à la vanilla et à la noisette pour ta fête. Demande aux invités s'ils veulent de la glace à la vanilla ou à la noisette.

Vous voulez de la glace à la vanilla ou à la noisette?

'You bought two flavors of ice cream for your party, but you're not sure which flavor to serve. Ask the guests at your party which flavor ice cream they'd prefer.

Do you all want chocolate or vanilla?'

31 sentences were recorded per speaker in the first task and approximately two minutes of extemporaneous speech in the second. Learners completed the tasks in both languages, while the host families completed them only in French.

2.3. Procedures

For the discourse completion task, the interviewer read the situational prompt for each item, and the participants read the target sentence in response. For the storytelling task, participants were asked to tell a story about a social event they recently attended, like a family gathering or party. The tasks were completed during a single interview session and were separated by approximately five minutes of time during which the participants could review the instructions for the upcoming task. Instructions for both tasks were printed on a prompt sheet and were explained by the interviewer at the beginning of the session.

2.3.1. Data analysis

Speaker responses to both tasks were recorded using a Shure SM51 condenser microphone and digitized at a sampling rate of 44.1khz and a depth of 24kbit/s. Responses to the DCT were segmented by hand and saved as separate files, yielding a total of 496 (16 x 31) phrases of varying length and tonal structure. Responses to the storytelling task were saved as a single wav file. Sound files for both tasks were then loaded into Praat [14] and segmented into phrases, words, syllables, and phones. Praat scripts were used to align phone boundaries to the text, and the boundaries were adjusted by hand to ensure accuracy.

Pitch curves were generated using Praat's built-in pitch tracker, and files containing f_0 perturbations or disfluencies were discarded. Several intonational features were then labeled by hand for each phrase, including f_0 minima and maxima, position of the late H, and position of the late L elbow. AP and IP boundaries were also marked in the French data to allow for the comparison of L1 and L2 phrasing. For the L2 speech, boundaries were marked both when speakers produced grammatical phrase-final pitch accents, and when they produced ungrammatical phrase-medial pitch accents.

Numerical data were exported to a spreadsheet and then analyzed using the R statistical package [15].

2.3.2. ToBI labeling

Learner varieties of intonation have been shown to be phonologically distinct from native varieties [16]. In this study, the speakers produced rises in the L2 that were sometimes phonetically similar to analogous rises in an L1, but sometimes not. Because of this ambiguity, tonal categories were not labeled in the L2 and were analyzed phonetically instead.

3. Results

3.1. Phrasing

Data from the discourse completion task show that learners produced more sentence-medial AP boundaries than native speakers. Mean counts for both boundary types are presented in the table below and organized by speaker group. Values for the French L1 speakers were tightly clustered (σ=3.7) and are thus reported as an aggregate mean, but values for the L2 speakers were more widely dispersed (σ=11.38). A Welch's *t*-test indicated that the difference between the means was statistically significant ($p<0.0002$), supporting Jun and Oh's findings for Korean that non-native speakers of French are generally less successful than native speakers at grouping words into phrases.

Speaker	French L1	French L2
1	78	85
2	77	117
3	80	108
4	73	115
5	74	96
6	82	94
7	75	100
8	71	112
	μ=76.3 σ=3.7	μ=103.4 σ=11.38

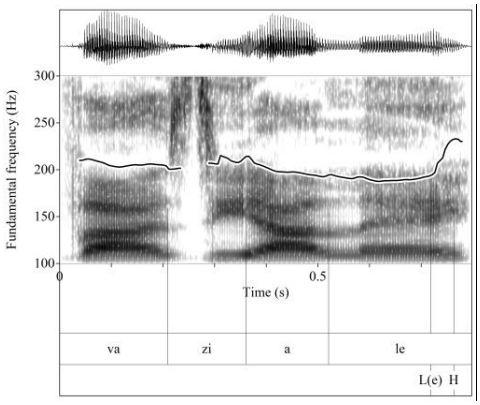
Table 1. *Number of sentence-medial APs produced by native and non-native speakers in the DCT.*

These results are likely due to two factors: the tendency for the non-native speakers to produce ungrammatical phrase-medial pitch accents, and their tendency to speak slower and with more disfluencies than the native speakers.

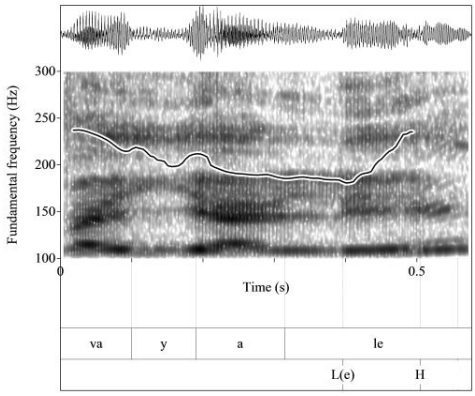
3.2. Tonal alignment

Results from both tasks show significant differences in the phonetic realization of rising tone sequences between the native and non-native speakers. In general, the L2 speakers realized the elbow in the late rise later than the L1 speakers. The placement of this elbow was also more variable. Figure 1 shows a spectrogram and pitch curve for the final four syllables of a sentence-medial AP in the phrase *Comment tu vas y aller* 'How are you getting there?' produced by a native (a) and a non-native (b) speaker. The low elbow in each figure is marked by an L(e), and f_0 is shown in hertz (Hz).

Figure 1. *Alignment of late rises in L1 and L2 French speech.*
a.



b.



The pitch curve in (b) is representative of a typical alignment pattern for late rises reported in [11], with the L elbow being located near the onset of the final syllable and the H target being reached near its end. The slope of the curve increases sharply about halfway through the syllable, but it is clearly rising throughout. By contrast, the pitch curve in (a) is marked by a much later rise in the final syllable, with the L elbow falling close to the H targetr near the end of the syllable. This pattern was common in the L2 speech, accounting for 45% of the total rises, and was in many cases produced with an even more pronounced delay.

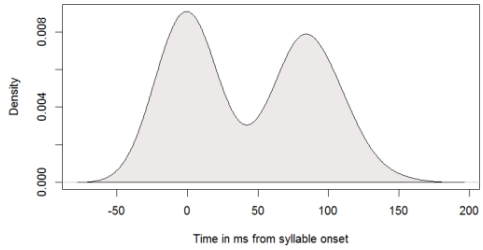
Table 2 presents the mean alignment in ms of the late elbow relative to the onset of the accented syllable. In calculating the means, positive values were entered for positions in the target syllable and negative values for positions in the preceding syllable. The French L2 values show late mean alignment relative to the French L1 values, which are closer to the syllable onset and similar to those reported by Welby [11].

Speaker	French L2 (μ, σ)		French L1 (μ, σ)	
1	32.1	39.3	10.5	8.3
2	51.6	58.6	15.6	10.3
3	45.3	33.2	8.5	9.6
4	39.8	37.4	21.3	11.4
5	52.5	49.3	12.3	9.3
6	61.8	56.7	11.2	14.2
7	33.1	36.5	9.3	16.3
8	71.0	58.8	18.1	15.8
All	48.4	46.3	13.35	11.9

Table 2. *Alignment in* ms *of the L-elbow in AP-final rises for French L2 and L1 speakers.*

The trend that emerges from these data is the close relationship between the mean and standard deviation for the L2 speakers' alignment patterns. In effect, the distribution was approximately bimodal, suggesting the speakers chose one of two options when realizing the rise: early alignment of the L, or late alignment of the L. In cases of early alignment, the elbow most often fell and occasionally before the syllable boundary. The alignment of the late H was relatively invariable, almost always occurring near the syllable boundary (mean latency from the right edge of the phrase was 1.2ms for all speakers, with σ =.53). A kernel estimation was used to calculate the probability distribution function for the alignment of the late L, shown in Figure 2. In addition to the bimodality, the distribution highlights the fact that the speakers tended to produce more early- than late-aligned L targets, on the whole.

Figure 2. *Probability density function for the alignment of the late L elbow in L2 speech.*



3.3. Pitch scaling

Overall pitch range, measured as 80% quantal range in Equivalent Rectangular Bandwidths (ERB) was lower for the learners (μ=2.1ERB) than for the native speakers (μ=4.3ERB); the difference was statistically significant (p=0.002). After z-score normalization, average values for the scaling of the late rises were shown to be significantly less variable for the learners as well, perhaps indicating a restriction in expressivity caused by a general processing constraint limiting the complexity of syntactic and pragmatic information that they can prosodically code.

4. Discussion and conclusion

The production data gathered for this study clearly suggest the presence of interlanguage effects in the non-native speakers' French intonation. In particular, the alignment and scaling of L2 AP-final rises differs significantly from that of the same rises in L1 French. One potential explanation for this difference is phonological. The L2 speakers essentially had two categorical choices when producing the AP-final rise: the scooped bitonal accent L*+H, and either the LH* or the L+H* (the phonetic distinction between these two was not clear in the data). If adult speakers can acquire L2 intonational phonetics, then we would expect advanced learners to produce mostly LH* and intermediate and beginning learners to alternate between the LH* and the L*+H. The results give some support to this hypothesis, with the learners, who self-rated their spoken proficiency in French as intermediate, producing the two contours fairly evenly.

The factors underlying this phonological interference are likely complex. Perceptual factors may play a role if the learners hear the French contour as a generic rise without perceiving (or perhaps recognizing as linguistically relevant) how consistently the tonal targets are aligned in L1 speech. However, pragmatic factors may also be involved, since the two pitch accents are generally assumed to have distinct meanings in English, with the scooped accent in particular signaling uncertainty or hesitation [see e.g. 17,18]; although the alignment of the late L elbow is potentially influenced by pragmatic factors [19], the variation does not appear to be as clearly categorical as it is in English. Perceptual and pragmatic factors are of course not mutually exclusive, and more investigation is needed to determine to what extent they both contribute to the phonological and phonetic characteristics of the learners' L2 intonation.

The findings suggest a number of directions for future research. First, production data from English are needed to determine whether rising intonation in the learners' L1 is affected by their exposure to the L2. Second, perception experiments are needed to look for interference between the learners' L1 and L2, e.g. whether they perceive L2 rises as distinct from L1 rises. Production data from the participants in this study will be gathered after they return from France to examine the effects of prolonged exposure to the L2 on their speech and add to the growing body of longitudinal research on prosodic acquisition. Based on the findings in [5] and [16], the expected result is for the speakers to make improvements to their phrasing, but not necessarily to the realization of the tone sequences. This last component also has the benefit of indirectly testing the effectiveness of language immersion as a pedagogical technique for improving L2 prosody, which is a relatively under-researched area of second language acquisition.

5. Acknowledgements

I would like to thank Meghan Armstrong and Keith Langston for their guidance and patience in helping me design the experiments for this study. I would also like to thank Diana Ranson for nurturing my interest in French, and the University of Georgia Graduate School and Office for the Vice President of Research for their generous financial support. Finally, thanks to Pauline Welby for creating the Praat script used to draw the pitch tracks in this paper.

6. References

[1] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233-277.

[2] Best, C. T. (1995). Learning to perceive the sound pattern of English. *Advances in infancy research*, *9*, 217-217.

[3] Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47-B57.

[4] Mennen, I. (2004). Bi-directional interference in the intonation of Dutch speakers of Greek. *Journal of Phonetics*, *32*(4), 543-563.

[5] Jun, S. A., & Oh, M. (2000, May). Acquisition of second language intonation. In *INTERSPEECH* (pp. 73-76).

[6] Mennen, I., Schaeffler, F., & Docherty, G. (2007). Pitching it differently: A comparison of the pitch ranges of German and

English speakers. *16th International Congress of Phonetic Sciences*, 1769-1772.
[7] Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology yearbook*, *3*(1), 5-70.
[8] Jun, S. A., & Fougeron, C. (2000). A phonological model of French intonation. In *Intonation* (pp. 209-242). Springer Netherlands.
[9] Jun, S. A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus*, *14*(1), 147-172.
[10] Delais-Roussarie, Elisabeth; Post, Brechtje; Avanzi, Mathieu; Buthke, Carolin; Di Cristo, Albert; Feldhausen, Ingo; Jun, Sun-Ah; Martin, Philippe; Meisenburg, Trudel; Rialland, Annie; Sichel-Bazin, Rafèu & Yoo, Hi-Yon (to appear: 2014). "Intonational phonology of French: Developing a ToBI system for French". In Frota, Sónia & Prieto, Pilar (eds.).*Intonational variation in Romance*. Oxford: Oxford University Press.
[11] Welby, P. (2004). The structure of French intonational rises: A study of text-to-tune alignment. In *Speech Prosody 2004, International Conference*.
[12] Post, B. (2000). *Tonal and phrasal structures in French intonation* (Vol. 34). Thesus.
[13] Prieto, Pilar (2001). 'L'entonació dialectal del català: El cas de les frases interrogatives absolutes', in A. Bover, M.-R. Lloret, and M. Vidal-Tibbits (eds.), *Actes del Novè Colloqui d'Estudis Catalans a Nord-Amèrica*. Barcelona: Publicacions de l'Abadia de Montserrat, 347-377.
[14] Boersma, Paul & Weenink, David (2014). Praat: doing phonetics by computer [Computer program].
[15] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
[16] Mennen, I., Chen, A., & Karlsson, F. (2010). Characterising the internal structure of learner intonation and its development over time. In Proceedings of The 6th International Symposium on the Acquisition of Second Language Speech (Newsounds 2010).
[17] Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 747-776.
[18] Hirschberg, J. (2004). Pragmatics and intonation. *The handbook of pragmatics*, 515-537.
[19] Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, *34*(3), 343-371.

Monosyllabic Lexical Pitch Contrasts in Norwegian

Niamh Kelly[1], *Rajka Smiljanić*[2]

[1,2]Department of Linguistics, The University of Texas at Austin, USA

niamh.kelly@utexas.edu, rajka@austin.utexas.edu

Abstract

This paper examines the lexical tonal accent contrast in monosyllabic words in the Trøndersk dialect of Norwegian. The results of a production experiment in which speakers produced the unmarked accent and the circumflex accent showed that the tonal distinction is characterised by a difference in f0 maximum, f0 height at onset, f0 minimum and its timing, and height of the final Accent Phrase H tone. The presence of the tonal accent contrast on monosyllabic words is unusual among dialects of Norwegian and Swedish.

Index Terms: lexical pitch accent, Norwegian, monosyllabic

1. Introduction

Scandinavian tonal accent contrasts have been described extensively, both impressionistically and experimentally [1–10]. The two contrasting tonal accents in Norwegian and Swedish are referred to as accent 1 and accent 2 [11]. Depending on the variety in question, the tonal contours of the two contrastive accents of Swedish and Norwegian may differ in their tonal makeup (e.g., LH vs HLH) or they may have the same tones but different timing in relation to the segmental string (e.g., both HLH) [12]. An example from the Oslo variety of Norwegian of words distinguished only by the tonal accent is *bønder* "farmers" (accent 1, LH) and *bønner* "beans" (accent 2, HLH) (the segments of both words are pronounced /ˈbønːər/).

Researchers agree that the Scandinavian accent contrast is generally only found on polysyllabic words [6, 13, 14] and some regard all monosyllabic words as having accent 1 [15, 16]. One explanation for this difference - likely related to the origin of the contrast - is that since accent 2 has a later timing, it needs a second syllable in order for the later tones to surface [13]. This is in contrast with tonal languages from a wide variety of language groups such as Burmese, Mandarin, !Xoo or Mixtec, where each syllable can bear a tone [17], but similar to other pitch accent languages such as Basque [18, 19] or Serbo-Croatian [20]. However, a small number of dialects of Norwegian and Swedish, including some of the Trøndersk dialects, have been described as having a tonal contrast surfacing on monosyllabic words, due to apocope, final vowel deletion [21–23]. This is referred to as a contrast between the unmarked monosyllabic accent and the circumflex accent and describes a difference in the shape of the contour between the accents. While previous research focused on the characteristics of the lexical contrast in polysyllabic words, only few acoustic studies examined in detail the nature of the contrast in monosyllabic words [23, 24]. Furthermore, even within the Trøndelag dialectal region, not all varieties (e.g., the city of Trondheim) have the contrast on monosyllabic words (Gjert Kristoffersen, p.c.). Given the rather limited number of monosyllabic minimal pairs that differ only in the accent feature, it is possible that this phenomenon will disappear. One aim of this study was thus to document this prosodically unusual variety.

The major goal of this work was to investigate whether the lexical tonal accent contrast surfaces on monosyllabic words in the Trøndersk dialect and, furthermore, what the exact acoustic correlates of the contrast are. This dialect is spoken in the Trøndelag region, a variety of East Norwegian spoken in central Norway.

In the Trøndersk dialect, disyllabic words with accent 1 or accent 2 both have the same shape (previously described as HLH [2, 25, 26]) with the tonal contour aligned later in accent 2 words. The circumflex accent occurs on words that are disyllabic in other varieties of Norwegian, such as infinitive forms of verbs [22], for example *å glimte* 'to gleam'. Although the circumflex accent most commonly occurs on words that originally had accent 2, they can also form from accent 1 words [24]. In the Trøndersk variety, these words are generally pronounced without the final (unstressed) vowel, reducing them to monosyllabic words which retain the underlying tones.

A small scale acoustic analysis of a different Trøndersk dialect, that of Ålvundeid, found that the circumflex accent is characterised by a longer vowel and a higher f0 onset than the unmarked monosyllabic accent [24]. In comparison, the unmarked monosyllabic accent is simply an L tone in East Norwegian (Gjert Kristoffersen, p.c.). In the Ålvundeid study, all target words were focused by elicitation. It is therefore not clear which tonal events characterise the accent contrast itself (if any) and which arise due to pragmatic focus. In addition, the Ålvundeid study examined only two male speakers from each of three age groups ('old' (over 70), 'mid' (38 and 50), and 'young' (20/21)). The current study aims to separate the effect of sentence intonation from the lexical tonal accent in order to examine the features that characterise the contrast on monosyllabic words in ten speakers of the Trøndersk dialect. Specifically, recordings were made of native speakers of the Trøndersk dialect reading sentences containing monosyllabic words with either the unmarked accent or the circumflex accent, in neutral (non-focused) intonation.

It was expected that the circumflex accent would have a longer vowel, higher f0 at vowel onset and later alignment of tones than the unmarked monosyllabic words.

2. Methods

2.1. Speakers

Ten native speakers (four males, six females) aged 18-45 of Trøndersk were recorded reading sentences containing the target words. The speakers were from a variety of towns south and west of Trondheim, where the circumflex accent is known to occur (Stian Hårstad and Jørn Almberg, p.c.), such as Tingvoll, Oppdal, Rennebu, Surnadal, Sunndal.

2.2. Stimuli

Target words contained only voiced sonorant consonants next to the vowel, such as *smil* "smile" (unmarked) and *'smile* "to smile" (circumflex). (Recall that, in this dialect, the final *e* is apocopated.) The words all contained the vowel /i/.

In Norwegian, the accent phrase (AP) high boundary tone (H%) [13, 27–29] is associated with the final unstressed syllable in the AP [29]. In order to control for the effect of this H%, the target word was always two unstressed syllables before the end of an AP. The target word also followed two or three unstressed syllables at the beginning of the sentence, which were AP-external [26]. In this way, the accent of the target word avoided being affected by either the accent or the final H% of a preceding AP [27]. The target word was always followed by a contrastively focused word in the following AP, in order to ensure that the target word did not receive narrow focus.

Examples (with target words in bold): (AP = accent phrase, IP = intonational phrase, IU = intonational utterance)

Unmarked:
Det var et lim i en film, men ikke i et stykke.
((Det var et (**lim** i en)$_{AP}$ (film,)$_{AP}$)$_{IP}$, (men ikke i et (stykke)$_{AP}$)$_{IP}$)$_{IU}$
"There was glue in a film, but not in a play."

Circumflex:
Jeg vil smile i en film, men ikke i et bilde.
((Jeg vil (**smile** i en)$_{AP}$ (film)$_{AP}$)$_{IP}$, (men ikke (i et bilde)$_{AP}$)$_{IP}$)$_{IU}$
"I want to smile in a film, but not in a photo."

For each accent there were five target words, each produced three times, giving 15 tokens per accent per speaker, a total of 300 tokens (15 tokens x 2 accents x 10 speakers). The words of the different accents were minimal pairs, where the accent was the only difference between the words.

2.3. Procedure

Speakers were recorded in the phonetics studio at the National University of Science and Technology (NTNU) in Trondheim, Norway. The experimenter was the first author. The sentences were presented one by one on powerpoint slides, with the speaker in control of when to move to the next sentence. The sentences were randomised and interspersed with sentences containing disyllabic target words. The sentences were in the same order for all speakers. They were written in the standard Bokmål orthography and also in a transcription of Trøndersk. Speakers were asked to speak as they would at home. None of the speakers had difficulty following the instructions.

2.4. Measurements and Analysis

In order to fully examine the F0 contour and its alignment with the segmental string, the following measurements were obtained (Figure 1 depicts and defines all the labelled landmarks):

1. F0 maximum (H)
2. F0 minimum (L)
3. F0 maximum timing ((H-V1)/(Vowel duration))
4. F0 minimum timing ((L-V1)/(Vowel duration))
5. F0 height at vowel onset (V1)
6. Vowel duration (C3-V1)
7. AP H% tone height (APH)
8. Boundary slope ((APH-LTP)f0/(APH-LTP)time)

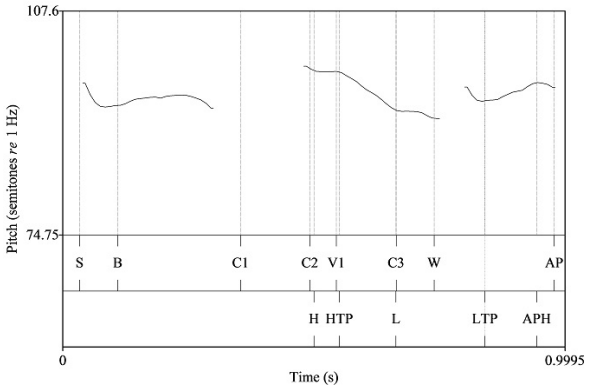


Figure 1: *Circumflex f0 contour of the word 'smile' showing measurement points.* S = beginning of the sentence; B = beginning of f0 rise; C1 = onset of target word; C2 = onset of second consonant (if present); V1 = vowel onset; C3 = onset of post-vocalic consonant; W = end of target word; AP = end of AP; H = f0 maximum, HTP = turning point from f0 maximum; L = f0 minimum; LTP = turning point from f0 minimum; APH = AP boundary tone.

All measurements were made on the target word. Duration was measured in milliseconds and f0 was measured in semitones in Praat [30]. Timing of f0 maximum and minimum were measured in milliseconds relative to vowel onset. These timings were then divided by the duration of the vowel, to get relative timing. This controlled for speaking rate differences. Boundary slope was measured as the difference in pitch between the turning point of the f0 minimum and the following AP boundary H%, divided by the duration between these two points.

Statistical tests consisted of a mixed model multiple linear regression analysis, conducted using the lme4 package in R [31]. The independent variable was accent and the dependent variables were the measures (f0 maximum and minimum and their timing, f0 at vowel onset, vowel duration, AP H% height, and slope of the rise to the AP-boundary tone). Speaker was included as a random effect. The reference level for accent was the unmarked accent. Since linear regressions do not produce *p*-values, the significance of accent as a predictor of each measure was calculated by a likelihood ratio test comparing a model that included accent as a predictor and one that did not [32]. This was conducted using the anova function in R, a likelihood ratio test for nested models.

3. Results

An average, time-normalised pitch track [33] for all target word productions by one female speaker (S16F) for the two accents is shown in Figure 2. This is representative of the other speakers. From this figure, it can be seen that the two accents are distinct. They differ in shape with the circumflex (blue line) having a higher f0 maximum and later f0 minimum timing than the unmarked accent. Furthermore, the circumflex accent has a higher f0 at onset.

Table 1 shows the statistical test results for each measure. The coefficient and t-values are from the linear regression tests

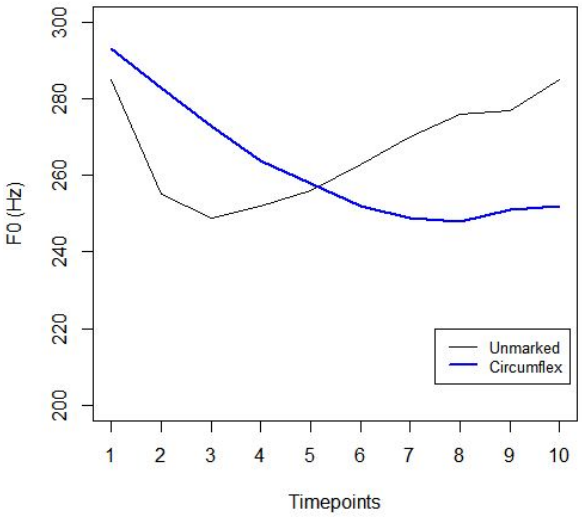


Figure 2: *Average, time-normalised f0 contour for one speaker.*

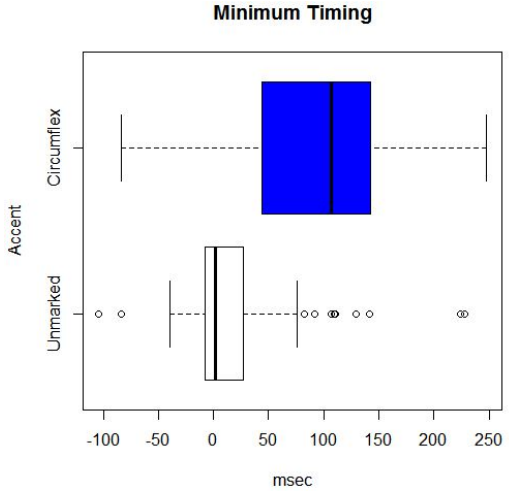


Figure 3: *F0 minimum timing by accent (unmarked in white, circumflex in blue).*

and the *p*-value is from the likelihood ratio test. A significant result (*) indicates that accent is a significant predictor of the measure. The sign of the coefficient tells the direction of the difference. Since the unmarked accent was the reference level, a positive coefficient means that for that measure, the circumflex accent has a higher average value than the unmarked accent, and a negative coefficient means the circumflex accent has a lower average value than the unmarked accent. For example, f0 minimum has a negative coefficient, meaning that the circumflex accent has a lower average f0 minimum than the unmarked accent.

Table 1: *Regression and likelihood ratio test results for each measure.*

Measure	Coef.	t-value	*p*-value
Vowel duration (msec)	-0.67	-0.12	0.906
F0 Maximum (st)	1.41	8.5	p<0.001*
F0 Minimum (st)	-0.47	-2.8	p<0.01*
F0 vowel onset (st)	1.71	9.9	p<0.001*
F0 max timing	-0.004	-0.07	0.956
F0 min timing	0.54	5.2	p<0.001*
Boundary slope	0.002	1.1	0.256
AP H% height	-0.5	3.3	p<0.01*

The results revealed that the two accents are differentiated by f0 maximum and f0 minimum height, f0 height at vowel onset, timing of f0 minimum (relative to vowel onset) and AP H% height. Table 2 shows the averages and standard deviations for each of these measures by speaker. (A negative number for timing means the f0 minimum occurred before vowel onset.)

The circumflex accent has a higher f0 maximum and f0 height at vowel onset, lower f0 minimum and AP H% tone, and later f0 minimum timing than the unmarked accent. Figure 3 shows the raw f0 minimum timing results pooled across speakers. The average f0 minimum timing for the unmarked accent is just after vowel onset, while that for the circumflex accent is over 100 milliseconds into the vowel.

4. Discussion

The results of this experiment confirmed that differences between the unmarked monosyllabic accent and the circumflex accent are present in Trøndersk. This is unlike many other varieties of Norwegian and Swedish in which this contrast is not found on monosyllabic words. The circumflex accent has a higher f0 maximum and onset, lower f0 minimum, later f0 minimum timing and lower AP H% tone than the unmarked monosyllabic accent. This indicates a wider pitch range and later timing for the circumflex accent. It should be noted that the height of the f0 maximum was not consistently found for all speakers and all words, and further testing showed that there were significant differences among the speakers for this measure. This suggests that this may not be as reliable a cue as the others for signalling the contrast. One reason for this was that for many tokens of the unmarked monosyllabic accent, there was no consistent point at which to mark the f0 maximum. The difficulty in reliably identifying the f0 maximum is consistent with the description of this accent as being just an L tone in this dialect.

While a previous acoustic study on the monosyllabic contrast [24] found vowel duration differences for the two accent types, speakers in this study did not use temporal cues to implement the lexical contrast. In the Oppdal dialect, there is a longer vowel in the circumflex accent, leading to an analysis of these syllables as being trimoraic [23]. The results from the current study suggest that this is not the case for all dialects of Trøndersk. The current results demonstrate that the circumflex accent can be realised without increasing vowel length.

It was also expected that the circumflex accent should have a steeper slope from the f0 minimum to the AP H% tone, due to the later timing of the f0 minimum for this accent; however, this was not found. The results instead revealed that the circumflex accent has a lower phrase boundary tone (H%) which in turn affected the slope. The circumflex accent thus has a later timing, lower f0 minimum and a lower AP H% tone. The lower AP H% tone could be due to the fact that the speaker simply does not have time to reach the natural H% target for the circumflex accent; however, this explanation seems unlikely, since the two

Table 2: *Mean results for each significant measure by speaker.* (The speakers marked F are female and M are male.) Results are in the form: Mean (SD)

	F0 max.		F0 min.		F0 onset		Minimum timing		APH	
Speaker	Unmark.	Circ.	Unmark.	Circ.	Unmark.	Circ.	Unmark.	Circ.	Unmark.	Circ.
S01F	95 (0.6)	96 (1.7)	93 (0.7)	94 (0.7)	93 (0.6)	95 (2.3)	-0.1 (0.1)	0.5 (0.7)	97 (0.4)	97(1.2)
S02F	95 (0.6)	99 (0.6)	92 (0.7)	91 (1.1)	93 (0.3)	99 (0.6)	0.9 (0.6)	0.9 (0.2)	98 (1)	97 (1.7)
S04F	94 (0.7)	94 (0.5)	92 (0.7)	93 (0.5)	92 (0.3)	93 (0.8)	0 (0.1)	0.1 (0.1)	96 (0.7)	96 (0.7)
S07M	86 (1.3)	89 (1.3)	88 (2.4)	87 (1.8)	88 (1.2)	89 (0.7)	-0.2 (0.4)	-0.4 (1)	90 (0.6)	89 (0.8)
S09M	86 (1.2)	88 (1.1)	84 (1.2)	83 (0.5)	85 (1.1)	86 (0.8)	-0.2 (0.4)	0.6 (0.3)	87 (0.8)	87 (1)
S10F	92 (0.9)	93 (1.3)	90 (0.7)	90 (0.9)	91 (0.5)	92 (1)	0.3 (0.4)	1.3 (1.2)	92 (0.5)	92 (0.8)
S12F	95 (0.7)	95 (0.7)	94 (0.6)	94 (0.5)	94 (0.6)	95 (0.8)	0.2 (0.3)	0.1 (0.3)	96 (0.8)	96 (2.1)
S13M	89 (1)	89 (0.8)	88 (1.2)	87 (1.2)	88 (0.7)	88 (0.9)	-0.1 (0.1)	0.8 (1.4)	90 (0.8)	89 (0.8)
S15M	83 (0.4)	85 (1.2)	82 (0.5)	80 (1.2)	82 (0.5)	84 (1)	0 (0)	1.2 (0.4)	86 (0.4)	82 (1.4)
S16F	95 (1)	98 (0.6)	96 (1)	95 (1)	96 (0.8)	97 (0.8)	0 (0.1)	0.1 (0.3)	98 (0.5)	97 (0.8)
Average	92	93	90	89	91	92	0.1	0.5	93	92

following syllables should be enough to reach this target, and the f0 contour did not reach any higher point after the AP.

The work presented here offers evidence of the lexical tonal accent contrast in monosyllabic words in Trøndersk. Future research will examine the realisation of both the unmarked monosyllabic accent and the circumflex accent in AP-final position and in a focus context, to determine how the contour is affected by sentence intonation and pragmatic factors. A comparison of those results with the findings of the current study will also indicate which cues are the most consistent in realising the monosyllabic contrast. Another direction of future research could be to compare the circumflex contour with that of accent 1 and 2 disyllabic words (depending on the accent of the word before apocopation), to examine whether the circumflex accent is indeed realised as a compressed form of the disyllabic tonal contour. This type of comparison could help to further elucidate the features that characterise the circumflex accent.

5. Conclusion

This investigation has shown that the circumflex accent is significantly different from the unmarked accent in monosyllabic words in the Trøndersk dialect of Norwegian. The acoustic analysis provides evidence that there is in fact a tonal accent contrast on monosyllabic words in this variety, at least in production. It remains to be determined whether listeners can perceive this tonal contrast on monosyllabic words and, furthermore, which of the cues indicated here are most salient for differentiating the accents.

This work provides an acoustic analysis of an unusual monosyllabic contrast in Scandinavian. Future work will examine how this contrast is affected by sentence intonation, thereby adding insight to how the lexical and phrasal levels of intonation interact.

6. Acknowledgements

We would like to thank Gjert Kristoffersen and his collaborators for their huge assistance in choosing target words and to Gjert for proof-reading the sentences and transcribing them into the dialect, as well as for comments on this paper. Special thanks to Wim van Dommelen for the use of the recording studio at NTNU, Trondheim. We also thank the anonymous reviewers for their helpful comments. The research used for this investigation was conducted with the support of the National Science Foundation Doctoral Dissertation Research Improvement Grant No. 1322700.

7. References

[1] J. Storm, *Norvegia. Tidsskrift for det norske folks maal og minder*. Kristiania: Grøndahl and Søn, 1884.

[2] K. Fintoft, *Acoustical Analysis and Perception of Tonemes in Some Norwegian Dialects*. Oslo: Universitetsforlaget, 1970.

[3] E. Gårding, "The Scandinavian word accents," in *Working Papers 8*. Phonetics Laboratory, Lund University, 1973.

[4] E. Gårding and P. Lindblad, "Constancy and variation in Swedish word accent patterns," in *Working Papers 7*. Phonetics Laboratory, Lund University, 1975, pp. 36–100.

[5] O. Lorentz, "Adding tone to tone in Scandinavian dialects," in *Nordic Prosody II*, T. Fretheim, Ed. Trondheim: Tapir, 1981, pp. 166–80.

[6] G. Kristoffersen, *The Phonology of Norwegian*. Oxford: Oxford University Press, 2000.

[7] W. A. Van Dommelen, "Toneme realization in two North Norwegian dialects," *Proceedings of Fonetik*, vol. 44(1), pp. 21–24, 2002.

[8] M. Segerup, "Word accent gestures in West Swedish," in *Proceedings from Fonetik 2003; Phonum 9*, M. Heldner, Ed., Univ. Umeå, 2003, pp. 25–28.

[9] J. Almberg, "Tonal differences between four Norwegian dialect regions - some acoustic findings," in *Nordic Prosody IX*, G. Bruce and M. Horne, Eds. Lund: Peter Lang, 2004, pp. 19–28.

[10] T. Riad, "Scandinavian accent typology," *Sprachtyp. Univ. Forsch. (STUF), Berlin*, vol. 59(1), pp. 36–55, 2006.

[11] G. Bruce, "Swedish word accents in sentence perspective," *Travaux de L'Institut de Linguistique de Lund*, vol. 12, 1977.

[12] C. Gussenhoven, *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press, 2004.

[13] E. Haugen and M. Joos, "Tone and Intonation in East Norwegian," *Acta Philologica Scandinavica*, vol. 22, pp. 41–64, 1952.

[14] T. Riad, "Towards a Scandinavian accent typology," in *Phonology and morphology of the Germanic languages (Linguistische Arbeiten 386)*. Tübingen: Niemeyer, 1998, pp. 77–109.

[15] E. Haugen, "Pitch accent and tonemic juncture in Scandinavian," in *Prosodi/Prosody*. Oslo: Novus Forlag, 1983, pp. 277–281.

[16] V. Felder, E. Jönsson-Steiner, C. Eulitz, and A. Lahiri, "Asymmetric processing of lexical tonal contrast in Swedish," *Attention, Perception, & Psychophysics*, vol. 71(8), pp. 1890–1899, 2009.

[17] M. Gordon, "A typology of contour tone restrictions," *Studies in Language*, vol. 25, pp. 405–444, 2001.

[18] K. Arregi, "Focus on Basque Movements," Ph.D. dissertation, MIT, 2002.

[19] J. I. Hualde, *Basque phonology*. London: Routledge, 1991.

[20] R. Smiljanić, "Lexical, pragmatic and positional effects on prosody in two dialects of Croatian and Serbian: An acoustic study," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2002.

[21] K. Elstad, "Det nordnorske cirkumsflekstonemet," in *Nordic Prosody*, E. Gårding, Gösta B., and R. Bannert, Eds., Lund, 1978.

[22] G. Kristoffersen, "Cirkumflekstonelaget i norske dialekter, med særling vekt på nordnorsk," *Maal og Minne*, vol. 2, pp. 37–61, 1992.

[23] ——, "Cirkumflekstonelaget i Oppdal," *Norsk Lingvistisk Tidsskrift*, vol. 29, pp. 221–262, 2011.

[24] J. Almberg, "The circumflex tone in a Norwegian dialect," in *Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim, August 19-21, 2000*, W. A. Van Dommelen and T. Fretheim, Eds. Frankfurt, Germany: Peter Lang, 2001, pp. 9–21.

[25] W. A. Van Dommelen and R. A. Nilsen, "Toneme realization in two East Norwegian dialects," *Proceedings of Fonetik, PHONUM*, vol. 9, pp. 21–24, 2003.

[26] G. Kristoffersen, "Tonal melodies and tonal alignment in East Norwegian," in *Nordic Prosody IX*, G. Bruce and M. Horne, Eds. Frankfurt am Main: Peter Lang, 2006.

[27] T. Fretheim, "Phonetically Low Tone-Phonologically High tone, and Vice Versa," *Nordic Journal of Linguistics*, vol. 10, pp. 35–58, 1987.

[28] ——, "Intonational phrases and syntactic focus domains," in *Levels of Linguistic Adaptation*, J. Verschueren, Ed. Amsterdam: John Benjamins, 1991, pp. 81–112.

[29] T. Fretheim and R. A. Nilsen, "Terminal rise and rise-fall tunes in East Norwegian intonation," *Nordic Journal of Linguistics*, vol. 12, pp. 155–181, 1989.

[30] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 5.3.03," http://www.praat.org/, 2011.

[31] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[32] B. Winter. (2013) Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [Online]. Available: http://arxiv.org/pdf/1308.5499.pdf

[33] Y. Xu, "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, 2013, pp. 7–10.

Taiwanese Tone Recognition Using Fractionalized Curve-fitting of Prosodic Features

Yu-lun Hsieh[1], *Ching-ting Chuang*[2], *Feng-fan Hsieh*[2], *Yueh-chin Chang*[2], *Wen-lian Hsu*[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Graduate Institute of Linguistics, National Tsing Hua University, Hsinchu, Taiwan
morphe@iis.sinica.edu.tw, d9644510@oz.nthu.edu.tw, ffhsieh@mx.nthu.edu.tw,
ycchang@mx.nthu.edu.tw, hsu@iis.sinica.edu.tw

Abstract

In this paper, we examined different methods of modeling prosodic features of tones, and their effects on a speaker-independent Taiwanese tone recognition system. Tones can be modeled either by plain or curve-fitted features. Plain features represent the original curve faithfully using pitch values, while curve-fitted features can be thought of as an approximation to the values using mathematical functions, such as a Legendre polynomial. In addition, durational information of tones was also proven effective in previous researches. Thus, we proposed a new approach of modeling Taiwanese tones using curve-fitted features extracted from fractions of the pitch curve, along with duration as an additional prosodic feature. Our experimental results showed that using these features in an SVM classifier could substantially improve the accuracy of tone recognition in Taiwanese. Besides, we provided an empirical perspective for theoretic studies on tonal neutralization.

Index Terms: Taiwanese, tone recognition, prosodic feature

1. Introduction

Sinitic languages such as Mandarin and Taiwanese are famous for their syllabic and tonal characteristic, which is different from western languages, such as English. The same syllable structure can carry different lexical tones to indicate different meanings. Tones provide critical information in speech recognition. It was argued that articulatory features such as segmental information, syllable structures and prosodic features may play an important role in tonal recognition in Mandarin [1].

Taiwanese, a relatively understudied language with fewer resources, is a dialect of the Southern Min languages widely spoken in Taiwan. Generally speaking, the structure of Taiwanese syllables is of the form 'CGVC', where 'C' stands for consonants, 'G' for glides, and 'V' for vowels [2]. Compared to Mandarin, Taiwanese has a more complicated tonal system. It has three tonal height contrasts, while Mandarin, on the other hand, only has two. There are seven tones in Taiwanese. Note that Tone 4 (or the *Yangshang* tone in traditional Chinese phonology terms) is missing because it was diachronically merged with Tones 3 or 6. The corresponding F0 curves of Tone 1 through 3 and 5 through 8, after normalizing the duration of the syllable, are shown in Figure 1.

Among them, Tone 7 and 8 are comprised of a stop consonant coda /p, t, k, ?/, and are called 'checked tones' or 'entering tones'. Durational differences between checked tones and other tones are significant, with checked tones being shorter. Furthermore, it was observed that the duration of a syllable might change drastically depending on different syllable structures. Duration is argued to be an important phonetic cue even in discourse [3]. Therefore, in order to characterize the tonal structure of Taiwanese, one has to consider not only the pitch curve but also other prosodic qualities such as duration. Moreover, the pronunciations of tones in Taiwanese were under the influence of tone sandhi, in which a citation tone changes to a sandhi tone according to its syntactic position. The details of this phenomenon were described in [4], but they are beyond the scope of this research.

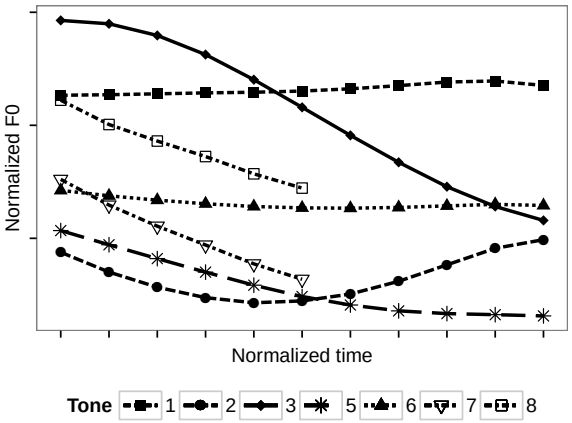


Figure 1: *Mean normalized F0 contour of seven Taiwanese tones.*

In the interest of comprehensive studies of the phonetics and phonology of Taiwanese language, a large speech corpus is required. However, a sizable Taiwanese speech database with accurate labeling is hard to come by. During the collection of such data by hand, one can find it to be time-consuming and error-prone. Thus, an automatic recognition system is crucial for building a large corpus for further studies.

Several researches have focused on different aspects of automatic processing of Taiwanese speech data. For example, in [5], a large vocabulary Taiwanese speech recognizer is built using HMM with raw pitch features in addition to a multiple pronunciation lexicon for sandhi tones. Specifically, two pitch smoothing techniques of the unvoiced regions, namely, random padding and exponential function linking between two consecutive pitch values, were compared to examine their abil-

ity to lower the character and utterance errors. Their results showed that using pitch information with exponential function for smoothing, in addition to the multiple pronunciation model, could significantly decrease the error rates in speech recognition. Another research focused on tone labeling of Taiwanese [6], in which both the citation and sandhi tones were jointly represented using statistical pitch contour models in order to eliminate contextual effects. It was proven to outperform the vector quantization method.

In this research, we want to focus on the recognition of Taiwanese tones using plain or curve-fitted features of the pitch contour along with durational feature. Following a similar approach in [7], we adopted the sub-sectioning method of splitting a pitch contour into different numbers of sections, and modeling them separately as our curve-fitted features. Also, the duration of a tone was included as an additional prosodic feature. We want to examine the effect of plain versus curve-fitted modeling of the F0 contour, as well as the effect of the number of sections and durational feature on tone recognition.

This paper is organized as follows. Section 2 describes our method of modeling and recognizing Taiwanese tones. Section 3 presents the experimental results along with some discussions. Finally, Section 4 concludes this paper.

2. Methods

We implemented several methods of extracting features from the F0 contour in order to compare the effect of them. First, raw F0 values were extracted by a Praat [8] script provided by [9]. Then, three types of feature sets, namely, plain, curve-fitted, and duration, were obtained from the raw F0 values to model different aspects of a tone. Plain and curve-fitted features were intended to capture the shape of the pitch curve, while the duration feature was included to describe the time-domain information. The following sections explain the definitions and extraction methods of these feature sets. Both the detailed and curve-fitted features were then paired with the duration features to train SVM classifiers and evaluation their performance on tone recognition.

2.1. Plain features

Plain features were simply the raw pitch values, in Hertz, computed from the audio. They precisely represent the original form of tones. There are 11 pitch values from equally-spaced points for each rhyme part of the syllable. The purpose of using these features is to provide fine-grained information of the pitch curve for tone model training.

2.2. Curve-fitted features

On the other hand, curve-fitted features were statistical pitch contour models used to capture the general characteristic of F0 variation within a tone. They can be further divided into two kinds. One is the method proposed by [10], in which a 3rd order orthogonal polynomial was used to represent the entire F0 contour. The basis polynomials, which are discrete Legendre polynomials, were normalized to the interval between 0 and 1. The detailed formulation is as expressed in (1).

$$
\begin{aligned}
\Phi_0(\frac{i}{N}) =& 1, \\
\Phi_1(\frac{i}{N}) =& \sqrt{\frac{12 \cdot N}{N+2}} \left(\frac{i}{N} - \frac{1}{2} \right), \\
\Phi_2(\frac{i}{N}) =& \sqrt{\frac{180 \cdot N^3}{(N-1)(N+2)(N+3)}} \cdot \\
& \left[\left(\frac{i}{N} \right)^2 - \frac{i}{N} + \frac{N-1}{6 \cdot N} \right], \\
\Phi_3(\frac{i}{N}) =& \sqrt{\frac{2800 \cdot N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}} \cdot \\
& \left[\left(\frac{i}{N} \right)^3 - \frac{3}{2} \left(\frac{i}{N} \right)^2 \right. \\
& + \frac{6 \cdot N^2 - 3 \cdot N + 2}{10 \cdot N^2} \left(\frac{i}{N} \right) \\
& \left. - \frac{(N-1)(N-2)}{20 \cdot N^2} \right]
\end{aligned}
\tag{1}
$$

for $0 \leq i \leq N$ where $N+1$ is the number of samples in the pitch contour. In this way, the original pitch values $\hat{f}\left(\frac{i}{N}\right)$ can be approximated by (2)

$$
\hat{f} = \left(\frac{i}{N} \right) \sum_{j=0}^{3} \alpha_j \cdot \Phi_j \left(\frac{i}{N} \right), 0 \leq i \leq N \tag{2}
$$

where

$$
\alpha_j = \frac{1}{N+1} \sum_{i=0}^{N} f \left(\frac{i}{N} \right) \cdot \Phi_j \frac{i}{N} \tag{3}
$$

Afterwards, the four coefficients $[\alpha_0, \alpha_1, \alpha_2, \alpha_3]$ from (3) were kept as one type of the curve-fitted features.

The other kind is a fractionalized fitting method similar to [7], in which the F0 curve was divided into four sections, and each section was represented by various parameters. Following this angle of approach, we used the 2nd order polynomial with the first three of the coefficients in (3). The F0 curve was first split into different numbers of equal-length sections, and each section was fitted separately. Figure 2 illustrates the difference between using a 3rd order polynomial to fit the entire curve, and fitting four sections separately using a 2nd order polynomial. We can see that a fractionalized fitting is more faithful to the original curve, while the higher order polynomial can capture the general shape of the curve. The resulting coefficients of the fitted functions were used as another set of curve-fitted features. The number of sections in a tone is another variable that we want to examine in this research.

2.3. Duration features

Lastly, the durations of each rhyme was used as another prosodic feature. As mentioned in Section 1, the duration information may be useful in distinguishing tones, especially the checked ones. We want to examine its effectiveness in distinguishing checked and non-checked tones, as well as other tones that were reported to have durational differences.

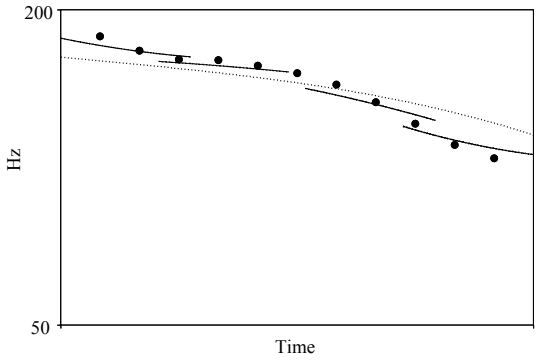


Figure 2: *Illustration of different curve-fitting methods on a series of samples of Tone 3. The dotted line represents using the whole series for fitting, and the solid lines represent fitting four sections separately.*

3. Experiments and Analysis

3.1. Experimental setup

We compiled a Taiwanese read speech corpus containing a total of 11,352 syllables from 10 different native speakers of Taiwanese. The recordings were conducted in a quiet room with a sampling rate of 44.1K and a background noise level lower than 30 dB. The wordlist consists of various categories of CVC combinations as described in Table 1, in which syllables with a stop coda were all carrying checked tones. The syllables represent real Taiwanese monosyllabic words, and were embedded in a carrier sentence to prevent influence from syntactic structures. Note that not all consonants and vowels in Taiwanese were included, as we are still in the process of building a comprehensive inventory now.

Table 1: *Inventory of our Taiwanese speech corpus.*

Position	Category	Inventory
Onset	Stop	p, t, k, ʔ, g
	Fricative	s
	Affricate	ts
	Liquid	l
	Nasal	m, n, ŋ
Vowel	Oral	i, e, a, ə, u, o, ɔ, ɤ
	Nasal	ã, ĩ
Coda	Stop	p, t, k, ʔ
	Nasal	m, n, ŋ

The total duration is about 11 hours. Since durational information may be used in training our tone models, the average duration is also reported in Table 2. The wave files were labeled by trained phoneticians using IPA symbols and numbers that denote surface tones. Pitch values from the rhyme part of the syllable were extracted using a Praat script "TimeNormalizedF0.praat" [9] from 11 equally spaced points. Then, different features described in Section 2 were obtained and used to train speaker-independent SVM classifiers using LibSVM [11], for comparing the effectiveness of our modeling methods.

Table 2: *Mean duration of seven Taiwanese tones in the corpus.*

Tone	Duration (ms)
1	202
2	233
3	183
5	180
6	209
7	122
8	107

3.2. Results and discussion

As shown in Table 3, the five-fold cross-validation accuracies of different feature sets were computed to evaluate the overall performance of our system. A few observations can be made from the results.

First, including the duration feature can indeed assist in identifying tones, with a 3% improvement in accuracy. It indicated that both frequency domain and time domain information are essential in the modeling of tones. Secondly, using curve-fitted features of the F0 curve outperforms using the plain features of the raw F0 values, which conforms to previous researches of tone modeling. The technique of dividing the curve into a number of sections was proven effective as well. By just splitting the curve into two sections, we can achieve a 0.5% increase of accuracy.

Table 3: *A comparison of tone recognition accuracy between different feature sets.*

Type	Feature	Accuracy (%)
Plain Features	Raw F0	77.10
	Raw F0 + duration	80.02
Curve-fitted Features	Entire curve fit + duration	80.08
	2-section fit + duration	80.51
	3-section fit + duration	80.32
	4-section fit + duration	**80.75**
	5-section fit + duration	80.60

The best performance was found in the feature set of 4-section polynomial fitting plus duration, with the accuracy of 80.75%. Notably, a higher partitioning of the F0 contour, i.e. 5-sections, resulted in a lower accuracy. It showed that the number of features is not positively related to accuracy, as the separation of a tone contour into too many sections might cause an overfitting effect that compromised the robustness of a model and its ability to identify the general characteristics of a tone. A 4-section method may be appropriate in that it can capture the left and right contextual variations resulting from neighboring segments using the two boundary sections, while the fluctuations in the center regions of a tone were well-represented by the remaining two sections. It has also been proven successful in [7], in which the language being studied is Mandarin. On the other hand, fitting the whole curve with a higher-order polynomial function is too coarse to be effective in representing a tone, and thus resulting in a lower accuracy.

In order to further analyze the effectiveness of our model on each of the seven tones, we trained the model using the complete set of F0 and duration data with the 4-section plus duration feature set, and then computed the accuracy for each tone. The results were shown in Table 4. As we can see, Tone 7 and 8

have considerably lower accuracies than others. It may simply be due to the fact that the sample size is too small for generating a robust model. However, there are in fact two kinds of checked tones in Taiwanese, one ending with /p, t, k/ and the other with /ʔ/. Previous studies showed that the realizations of them are slightly different [4]. For a deeper understanding and modeling of these two checked tones, additional data as well as research on the modeling techniques are required.

Table 4: *Recognition accuracy grouped by tone types.*

Tone	Number of correct/total syllables	Accuracy (%)
1	2204/2525	87.29
2	805/939	85.73
3	2033/2193	**92.70**
5	1737/2117	82.05
6	2200/2645	83.18
7	441/588	75.00
8	154/344	44.77

Nonetheless, for the non-checked Tone 1 to 6, the best performance was found on Tone 3. It could be attributed to the unique shape and range of the F0 curve, as depicted in Figure 1, along with a shorter duration that gave rise to a more distinctive tone model. Contrastively, the relatively lower accuracy of less than 85% occurred in Tone 5 and 6. It could be accounted for if we look at the wrong predictions of the classifier, as explained separately below.

- For Tone 5, we found that the most common mistakes were Tone 3 and 6, each occurred about 140 times. The similar pitch height of Tone 5 and 6 might have caused a confusion for the recognizer. As for the other two tones with comparable pitch height, namely, Tone 7 and 8, they can be easily distinguished by duration. Meanwhile, Tone 5 and 3 may have been indivisible because they were alike in both shape and durational feature.
- For Tone 6, the most common errors were Tone 1 and 5, with around 260 and 140 occurrences, respectively. The analogous reasoning above could be applied to explain the indistinguishability between Tone 6 and 1, in that their shape, height, and duration were all comparable. The multiple resemblances between them could have contributed to the errors.

Another perspective on the lower accuracy group is that there may be mutual affinity among them. In fact, previous study [12] showed that, in some dialects of Southern Min, Tone 5 and 6 were merged or 'neutralized'. Our findings could lend support to the theory that tonal neutralization is the result of the similarity and difficulty in maintaining contrast between tones.

In sum, a more robust model with the capability to tackle with these problems is required to improve the accuracy of our system. Moreover, a full-fledged Taiwanese tone recognition system must be able to distinguish sandhi tones from surface tones, and our system is yet to achieve this goal. A more sophisticated modeling scheme is necessary to incorporate such complications.

4. Conclusions

In this paper, we examined the effect of fractionalizing prosodic features on Taiwanese tone recognition, and proposed a new approach to modeling Taiwanese tones. Our results showed that using curve-fitting of four fractions of F0 values, and including the duration feature into model training were useful means of improving the accuracy of tone recognition. By further analyzing the outcomes, we provided an empirical point of view for theoretic studies on tonal neutralization. Future work can be done on investigating the effect of other prosodic and articulatory features, such as consonant context or energy. Since previous research suggests that these factors may play a role in tone recognition [1], incorporating them into Taiwanese speech recognition systems could be fruitful. In addition, expanding our corpus to include more tokens of checked tones is necessary for improving the accuracy. Furthermore, the current system only dealt with the surface tones. We will have to derive a more extensive model to resolve the problem of recognizing sandhi tones in Taiwanese.

5. Acknowledgments

The Taiwanese recording corpus was provided by the National Tsing Hua University Phonetics Lab, and was supported in part by the National Science Council grant of Taiwan (NSC 97-2410-H-007-025 and NSC 99-2410-H-007-050). The authors would like to thank Dr. Yu Tsao for insightful remarks and assistance, as well as comments from two anonymous reviewers.

6. References

[1] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4741–4744.

[2] L. W. R. Cheng and S. J. Cheng Xie, *Phonological Structure and Romanization of Taiwanese Hokkien*. Taipei: Student Book Company, 1977.

[3] S.-F. Wang and J. Fon, "Durational cues at discourse boundaries in Taiwan Southern Min," in *Proc. 6th International Conference on Speech Prosody*, 2012.

[4] R. L. Cheng, "Tone sandhi in Taiwanese," *Linguistics*, vol. 6, no. 41, pp. 19–42, 1968.

[5] D.-C. Lyu, M.-S. Liang, Y.-C. Chiang, C.-N. Hsu, and R.-Y. Lyu, "Large vocabulary Taiwanese (Min-nan) speech recognition using tone features and statistical pronunciation modeling," in *Proc. 8th EuroSpeech*, 2003.

[6] W.-C. Kuo, Y.-R. Wang, and S.-H. Chen, "A model-based tone labeling method for Min-nan/Taiwanese speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 505–8.

[7] Y. Tian, J. L. Zhou, M. Chu, and E. Chang, "Tone recognition with fractionized models and outlined features," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2004.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.3.55)," 2013.

[9] Y. Xu, "Timenormalizef0.praat," 2009.

[10] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. On Communications*, vol. 38, no. 9, pp. 1317–1320, 1990.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[12] C.-T. Chuang, Y.-C. Chang, and F.-F. Hsieh, "Complete and not-so-complete tonal neutralization in Penang Hokkien," in *Proc. International Conference on Phonetics of the Languages in China*, W.-S. Lee, Ed., 2013, pp. 54–57.

Comparison of Pitch Range and Pitch Variation in Slavic and Germanic Languages

Bistra Andreeva[1], *Grażyna Demenko*[2], *Magdalena Wolska*[3], *Bernd Möbius*[1], *Frank Zimmerer*[1], *Jeanin Jügler*[1], *Magdalena Oleskowicz-Popiel*[2], *Jürgen Trouvain*[1]

[1] Computational Linguistics & Phonetics, Saarland University, Germany
[2] Department of Linguistics, Adam Mickiewicz University, Poland
[3] LEAD, Eberhard Karls University Tübingen, Germany

[andreeva, moebius, zimmerer, juegler, trouvain]@coli.uni-saarland.de,
lin@amu.edu.pl, magdalena.jastrzebska@speechlabs.pl, magdalena.wolska@uni-tuebingen.de

Abstract

This study presents the results of a large-scale comparison of various measures of pitch range and pitch variation in two Slavic (Bulgarian and Polish) and two Germanic (German and British English) languages. The productions of twenty-two speakers per language (eleven male and eleven female) in two different tasks (read passages and number sets) are compared. Significant differences between the language groups are found: German and English speakers use lower pitch maxima, narrower pitch span, and generally less variable pitch than Bulgarian and Polish speakers. These findings support the hypothesis that linguistic communities tend to be characterized by particular pitch profiles.

Index Terms: pitch range, pitch variation, cross-language differences, Bulgarian, Polish, German, British English

1. Introduction

Several studies over the past decades have shown that linguistic communities (different social groups within a single language or speakers of different languages) tend to be characterized by particular pitch profiles (pitch range and pitch variation, see [7] for a review). Luchsinger and Arnold [14] found that Puerto Rican girls in New York City and native American women use fundamental frequency (*f0)* differently. While Puerto Rican girls tend to speak on a rather high pitch, many American women prefer to speak on a low pitch level. Dialects of a language can also differ with respect to the use of *f0* (e.g. [6, 29]). Various cross-linguistic studies also indicate language specific differences with respect to *f0*. Comparing typologically different languages (English, Spanish, Japanese, Tagalog), Hanley et al. [9] and Hanley and Snidecor [10] found that the fundamental frequency of English males had the lowest median *f0*. Later studies compared Polish vs. English [17], Mandarin vs. English [4, 11], British English vs. German [18, 19], or Russian vs. German [20]. Some studies showed that bilingual speakers differ when speaking their two languages. For example, bilingual English/Japanese speakers used a higher pitch in Japanese than in English [8, 28, 30]. These findings demonstrate that such differences need not be due to physiological differences between speakers of different languages.

Ohala and Gilbert's [21] report on experiments in which listeners can identify their own language (Japanese, Cantonese and English) based solely on prosodic cues (*f0*, amplitude and timing characteristics). It has further been found that some languages are discriminable purely by their fundamental frequency ([23] for English and Japanese, [15] and [16] for English and French and [5] for English and Dutch).

Language specific components have also been found to be important in the perception and production of paralinguistic aspects ([13] for politeness in Japanese and English, [3] for 'confident', 'friendly', 'emphatic' and 'surprised' in British English and Dutch).

However, it is difficult to compare the data reported in these publications, because most studies have been limited to either male or female (mostly small numbers of) speakers, the analyses were based on different discourse types, or the methods for *f0* estimation were different.

The aim of this study is to lay a foundation for a large-scale quantitative analysis of the fundamental frequency (level and span) of speakers of two typologically different language groups (Slavic: Bulgarian and Polish, and Germanic: German and (British) English). The analysis presented here is based on the assumption that pitch range and pitch variation in linguistically homogeneous communities will cluster within each community, but might differ across communities.

2. Material and Methods

Two Slavic (Bulgarian and Polish) and two Germanic (German and English) languages are in the focus of this study. The material analyzed is continuous read speech taken from two comparable multi-lingual speech databases, EUROM-1 (for German and English) [2] and BABEL (for Bulgarian and Polish) [25, 26]. We used a subset of the data, consisting of 3 blocks of 20 numbers (from 0 to 9999) and 3 cognitively linked short passages, containing 5 thematically connected sentences, read by 22 speakers (11 male and 11 female) per language. The passages were based on identical, real-life topics for the different languages, freely translated and adapted for Bulgarian, German and Polish from the original English texts. The overall length of the analyzed material per language is about 60 minutes.

3. f_0 Measures

Pitch values were collected at 0.01 seconds time steps for the male and 0.005 seconds time steps for the female speakers using the RAPT algorithm [27] implemented in the program 'get_f0' from the ESPS software package. The automatically extracted *f0* values were verified and manually corrected, if necessary. Irregular voiced stretches of speech due to laryngealization were excluded from further analyses.

According to Ladd [12], *f0* values can be attributed to two partially related but distinct characteristics of a speaker's performance: (a) pitch level, i.e. the overall height of the speaker's voice, and (b) pitch span, i.e. the range of frequencies covered by the speaker. To analyze the cross-language differences in pitch range and variation, the following distributional measures were calculated: mean and median *f0* values for level and interquartile range (IQR) and the simple pitch excursion for span, whereas the latter was simply computed as the difference between maximum and minimum pitch values over a passage or number block. The obtained Hertz measurements for span were additionally converted to semitones by means of the formula [24]:

39.863 * log10(Maximum/Minimum).

The measures describing the variation and shape of the *f0* distribution were standard deviation (SD), kurtosis and skewness (in Hz).

4. Results

Means and standard deviations for each of the distributional measures for pitch level and span are presented in Table 1, organized by language, speaker sex, and task.

As a first step towards determining the differences, linear mixed models with the respective measure as dependent variable, speaker as random factor, and native language (Bulgarian/Polish/English/German), gender (male/female) and task (passage/number set) as fixed factors, as well as all their possible interactions, were computed for each dependent variable in separate analyses. Separate Tukey post-hoc tests were carried out per variable, if appropriate. The confidence level was set at α=0.05.

4.1. Passages

Predictably, gender had a significant main effect on mean (F [1, 80] = 520.32, p<0.001) and median *f0* (F [1, 80] = 480.50, p<0.001), IQR (F [1, 80] = 70.47, p<0.001), minimum *f0* (F [1, 80] = 266.57, p<0.001), maximum *f0* (F [1, 80] = 341.84, p<0.001) and SD (F [1, 80] = 94.69, p<0.001), with females having significantly higher *f0* values (cf. Figure 1 for level measures). Gender did not differ in skewness, kurtosis and *f0* span measured in semitones.

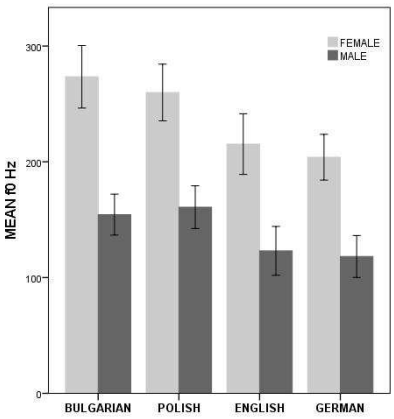


Figure 1: *Mean f0 pooled over all male and female speakers and all tasks.*

Table 1. Means and standard deviations for the distributional measures by language, speaker sex and task. The values for each measure are given in Hz except for the second span measure which is in semitones.

	measure	Passages		Numbers	
		M	*F*	*M*	*F*
BG	mean	154 (17)	275 (26)	155 (19)	272 (28)
	median	154 (17)	273 (26)	159 (20)	278 (29)
	IQR	43 (13)	72 (17)	28 (9)	55 (16)
	minimum	81 (11)	152 (24)	83 (17)	165 (24)
	maximum	231 (34)	428 (42)	195 (23)	352 (40)
	span	151 (33)	276 (42)	112 (21)	187 (30)
	span (ST)	18.3 (3.1)	18.0 (2.7)	14.9 (3.0)	13.1 (1.8)
	SD	29 (8)	52 (9)	22 (5)	40 (8)
	skewness	-.02 (.39)	.18 (.24)	-.95 (.54)	-.50 (.38)
	kurtosis	-.27 (.71)	-.25 (.39)	.98 (1.5)	-.19 (.78)
PL	mean	157 (18)	259 (21)	165 (19)	260 (28)
	median	156 (19)	254 (21)	166 (20)	259 (31)
	IQR	38 (11)	62 (15)	51 (18)	73 (18)
	minimum	82 (8)	146 (23)	92 (21)	165 (16)
	maximum	246 (34)	437 (56)	231 (27)	382 (50)
	span	165 (33)	291 (59)	139 (30)	217 (43)
	span (ST)	19.0 (2.7)	19.1 (3.6)	16.2 (4.0)	14.5 (2.2)
	SD	30 (8)	50 (9)	32 (10)	46 (9)
	skewness	.10 (.51)	.49 (.43)	-.18 (.41)	.10 (.35)
	kurtosis	.28 (.57)	.52 (1.0)	-.77 (.40)	-.76 (.46)
DE	mean	120 (18)	206 (21)	116 (18)	202 (19)
	median	119 (19)	204 (22)	117 (18)	204 (18)
	IQR	24 (8)	44 (13)	15 (5)	28 (6)
	minimum	82 (14)	137 (30)	85 (15)	144 (24)
	maximum	181 (33)	298 (30)	149 (22)	264 (25)
	span	100 (28)	161 (39)	64 (11)	120 (24)
	span (ST)	13.8 (2.9)	13.9 (4.5)	9.7 (1.4)	10.6 (2.6)
	SD	18 (6)	30 (8)	11 (3)	20 (4)
	skewness	.44 (.49)	.26 (.29)	-.23 (.51)	-.04 (.48)
	kurtosis	.35 (1.1)	-.29 (.41)	.17 (.75)	.15 (1.4)
EN	mean	127 (23)	218 (23)	119 (19)	213 (29)
	median	125 (22)	214 (25)	119 (20)	212 (31)
	IQR	30 (13)	41 (12)	27 (15)	32 (14)
	minimum	84 (14)	155 (22)	79 (10)	160 (31)
	maximum	205 (52)	330 (44)	180 (45)	294 (42)
	span	121 (44)	175 (47)	101 (43)	134 (36)
	span (ST)	15.1 (3.2)	13.1 (3.4)	14.0 (3.9)	10.7 (3.1)
	SD	23 (9)	32 (8)	19 (10)	23 (8)
	skewness	.67 (.36)	.72 (.47)	.41 (.46)	.47 (.48)
	kurtosis	.50 (.91)	.73 (1.1)	.26 (1.7)	.23 (.90)

However, over and above the expected gender effect, there was also a significant main effect of language on all measurements except on minimum *f0*, where the speakers are near the floor of their physiological *f0* range. Separate post-hoc tests showed that Bulgarian and Polish speakers had a significantly higher mean *f0* (F [3, 80] = 33.07, p<0.001), median *f0* (F [3, 80] = 32.60, p<0.001), IQR (F [3, 80] =

21.06, p<0.001), maximum *f0* (F [3, 80] = 33.29, p<0.001), *f0* span in semitones (F [3, 80] = 17.05, p<0.001) and SD (F [3, 80] = 26.96, p<0.001) than English and German speakers. We found a positively skewed *f0* distribution for the four languages. This implies that the most frequent *f0* observation occurs lower than the mean. The skewness values for English speakers were significantly higher than those for German, Polish and Bulgarian speakers (F [3, 80] = 11.51, p<0.001). English speakers also had a higher kurtosis than German and Bulgarian speakers, and Polish speakers had a higher kurtosis than Bulgarian speakers (F [3, 80] = 8.33, p<0.001). This reflects the fact that *f0* in Bulgarian and German is distributed over a narrower area (cf. Tables 1 and 2).

The statistical analysis revealed a significant interaction between language and gender for mean *f0* (F [3, 80] = 3.10, p<0.05), maximum *f0* (F [3, 80] = 6.15, p<0.001) and SD (F [3, 80] = 3.76, p<0.05). In the passages, the speakers of the Slavic group used higher average *f0* and higher maximum values and showed a larger SD (possibly indicating more liveliness) than the speakers in the Germanic group. The only exceptions to this finding are the English male speakers with respect to their maximum *f0* values and SD (cf. Figure 2a). Thus, the English male speakers used the same maximum *f0* values and had the same SD as the German male speakers and the male speakers from the Slavic group. Figures 2a and 2b display the mean and maximum pitch values as well as SD for male and female speakers in the four languages.

4.2. Number Blocks

In these analyses, as expected, we found again that women had a significantly higher mean *f0* (F [1, 80] = 424.13, p<0.001), median *f0* (F [1, 80] = 379.82, p<0.001), IQR (F [1, 80] = 43.09, p<0.001), maximum *f0* (F [1, 80] = 339.89, p<0.001), minimum *f0* (F [1, 80] = 315.98, p<0.001) and SD (F [1, 80] = 51.78, p<0.001) . In contrast to the findings for the passages, in the number task male speakers used a larger frequency range in semitones (F [1, 80] = 6.97, p<0.01) than females. This result may be partially attributed to the fact that speakers tend to use quite idiosyncratic intonation patterns for the number blocks: some speakers prefer continuation rises to separate the blocks, other speakers tend to use falling intonation to end a block. Gender was also significant for skewness (F [1, 80] = 8.32, p<0.05).

Table 2. *Language-group differences for the f0 measures on the basis of Tukey post-hoc comparisons.*

f0 measure	significant language-group differences	
	passages	number blocks
mean *f0*	BG = PL > EN = DE	BG = PL > EN = DE
median *f0*	BG = PL > EN = DE	BG = PL > EN = DE
min *f0*	N.S.	N.S.
IQR	BG = PL > EN = DE	PL > BG > EN = DE
max *f0*	PL = BG > EN = DE	PL > BG > EN > DE
span ST	PL = BG > EN = DE	PL = BG > BG = EN > DE
SD	BG = PL > EN = DE	PL > BG > EN = DE
skewness	EN > DE = PL = BG	EN > PL = DE > BG
kurtosis	EN = PL > PL = DE > DE = BG	BG = EN = DE > PL

Again, a significant main effect of language was found in all measures except *f0* minimum: mean *f0* (F [3, 80] = 37.75, p<0.001), median *f0* (F [3, 80] = 36.73, p<0.001), IQR (F [3, 80] = 44.68, p<0.001), maximum *f0* (F [3, 80] = 35.52, p<0.001), *f0* span in semitones (F [3, 80] = 16.27, p<0.001), SD (F [3, 80] = 42.57, p<0.001), skewness (F [3, 80] = 31.87, p<0.001) and kurtosis (F [3, 80] = 9.82, p<0.001). However, the post-hoc tests yielded different language groupings (cf. Table 2).

Significant interactions between gender and language were found for IQR (F [3, 80] = 3.30, p<0.05), SD (F [3, 80] = 3.71, p<0.05) and kurtosis (F [3, 80] = 2.99, p<0.05).

4.3. Level vs. Span

The scatter plots in Figures 3 and 4 provide a visual representation of *f0* span (in Hz) and level (mean *f0* Hz) for all speakers of the four languages. The figures show that some speakers have a wide span but differ in level or vice versa, i.e. some speakers have a similar level but differ in span. The English and German speakers cluster in the lower left corner of the level/span plane, while the Bulgarian and Polish speakers cluster mostly in the higher right sector, which indicate that Slavic speakers may are more expressive than Germanic speakers in terms of span and level. The scatter plots also illustrate the different strategies of some speakers with respect to task type.

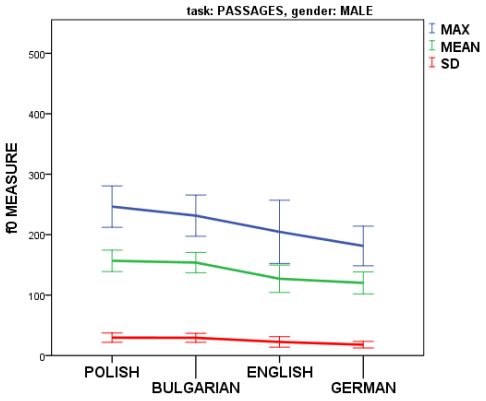


Figure 2a: *Mean and maximum f0 values and SD for male Bulgarian, Polish, English and German speakers.*

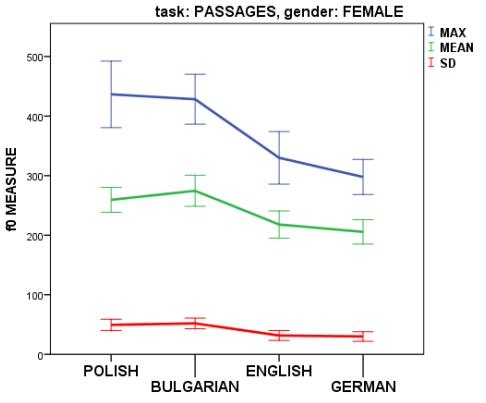


Figure 2b: *Mean and maximum f0 values and SD* for *female Bulgarian, Polish, English and German speakers.*

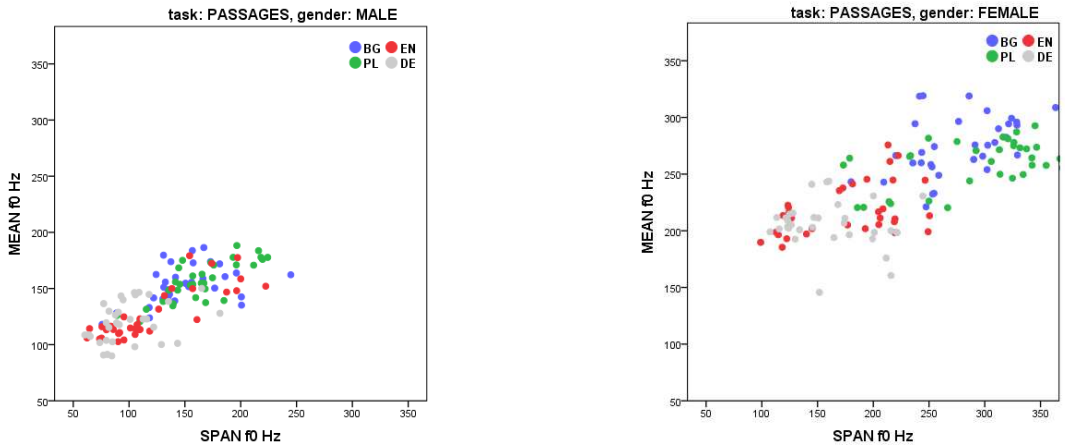


Figure 3: *Scatter plot showing span and level from the passages for male (left panel) and female (right panel) speakers.*

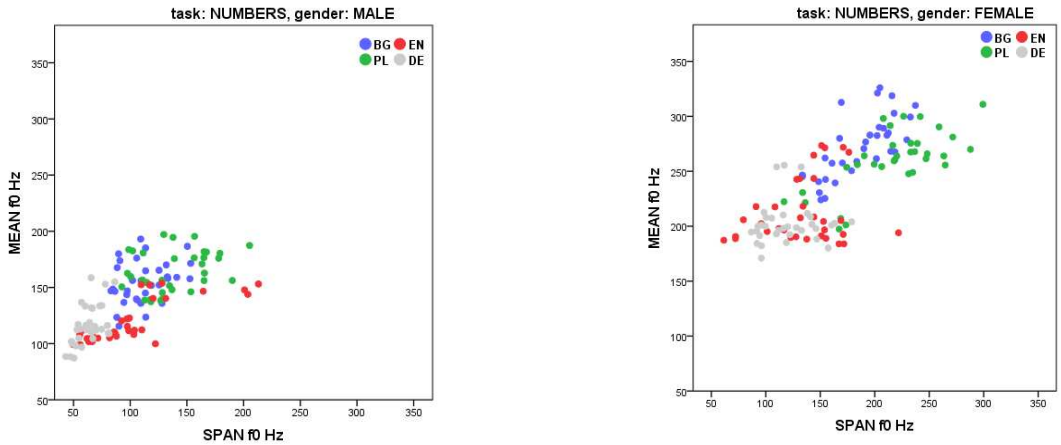


Figure 4: *Scatter plot showing span and level from the number blocks for male (left panel) and female (right panel) speakers.*

5. Discussion and Conclusions

This paper contributes to the growing number of studies on cross-language differences in pitch range and pitch variation. Our results confirm the hypothesis that linguistic communities tend to be characterized by particular pitch profiles. German and English speakers use a considerably lower level, narrower span, and generally less variable pitch than Bulgarian and Polish speakers. Gender also plays a significant role in *f0* variation. In the present study a distinctive frequency region as well as different mean frequencies were found for male and female speakers (about 139 Hz and 238 Hz respectively). However, the differences in mean frequencies were on the average less than one octave: 10.14/9.72 semitones for Bulgarian, 8.63/7.86 semitones for Polish, 9.32/9.64 semitones for German and 9.51/10.11 semitones for English (first number for passages, second one for number blocks).

Systematic differences between tasks were observed which appear to be attributable to differing strategies that speakers employ when reading short stories versus lists of numbers. Inter-speaker variability was considerably greater for the number lists. The syntactic-semantic structure of the story seems to constrain the speakers' prosodic options.

Our results do not corroborate the results reported by Mennen and colleagues [18, 19]; the female English and German speakers do not differ with respect to level and span. But distributional measures may in fact not be able to capture significant cross-linguistic differences (cf. [1, 19, 22]). In future work we expect to refine our measures of pitch range, by including linguistically based measures which were found to be better predictors of differences in pitch range and pitch variation across speakers and languages, and also by adding data from more speakers, including Bulgarian and Polish L2 speakers of English, more languages, as well as spontaneous speech data.

6. Acknowledgements

This research was partially supported by Research Grant UMO-2012/04/M/HS2/00551 from the NCN (Polish National Research Center).

We would like to thank Ryszard Gubrynowicz (Speech Acoustics Laboratory, Institute of Fundamental Technology Research, Polish Academy of Science) and Snezhina Dimitrova (English Department, Sofia University “St. Kliment Ohridski”) for kindly providing the Babel databases for Polish and Bulgarian, respectively.

7. References

[1] Campione, E. , and Véronis, J. (1998). A statistical study of pitch target points in five languages. *Proceedings of ICSLP'98*, 1391-1394.

[2] Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; Veld, C. and Zeiliger, J. (1995). Eurom - a spoken language resource for the EU. In *Eurospeech' 95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology*, 1, Madrid., 18-21 September 1995, 867–870.

[3] Chen, A, Gussenhoven, C., and Rietveld, T. (2004). Language-specificity in the perception of paralinguistic intonational meaning, *Language & Speech* 47, 311–349.

[4] Chen, G. T. (1972). *A comparative study of pitch range of native speakers of Midwestern English and Mandarin Chinese: An acoustic study*, doctoral dissertation, University of Wisconsin-Madison, Madison.

[5] de Pijper, J. R. (1983). *Modelling British English intonation*, Dordrecht - Holland: Foris.

[6] Deutsch, D., Le, J., Shen, J., and Henthorn, T. (2009). The pitch levels of female speech in two Chinese villages. *Journal of the Acoustical Society of America*, April, 125, EL208.

[7] Dolson, M. (1994). The pitch of speech as a function of linguistic community, *Music Perception* 11 (3), 321–331.

[8] Graham, C., (2013). Revisiting f0 Range Production in Japanese-English Simultaneous Bilinguals. *Annual Report of UC Berkeley Phonology Lab*, 110–125.

[9] Hanley, T.D., Snidecor, J.C., and Ringel, R. (1966). Some acoustic differences among languages, *Phonetica* 14, 97–107.

[10] Hanley, .D. and Snidecor, J.C. (1967). Some acoustic similarities among languages, *Phonetica* 17, 141–148.

[11] Keating, P. & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin, *Journal of the Acoustical Society of America* 132, 1050–1060.

[12] Ladd, D.R. (1996). *Intonational Phonology*. Cambridge: Cambridge University Press.

[13] Loveday, L. (1981). Pitch, politeness and sexual role: an explanatory investigation into the pitch correlates of English and Japanese formula,. *Language and Speech* 24, 71–89.

[14] Luchsinger, R. and Arnold, G. (1965). *Voice-Speech-Language*. Constable&Co Ltd., London.

[15] Maidment, J. A. (1976). Voice fundamental frequency characteristics as language differentiators. *Speech and hearing: Work in progress*, University College London, 74–93.

[16] Maidment, J. A. (1983). Language recognition and prosody: further evidence, *Speech, hearing and language: Work in progress*, University College London 1, 133–141.

[17] Majewski, W., Hollien, H., and Zalewski, J. (1972). Speaking fundamental frequency of Polish adult males, *Phonetica* 25, 119–125.

[18] Mennen, I., Schaeffler, F., & Docherty, G. (2007). Pitching it differently: A comparison of the pitch ranges of German and English speakers. *16th International Congress of Phonetic Sciences* (ICPhS XVI), Saarbrücken,1769-1972.

[19] Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: a comparison of English and German *Journal of the Acoustical Society of America* 131(3), 2249–2260.

[20] Nebert, Augustin Ulrich (2013). *Der Tonhöhenumfang der deutschen und russischen Sprechstimme. Vergleichende Untersuchung zur Sprechstimmlage.* Hallesche Schriften zur Sprechwissenschaft und Phonetik, Band 46. Frankfurt/M.

[21] Ohala, J. J., and Gilbert, J. B. (1979). Listeners' ability to identify languages by their prosody, in P. Léon and M. Rossi (eds.) *Problèmes de Prosodie*, Didier, Ottawa, 123–131.

[22] Patterson, D. (2000). *A Linguistic Approach to Pitch Range Modelling*. Ph.D. dissertation, University of Edinburgh.

[23] Ramus, F., and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America* 105 (1), 512–521.

[24] Henning Reetz (1999): *Artikulatorische und akustische Phonetik*. Wissenschaftlicher Verlag, Trier.

[25] Roach, P.; Arnfield, S. and Hallum, E., (1996). BABEL: A multi-language speech database. In *Proceedings of SST-96: Speech and Science Technology Conference*, Adelaide, 351–4.

[26] Roach, P., Arnfield, S., Barry, W.J., Dimitrova, S., Boldea, M., Fourcin, A., Gonet, W., Gubrynowicz, R., Hallum, E., Lamel, L., Marasek, K., Marchal, A., Meister, E., Vicsi, K. (1998). Babel: a database of Central and Eastern European languages, *Proceedings of the First International Conference on Language Resources and Evaluation*, vol. 1, 28-30 May 1998, Granada, Spain, pp. 371–374.

[27] Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT). In Kleijn, W. B. and Paliwal, K. K. (eds.), *Speech Coding and Synthesis*. New York: Elsevier.

[28] Todaka, Y. (1993). *A cross-language study of voice quality: bilingual Japanese and American speakers*, doctoral dissertation, University of California, Los Angeles, pp. 145–147.

[29] Torgerson, R. C. (2005*). A comparison of Beijing and Taiwan Mandarin tone register: An acoustic analysis of three native speech styles*, master's thesis, Brigham Young University, 73–82.

[30] Yamazawa, H., and Hollien, H. (1992). Speaking fundamental frequency pattern of Japanese women, *Phonetica* 49, 128–140.

Silent reading and prosodic structure constraints

Philippe Martin

UMR 7110, LLF, UFRL, Université Paris Diderot, ODG, Place Paul Ricœur, 75013 Paris, France
philippe.martin@linguist.univ-paris-diderot.fr

Abstract

Silent reading of written texts involves normally a process of subvocalization, i.e. the presence of a voice reading the text in the head of the reader speaking to her/himself. This process includes not only the sequences of syllables corresponding to the written material, but also sentence intonation. Since subvocalization cannot be eliminated other than by changing the status of each word into a pictographic function (as it may be the case for a STOP road panel sign for example), it is argued here that sentence intonation is essential to language comprehension, and more specifically to the conversion of sequences of syllables into higher order linguistic units (corresponding to accent phrases AP in the Autosegmental-Metrical model).

Consequently, reading and in particular silent reading is constrained by the same rules than the prosodic structure in general, and specifically to the minimal duration of accent phrases. This minimal value, occurring when AP's contain only one syllable, is about 250 ms, a value which corresponds to the minimal period value of Delta brain waves [4], [11]. Therefore, this AP minimal duration limits also the maximal number of AP that could be processed in silent reading, i.e. about 240 per minute, which corresponds to the maximal number of words per minutes experts in fast reading can process while keeping a reasonable level of comprehension, i.e. about 800 wpm.

Index Terms: silent reading, prosodic structure, subvocalization, Delta waves, Theta waves.

1. Introduction

When we read, either silently or aloud, we generate speech sounds according to the reduced information given in the written text. In this process, we also generate a prosodic structure, which hierarchically organizes accent phrases AP (minimal units of intonation containing a single lexical or group stress), into prosodic groups, called in the Autosegmental-Metrical model ip (intermediate intonative phrases) and at a higher level IP (Intonation Phrases), whose sequences in turn constitute the whole utterance intonation.

It is noticeable that this prosodic structure (re)generation is essential to help the reader to understand the text. Therefore, the whole reading process is constrained by the rules governing the elaboration of the sentence prosodic structure when speaking either silently or aloud, and in particular the minimal and maximal duration of accent phrases [11].

In silent reading, it is difficult to proceed without subvocalization, i.e. without hearing a voice in one's head that corresponds to a voice reading the text aloud. For this reason, silent reading may be subject to similar constraints than reading aloud (let aside articulatory constraints). These constraints may interact or even supersede constraints established for eye movement while reading. In particular, they may lead to a new explanation pertaining to the maximum number of words that can be processed in fast reading.

2. Eye movement

When reading, the eye proceeds in saccades (short rapid movements) to scan the text, jumping in steps varying from 1 to 20 characters with an average of 7 to 9 characters (forward and backward). In the process, the most frequent fixations are given by verbal forms and punctuation marks (dots, commas, semicolons, question marks, etc.). The eye jumps then constantly to spot these markers, which will constitute the bases for the prosodic structure to build [10].

Most of the laboratory speech research on sentence intonation actually investigate this process thoroughly on read speech, before considering spontaneous, non-prepared speech prosodic features. For example, if a dot normally ending written text sentences is associated with a falling conclusive prosodic marker, the correspondence of the other punctuation marks and the verbal forms must be dynamically associated with a proper prosodic contour (a Tone Boundary in the Autosegmental-Metrical model).

The saccades allows to eye to focus on fixation point in 20 ms to 40 ms, whereas the fixation point last between 100 ms and 500 ms [17]. The fixation state of the eye allows the fovea, the central part of the retina, to scan the selected written information with high resolution, whereas peripheral information is viewed with fewer details (Fig. 1).

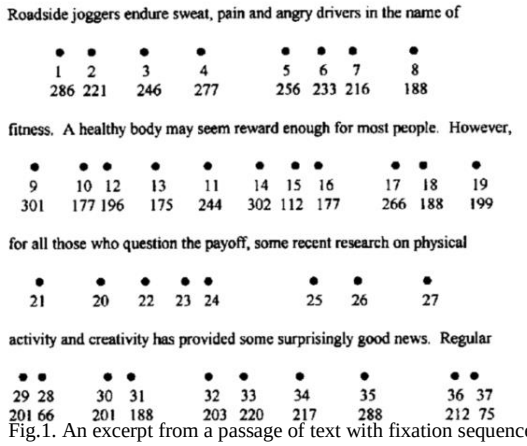


Fig.1. An excerpt from a passage of text with fixation sequence and fixation durations. The dots below the words indicate the fixation location, the first number below a dot its rank in the sequence, and the second number below a dot its duration of fixation (from [17]).

Due to the complex muscular mechanisms for speech generation, oral (i.e. aloud) reading is slower than silent reading. However, the puzzling aspect of silent reading lies in its limitations. Despite a number of questionable commercial claims stating that fast readers could read up to 3000 words per minute (about 50 words/sec…), the fast reading process is limited by subvocalization, the effect to hear a voice in one's head while reading silently (which was curiously attributed by some to the way we learn to read at school [13]).

3. Subvocalization

Subvocalization does not pertain to the mechanical control of articulators muscle control, but to the perception of the speech signal, which is recovered by reading. The invention of writing has precisely this function, allowing not only reading aloud but also silently, i.e. "to talk to oneself in silence". Indeed, writing is a shorthand notation system of speech sound and not of articulatory movements, contrary to what claim supporters of the motor theory of speech perception [8]. There is apparently no writing system (even API) referring directly to articulatory configurations.

Other systems such as pictograms bypass the generation of speech sounds by associating directly significant and signifier to access their signification without going through language units be syllables, words, prosodic words, syntagms, etc. A road STOP sign may indeed be read aloud or silently, but is more frequently directly associated with its meaning, i.e. to give way on the road.

Likewise, dates written with numbers, e.g. 1789, may be read as "seventeen hundred eighty nine", but the constant use of symbols not corresponding directly to syllables and words leads more frequently to a direct access to its signifier. The passage to the status of pictogram depends of course of the familiarity of the reader with the object and its frequency of occurrence.

Writing systems using ideograms, for example Mandarin, also involve subvocalization in silent reading. Learning Mandarin without being concerned by ideograms pronunciation would be difficult, as many words are plurisyllabic, implying for such reader to deal with combination of pictograms [9]. However, one could associate other sounds to ideograms, such as English words for example, but the mediation of some speech sound seems difficult to avoid, although not impossible in principle.

Commercial US based fast reading "schools" claim that they can remove subvocalization, or at least minimize it. The subliminal idea is to transform every word into a pictogram, so when read it will not be pronounced silently. Other techniques recommend to use a pencil to determine eye fixation targets and accelerate the number of saccades. An application even proposes to display only lexical words sequentially on a computer screen with a user adjustable speed (this approach incidentally corresponds to the definition of accent phrases in the autosegmental-metrical model, i.e. one lexical word for each accent phrase). Comprehension should then be achieved without any prosodic structure and no syntactic structure linking the read words together by simple concatenation and no hierarchical structure.

Faster readers claim speed from 400 wpm (words per minutes) to 800 wpm. With an average number of about 3 (written) words per accent phrase, 800 wpm convert into about 266 AP's per minute, or 266/60 = 4.4 AP's per second. So the minimal average duration between silently read AP's would be about 225 ms. 800 wpm for the best observed performance for speed readers [2].

4. Accent phrases

Recent studies on spontaneous speech show that we speak and read by accent phrases and not word by word [1]. Accent phrases (AP's, aka prosodic word, stress groups, temporal groups, etc.) are minimal units of prosody contain only one (lexical or group) stress. It is also claimed that accent phrases necessarily contain either a verb, a noun, an adjective or an adverb together with grammatical words, but the analysis of non-prepared speech (i.e. spontaneous) data invalidated this hypothesis [12].

Accent phrase duration measured on various styles of spontaneous speech show that the shortest values, corresponding to a single syllable accent phrase, is about 250 ms, even if the single syllable is reduced to a single vowel, which would otherwise take some 100 ms to 150 ms when unstressed [12]. This minimal duration seems to be a limit under which the syllable ceased to be perceived as prominent (i.e. stressed).

The longest duration of accent phrase is about 1200 ms, which corresponds to an average of 7 syllables. This implies that sequences of more than 7 syllables or so must contain more than one stressed syllable, as in the English word *paraskevidekatriafobia* (the fear of Friday 13) realized with two or three stressed syllables: *paraske'**vi**dekatriafo'**bi**a* or *pa'raskevide'katriafo'**bi**a* for example. The question is: why do we have these lower and upper duration limits for accent phrases?

5. Theta and Delta brain waves

In the years 1930-40, researchers observed that the human brain consisted of a very large number of neurons (in the order of 100 billions) interconnected in groups in specific regions of the brain mass. These interconnections allow a transfer of chemically stored information in each neuron. These transfers induce variations of a small electric potential (in the µV range), that can be observed through captors positioned on subjects skull (electroencephalography, or EEG). These electrical variations are called evoked potential as they result from a sensory stimulation, auditory, visual or other.

Electrical activity produced by transfers of group of neurons to other groups of neurons is not done haphazardly. First, they operate in specific frequency ranges linked to specific cognitive activities, and secondly they can be synchronized in phase in each frequency range. Greek letters designate specific frequency ranges: Alpha, Delta, Delta, Gamma… The range of interest here are Delta, varying from 1 to 4 Hz, and Theta, varying from 4 to 10 Hz.

Evoked potential is usually observed with a relatively large number of captors (from 32 to more than 256 today) placed around the subject skull according to location standards. EEG signals are stored in real time and analyzed into the frequency domain with either a (Fast) Fourier or Wavelet transform. The resulting representation is very similar to spectra obtained in frequency speech analysis, the frequency ranges being of course much lower.

The constraints governing temporal groups observed on prosodic structures of both read and spontaneous speech can find a justification – and an explanation – in recent neurophysiological research work on speech ([4], [5], [11]). These studies, based essentially on EEG, investigate the possible correlations that may exist between brain activity and the perception and linguistic treatment of the information by listeners.

Researchers in neurolinguistic for instance, demonstrated with this technique of investigation the precedence of prosodic over syntax treatment [19]. Experiments described in [14] and [4] showed that the speech flow was segmented thanks to prosodic tags and with direct identification of already

memorized units. Two complementary processes would explain the conversion of syllabic flow into higher order units, which would result preferential lateralization to the right hemisphere for the prosodic information, and the left hemisphere for language information already stored.

Following proposals put forward in [3], [4] and [5], these observations lead to the following hypothesis. We know that the waves of the cortex Theta and Delta (among others) govern the flow of information transfer between neuronal groups. Delta wave frequencies ranging from 1 to 4 Hz, while those of Theta waves range from 4 to 10 Hz. These values suggest that Delta waves are responsible for the timing of the transfer of syllabic sequences, the syllables storage in short-term memory being synchronized by Theta waves.

Figures 2 and 3 below illustrate the difference pertaining to EEG recordings resulting from a stimulus of unstructured pure tones (Fig. 1) and a structured sequence organized in 4 chunks of 3 pure tones, the last tone of each chunk with a longer duration (Fig. 2).

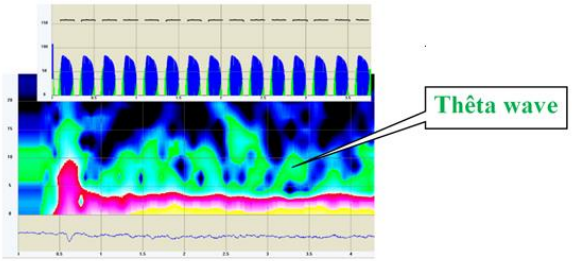


Fig. 2. Example of EEG spectral analysis (channel 28 or Pz) of evoked potential for a stimulus of a sequence of pure tones (top of the figure). Spectral analysis (bottom) shows Theta waves (in the range 4 Hz-10Hz) with no temporal structure [11].

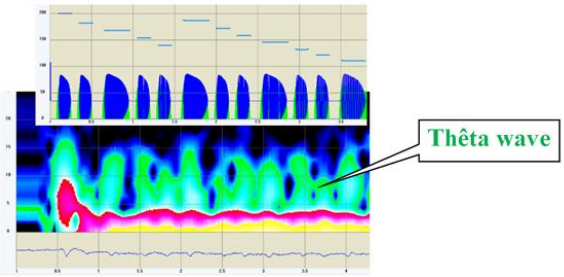


Fig. 3. Example of EEG spectral analysis (channel 28 or Pz) of evoked potential for a stimulus of a sequence of pure tones structured in groups of 3, the third tone being longer (top of the figure). Spectral analysis (bottom) shows Theta waves (in the range 4 Hz-10Hz) organized in a temporal structure corresponding to the stimulus structure [11].

Fig. 2 and 3 above suggest that temporal groups can only be perceived by the listener if their conversion is triggered by Delta waves (which may also synchronize Theta waves). This process is therefore constrained by the Delta wave properties, and in particular by its frequency properties. This hypothesis would account for 1) the extent of variation of the durations of stress groups, ranging from 250 ms to about 1200 ms (variation range of wave periods Delta) and 2) variation periods of Theta waves, from 100 ms to 250 ms.

6. Conclusion

Delta brain waves, whose periods vary from 250 ms to 1200 ms (about 1 Hz to 4 Hz), synchronize the conversion of sequences of syllables stored in short-term memory into higher linguistic units. This hypothesis is validated by the minimal and maximal duration of accent phrases (stress groups) which correspond to the Delta period variations. Furthermore, the average duration of syllables decreases linearly with their number in an accent phrase, from about 250 ms to 100 ms in accent groups of 1 to 7 syllables [11].

As no actual acoustical speech production is involved, silent reading is much faster than reading aloud, where multiple muscular command must be executed. Still, although eye saccade and eye fixation can operate much faster, for example in the reading of pictograms, subvocalization essential in silent reading limits the reading speed. Indeed, since subvocalization implies the generation of sentence prosodic structures as well as sequences of syllables, a prosodic constraint limiting the minimum duration of accent phrase to about 250 ms limits also the speed of the silent reading process, which has to go necessarily through this prosodic structure regeneration process. These values correspond tightly to the fastest reading performances cited in the literature [2], i.e. about 800 wpm or 4 AP per second.

7. References

[1] Blanche-Benveniste, Claire (2003) La naissance des syntagmes dans les hésitations et répétitions du parler, in Araoui J.L. ed. *Le sens et la mesure. Hommages à Benoît de Cornulier*, Honoré Champion, Paris, 2003, 40-55.

[2] Dunning, Brian (2010) Speed Reading, *Skeptoid Podcast.* Skeptoid Media, Inc., 26 Oct 2010. Web. 12 Dec 2013. http://skeptoid.com/episodes/4229

[3] Friederici, Angela & Wartenburger, Isabell, 2010, Language and brain, *Cognitive Science*, (10) 150-159.

[4] Ghitza1, Oded, Giraud, Anne-Lise and Poeppel, David (2013) Neuronal oscillations and speech perception: critical-band temporal envelopes are the essence, *Frontiers in Human Neuroscience*, www.frontiersin.org, January2013, Volume 6, Article340.

[5] Gilbert, Annie & Boucher, Victor (2007) What do listeners attend to in hearing prosodic structures? Investigating the human speech-parser using short-term recall, *Proc. Interspeech 2007*: 430-433.

[6] Giraud, Anne-Lise and David Poeppel (2012) *Cortical oscillations and speech processing: emerging computational principles.* Nature neuroscience E-pub, doi: 10.1038/nn.3063.

[7] Just, Marcel Adam and Patricia A. Carpenter (1987) *The Psychology of Reading and Language Comprehension.* Boston: Allyn and Bacon, 1987.

[8] Liberman, Alvin M. and Ignatius G. Mattingly (1985) The motor theory of speech perception revised, *Cognition* 21 (1): 1–36

[9] Marshall Unger, James (2003) *Ideogram: Chinese Characters and the Myth of Disembodied Meaning*, University of Hawai'i Press, 216 p.

[10] Martin, Philippe (2011) *Ponctuation et structure prosodique*, Langue Française, 2011, nº 172, 99-114.

[11] Martin, Philippe (2013) Contraintes phonologiques de l'intonation de la phrase réinterprétées à la lumière des recherches récentes en neurophysiologie, *La Linguistique*, 2013/1.

[12] Martin, Philippe (2014) Spontaneous speech corpus data validates prosodic constraints, submitted to *Speech Prosody 2014 Conference*, Dublin 2014.

[13] Nowak, Paul (2012) *Speed reading tips: 5 ways to minimize subvocalization,* http://www.irisreading.com/speed-reading/speed-reading-tips-5-ways-to-minimize-subvocalization/

[14] Obrig, Hellmuth, Rossi, Simone, Telkemeyer, Silke & Watenburger, Isabell, 2010, From acoustic segmentation to

language processing: evidence from optical imaging, *Front. Neuroenerg*, 2:13.
[15] Rayner, Keith and Alexander Pollatsek (1989) The Psychology of Reading, Lawrence Erlbaum Associates, Hillsdale, NJ, 544 p.
[16] Rayner, Keith, Barbara Foorman, Charles A. Perfetti, David Pesetsky, and Mark S. Seidenberg (2001) How Psychological Science Informs the Teaching of Reading. *Psychological Science in the Public Interes*t **2** (2): 31-74.
[17] Reichle, Erik D., Pollatsek, Alexander, Fisher, Donald L. and Keith Rayner (1998) Toward a Model of Eye Movement Control in Reading, *Psychological Review* 1998, Vol. 105, No. 1, 125-157
[18] Sereno, Sara and Keith Rayner (2003) Measuring word recognition in reading: eye movements and event-related potentials. *Trends in Cognitive Science* **7** (11): 489 - 493.
[19] Steinhauer, Karsten, Alter, Kai & Friedrici & Angela D., 1999, Brain potentials indicate immediate use of prosodic cues in natural speech processing, *Nature Neuroscience*, 2(2) 191-196.

Rising intonation in spontaneous French: how well can continuation statements and polar questions be distinguished?

Emma Valtersson [1], *Francisco Torreira*[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Emma.Valtersson@mpi.nl, Francisco.Torreira@mpi.nl

Abstract

This study investigates whether a clear distinction can be made between the prosody of continuation statements and polar questions in conversational French, which are both typically produced with final rising intonation. We show that the two utterance types can be distinguished over chance level by several pitch, duration, and intensity cues. However, given the substantial amount of phonetic overlap and the nature of the observed differences between the two utterance types (i.e. overall F0 scaling, final intensity drop and degree of final lengthening), we propose that variability in the phonetic detail of intonation rises in French is due to the effects of interactional factors (e.g. turn-taking context, type of speech act) rather than to the existence of two distinct rising intonation contour types in this language.

Index Terms: rising intonation, question intonation, continuation intonation, polar questions, French

1. Introduction

One of the main challenges to the study of human communication is the lack of a one-to-one mapping between pragmatic function and linguistic form. In this study, we tackle this issue by investigating the prosody of continuation statements and polar questions in French, two utterance types that often exhibit a similar form (i.e. SVO word order and final rising intonation) in spite of their markedly different functions (e.g. keeping vs. yielding the floor, conveying vs. requesting information). Despite the syntactic and intonational similarities of these two utterance types, it is possible that their prosodic characteristics can distinguish them robustly, and that participants in a conversation do not need to rely solely on context in order to interpret their pragmatic function. Our aim in this study is to examine whether a clear distinction can be made between the prosody of the two kinds of utterances in a corpus of spontaneous French.

Several claims regarding the prosodic distinction between rising continuation statements and rising polar questions have been made in the French intonation literature. Based on introspection, [1] claimed that these two utterance types both have final rising pitch, but also that they are differentiated by the scaling and shape of the final pitch rise, with questions reaching a higher pitch maximum and exhibiting a more concave final pitch contour than continuation statements. Two later studies based on spontaneous and read speech [2, 3] supported this distinction in the scaling of the final pitch maximum, but did not find evidence for a distinction in terms of contour shape. Moreover, [3] found a different intensity profile and a shorter duration of the final vowel for continuation statements than for questions.

However, [4], based on the variability observed for both continuation statements and polar questions, claimed that a single underlying rising intonation pattern is compatible with both functions. The possibility that rising polar questions and continuation statements share the same phonological pattern has also been reflected in more recent accounts of French intonation within the Autosegmental-Metrical framework. It has been suggested that the final rise found in both major continuations and polar questions coincides with the end of the highest level of phrasing (i.e. the Intonation Phrase) [5], and that they both consist of the same tonal elements (H*H%) [6].

In summary, there is disagreement in the French intonation literature about the exact prosodic characteristics of rising continuation statements and rising polar questions, and about their possible phonological distinction. In this study, we assess the extent to which these two kinds of utterances can be distinguished phonetically on the basis of their prosody using data from a corpus of French spontaneous conversation.

2. Method

Our data come from the Nijmegen Corpus of Casual French [7], which consists of 23 casual conversations among groups of friends recorded with head-mounted microphones in a sound-attenuated room. We used the initial dyadic part of each conversation, which had a duration ranging from 4 to 40 minutes, and provided a total of around seven and a half hours of spontaneous conversation. Twelve dyads were composed by female speakers, and eleven by male speakers. All speakers originated from Central or Northern France, and were mostly university students, with ages from 18 to 27.

The two authors independently inspected all the conversations in the data and identified cases of rising intonation using Elan and Praat software [8, 9]. We only considered utterances with rising intonation, which were immediately followed by a pause. The reasons for this were to avoid the effects of tonal and segmental coarticulation in the utterance-final syllable, and to make sure that the final rises in our data were all produced in a comparable context in terms of prosodic phrasing. It should also be noted that we did not mark cases with final plateaus, which we observed in tag questions [5] and enumerations [10]. After finishing a first annotation pass, we created a final dataset including all cases of rising pre-pausal intonation that were identified by the two annotators (n = 320). Then, for each utterance, we annotated whether it was a question or a continuation statement in the context of the conversation. Questions typically involved knowledge within the listener's epistemic domain [11] (e.g. 'You went to Paris yesterday?') and were followed by a turn transition and an answer, whereas continuation statements involved knowledge within the speaker's epistemic domain (e.g. 'I went to Paris yesterday') and were usually not followed by a turn-transition. No ambiguous cases were observed when the conversational context was taken into account. In total, we included a total of 126 questions and 190 continuation statements in our dataset.

Several acoustic measurements were performed with Praat [9] in the final syllables of the utterances, where differences in

the final prosody of continuation statements and polar questions were likely to be observed according to previous studies [1,2,3]. We measured the minimum and maximum pitch values (semitones re 100 Hz) in the final vowel of the utterance, which was always present regardless of the structure of the final syllable (e.g. open or closed, with or without an onset). These pitch minimum and maximum values corresponded to the start and end of the intonation rise throughout the final vowel and can be used to investigate differences in scaling between the intonation rises of continuation statements and polar questions. As an estimate of the pitch register used prior to the rise, we measured the median pitch in the interval from the start of the utterance up to the start of the final syllable. In order to control for speaker-based variation in pitch, we normalised pitch values with reference to the median value calculated in an excerpt of ten minutes for every speaker. The shape of the pitch rise has been proposed as being important [1] as well as irrelevant [2, 3] for distinguishing between continuation statements and polar questions in French. In order to examine this feature, we took pitch measures every ten ms throughout the final vowel, and approximated these values with a quadratic equation using least-squares linear regression as in [12, 13, 14]. For each pitch rise, we fitted a linear model $y = a + bx + cx^2$, where a represents the initial pitch value, b the initial slope and c the curvature of the rise. Convex pitch rises are fitted by models with negative quadratic coefficients, whereas concave pitch rises are fitted by models with positive quadratic coefficients. In addition to the pitch measurements above, we calculated the final drop in intensity between the intensity peaks in the last two vowels. Furthermore, we also measured the duration of the final vowel, which was always present regardless of the structure of the final syllable (e.g. open or closed, with or without an onset). In order to control for variability in final vowel duration due to differences in speech rate across tokens, we also calculated the speech rate over the last intonational phrase up to the final syllable, so that it could be included as a covariate in our statistical analyses.

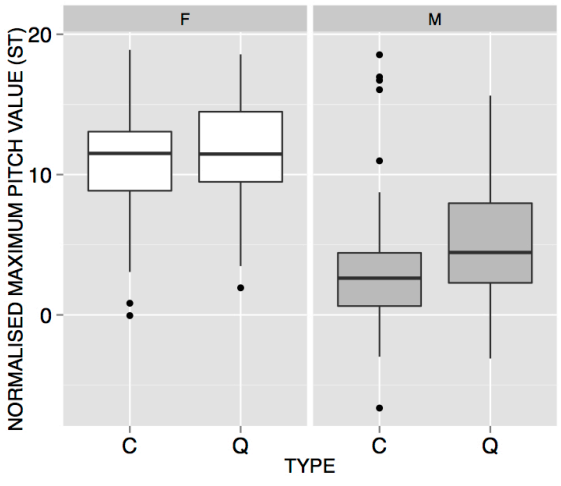


Figure 1: *Speaker-normalised maximum pitch in final vowel (semitones, re 100 Hz) for the two kinds of utterances (C = Continuation statement, Q = Question) for each gender (F = Female, M = Male).*

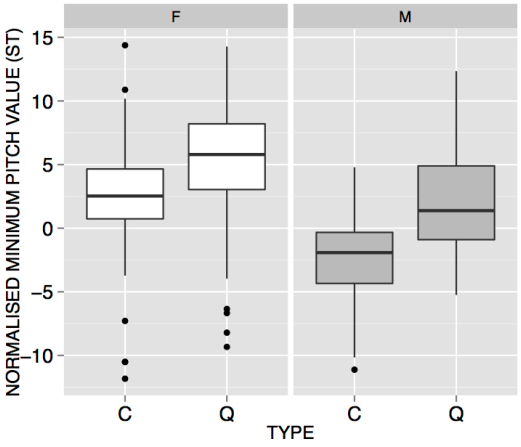


Figure 2: *Speaker-normalised minimum pitch in final vowel (semitones, re 100 Hz) for the two kinds of utterances (C = Continuation statement, Q = Question) for each gender (F = Female, M = Male).*

3. Results

We investigate whether the final prosody of continuation statements and rising polar questions can be clearly distinguished in spontaneous French. Using regression modelling, we first compare the pitch, duration and intensity characteristics between the two kinds of utterances. We focus on the final syllables of the utterances, where relevant differences are expected to occur according to previous literature. Following this, we evaluate the degree of overlap and separation between the two kinds of utterances using a leave-one-out cross-validation procedure.

3.1. Phonetic comparison

Figure 1 shows boxplots of maximum pitch values (i.e. the end of the pitch rise) as a function of utterance type, separated by speaker gender. This figure shows that, excepting a small number of outliers, questions tend to end in higher pitch compared to continuations for males, but not for females. It also shows a clear gender difference, with females having higher values in general. These differences were confirmed by a statistically significant interaction between utterance type and speaker gender in a regression model with maximum pitch as the dependent variable ($\beta = 1.7$, $t = 2.24$, $p < .05$).

Figure 2 shows boxplots of minimum pitch values in the final intonation rise as a function of utterance type, separated by speaker gender. As for maximum pitch, females appeared to have higher values than males in general. Regarding utterance type, questions appear to exhibit a higher minimum pitch value for both females and males. These differences were confirmed by a regression model with minimum pitch as the dependent variable, and utterance type and gender as predictors (utterance type: $\beta = 3.65$, $t = 7.52$, $p < .0001$; gender: $\beta = -4.47$, $t = -9.36$, $p < .0001$). Similarly to minimum pitch, the pitch median up to the final syllable of the utterance also exhibited a higher value for questions. A regression model with the same predictors as for minimum pitch supported this observation (utterance type: $\beta = 2.05$, $t = 6.50$, $p < .0001$; gender: $\beta = -4.39$, $t = -14.12$, $p < .0001$).

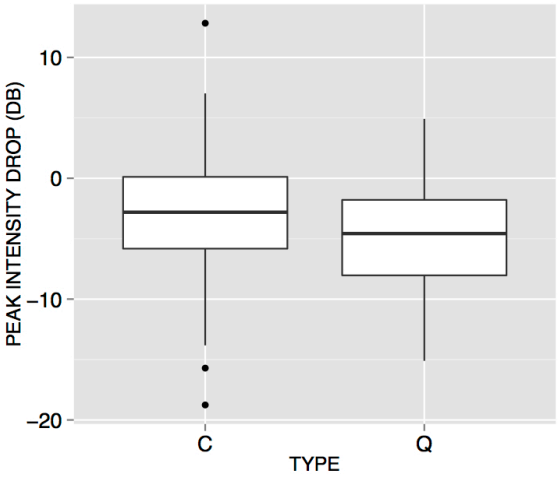


Figure 3: *Intensity drop in final vowel in seconds (dB) for the two kinds of utterances (C = Continuation statement, Q = Question).*

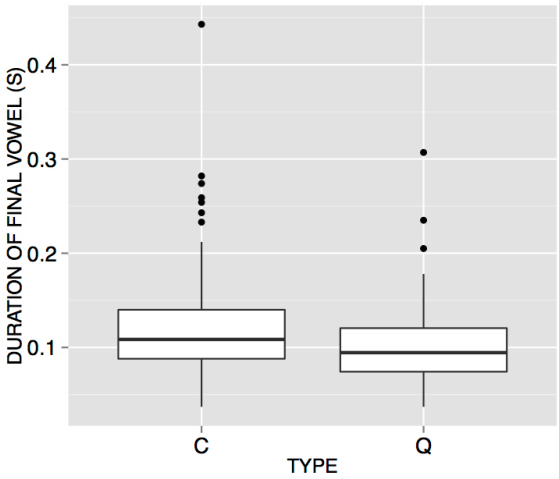


Figure 4: *Duration of the final vowel in seconds (s) for the two kinds of utterances (C = Continuation statement, Q = Question).*

In order to investigate the shape of the rise (i.e. convex vs. concave) we used quadratic equations modelling its trajectory as explained in the Methods section. We fitted a regression model with the quadratic coefficients of these equations, which capture the degree of curvature of pitch rises, as the dependent variable, and utterance type as predictor. The statistical analysis revealed that polar questions and continuation statements did not differ statistically in the curvature of their rises (p = .33). Inspection of the data revealed that the two utterance types could exhibit negative and positive curvature values, indicating that they could both have convex or concave final pitch rises.

Figure 3 displays boxplots of peak intensity in final syllable relative to the peak intensity in the previous syllable. The figure shows that questions tend to exhibit a lower final intensity peak compared to continuation statements. A regression analysis confirmed this difference (β = -1.89, t = -3.71, p < .001).

Finally, we investigated if the duration of the final vowel varies with utterance type. Figure 4 shows boxplots of final vowel duration for the two utterance types. We can see in this figure that questions tend to be produced with a slightly shorter final vowel in comparison to continuation statements. We fitted a regression model with duration as the response, utterance type as the main predictor, and also speech rate as a covariate, since we wanted to control for duration variability related to this factor. This model yielded a statistical difference of roughly 10 ms between continuation statements and polar questions (β = -0.012, t = -2.36, p < 0.05).

In summary, we have found differences between continuation statements and polar questions in terms of several pitch, duration and intensity measures. However, our data also indicate that these two utterance types overlap considerably in these prosodic features, in particular in the case of maximum pitch for females, and intensity and duration for both genders.

The question arises therefore how well the two utterance types can be distinguished on the basis of the several relevant phonetic features in the comparison. We address this question in the following subsection.

3.2. Cross-validation procedure

In order to estimate the degree to which continuation statements and polar questions can be separated on the basis of the cues identified in the previous subsection, we performed a leave-one-out cross-validation of the phonetic cues that appeared to be relevant with our data. This procedure simulated predicting new unseen data using our dataset in the following way: utterance type was predicted for each token in our dataset by means of logistic regression models trained on the rest of the dataset including several features, yielding a percentage of correct classifications for each different model. We used regression models that included several combinations distinguishing the two utterance types in the previous subsection. Models with pitch cues as predictors also included speaker gender as a covariate, whereas models including final vowel duration also included speech rate.

Our automatic classifications yielded different accuracy levels dependent on the model used. The minimum value of the pitch rise offered an accuracy of 72%, close to that of a model comprising all features (76%). The accuracy afforded by maximum pitch, on the other hand, was significantly lower (60%). Combining the two final pitch cues (i.e. minimum and maximum pitch) allowed for a marginal increase in accuracy over the model with minimum pitch alone (73%). Interestingly, the pitch median before the last syllable, which we used as an estimate of the pitch register of the utterance prior to the final syllable, was almost as good a predictor of utterance type as the final pitch cues, with an accuracy of 70%. Regarding non-pitch cues, intensity and duration both provided moderate improvements over chance level (61% for intensity, 67% for duration, and 67% for a model combining duration and intensity).

These results show that not all phonetic features are equally useful for distinguishing between the two utterance types. In particular, the minimum pitch value at the beginning of the pitch rise was the best cue for the contrast between continuation statements and polar questions, performing almost as well as a the full model with all cues, whereas the final pitch maximum only led to a moderate gain over chance level.

4. Discussion and conclusion

This study has investigated whether a clear distinction can be made between the prosody of French continuation statements and polar questions, which both exhibit rising intonation patterns previously claimed to be distinct or similar by different authors [1, 4]. Our comparative analysis in Section 3.1 has shown phonetic differences between the prosody of the two kinds of utterances in terms of several prosodic features. However, a great amount of phonetic overlap between the phonetic realizations of the two utterance types has been observed. This has been evidenced in Section 3.2 by the fact that roughly a quarter of the data was wrongly classified by a logistic regression model containing all relevant prosodic features identified in Section 3.1.

Regarding the phonetic differences, we have observed that the maximum pitch value in the final intonation rise tended to be higher for questions than for continuation statements, but only for male speakers. A more consistent difference was observed between the two utterance types in the minimum pitch value at the beginning of the rise, and in the pitch register prior to the final syllable of the utterance. As for the shape of the rise, we did not observe any differences between continuations and questions. On the other hand, the duration and peak drop intensity of the final vowel showed differences between the two utterance types. Questions tended to have a shorter final vowel and end in a lower peak intensity target than continuations.

Our findings therefore only partially agree with those from past studies. The higher maximum pitch in the rise for questions is in line with the findings of [1, 3], and the more pronounced final drop in intensity for questions agrees with the observations of [3]. Regarding the shape of the final pitch rise, which had been suggested as a relevant feature by [1], we did not observe any consistent differences between the two utterance types, in line with [2, 3]. In our data, final intonation rises in both continuation statements and polar questions could adopt slightly concave or convex shape.

On the other hand, the minimum pitch value in the intonation rise, a feature judged to be irrelevant by [1] and not investigated in detail by [3], clearly offered the best cue for distinguishing between continuation statements and polar questions. Our findings regarding the duration of the final vowel, which was the second-best cue to the contrast between continuations and questions, were also opposed to the previous literature [3]. In our study, polar questions exhibited shorter, not longer, final vowels than continuation statements.

These discrepancies between our findings and those of previous studies may be due to the fact that studies in the past were mostly based on introspection and read speech data, whereas ours used utterances extracted from spontaneous conversations more directly affected by the interaction between speakers. In our study, for instance, the fact that polar questions tended to have less final lengthening than continuation statements may be due to the fact that questions typically yield the floor to the interlocutor, and that their final part usually marks the end of the speaker's turn. This would be in line with observations from studies on English dialogue, which have shown that turn-final utterances tend to exhibit less final lengthening and a more marked final drop in intensity than turn-medial utterances [15, 16]. Since these differences in duration and intensity were observed in utterances with different intonation patterns, it can be concluded that they were not directly related to differences in intonation contour choice, but rather to differences in turn-taking actions (i.e. turn-yielding vs. turn-keeping). It is therefore likely that the observed differences in duration and intensity between continuation statements and polar questions in our French data are due to the different turn-taking contexts in which these utterances tend to occur.

Interactional factors may also be the reason why the minimum pitch value at the start of the rise and the pitch register at which the utterance was spoken before the final syllable, features neglected in previous research, provide the best cues in distinguishing continuation statements from polar questions in our corpus. It has often been proposed that questioning utterances in general, not only those with final rising intonation tend to exhibit higher pitch registers compared to assertive utterances [e.g. 17, 18, 19 among others]. Our finding that the minimum pitch at the start of the final rise and the pitch median before the final syllable were consistently different between continuation statements and polar questions is most likely due to the fact that the pitch register throughout the utterances is generally higher in questions than in continuations.

Our findings therefore appear to be in line with previous proposals [4, 6] that a single rising intonation pattern (e.g. H*H%) is compatible with both continuation statements and polar questions in French. We have observed an important amount of overlap in our data, showing that a clear phonetic distinction cannot be drawn between the prosody of rising continuation statements and polar questions in French. Moreover, we have argued that the phonetic differences observed between the two utterance types (i.e. overall F0 scaling, final intensity drop, degree of final lengthening), are likely to be due to the effects of interactional factors (i.e. turn-taking context). For these reasons, we conclude that continuation statements and polar questions in French both make use of the same rising contour type, and that this contour type is subject to contextual variation which may help cue different speech acts during conversation (signalling that one's turn is not complete vs. asking a question).

5. Acknowledgements

The contribution of the second author was made possible thanks to the financial support of the Language and Cognition Department at the Max Planck Institute for Psycholinguistics, Max-Planck Gesellschaft, and a European Research Council's Advanced Grant (269484 "INTERACT") to Steve Levinson.

6. References

[1] Delattre, P., “Les dix intonations de base du français”, French Review, 40(1):1-14, 1966.

[2] Grundstrom, A., ”L’intonation des questions en français standard”, in Interrogation and Intonation, Studia Phonetica, 8, 19-49, 1973.

[3] Rossi, M., Di Cristo, A., Hirst, D., Martin, P. and Nishinuma, Y., ”L'intonation: de l'acoustique à la sémantique.”, 1981.

[4] Di Cristo, A., “Intonation in French”, in D. J. Hirst & A. Di Cristo [Ed], Intonation Systems. A Survey of Twenty Languages. 195-218, Cambridge University Press, 1998.

[5] Jun, S. A. and Fougeron, C., “Realizations of accentual phrase in French intonation”, Probus, 14(1):147-172. 2002.

[6] Post, B., “French tonal structures”. Speech Prosody, 2002.

[7] Torreira, F., Adda-Decker, M. and Ernestus, M., ”The Nijmegen corpus of casual French”, Speech Communication, 52(3):201-212. 2010.

[8] Brugman, H. and Russel, A., “Annotating Multimedia/ Multi-modal resources with ELAN. Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation. 2004.

[9] Boersma, P. and Weenink, D., ”Praat: doing phonetics by computer ” (Version 5.3. 51). http://www. praat.org. 2013.

[10] Portes, C., Bertrand, R. and Espesser, R. “Contribution to a grammar of intonation in French. Form and function of three rising patterns”, Nouveaux cahiers de linguistique française, 28:155-162. 2007.

[11] Heritage, J., ”Epistemics in action: Action formation and territories of knowledge”, Research on Language & Social Interaction, 45(1):1-29. 2012.

[12] Andruski, J. E. and Costello, J., ”Using polynomial equations to model pitch contour shape in lexical tones: an example from Green Mong”, Journal of the International Phonetic Association, 34(2):125-140. 2004.

[13] Grabe, E., Kochanski, G. and Coleman, J., ”Connecting intonation labels to mathematical descriptions of pitch”, Language and speech, 50(3):281-310. 2007.

[14] Torreira, F., ”Tonal realization of syllabic affiliation in Spanish”, ICPhS XVI, Saarbrücken, 6-10. 2007.

[15] Gravano, A. and Hirschberg, J., “Turn-taking cues in task-oriented dialogue”, Computer Speech & Language, 25(3):601-634. 2011.

[16] Duncan, S., “Some signals and rules for taking speaking turns in conversations”, Journal of personality and social psychology, 23(2):283-292. 1972.

[17] Bolinger, D., Intonation and its uses: Melody in grammar and discourse. Stanford University Press, 1989.

[18] Ohala, J. J., “An ethological perspective on common cross-language utilization of F0 of voice”, Phonetica, 41(1):1-16. 1984.

[19] Gussenhoven, C., “Intonation and interpretation: phonetics and phonology”, Speech Prosody. 2002.

Intonation and focus marking in Ulyap Kabardian

Ludger Paschen[1]

[1] Department of Linguistics, Ruhr-University Bochum
ludger.paschen@rub.de

Abstract

This paper presents a pilot study that aims at establishing a model for the intonation of Ulyap Kabardian in the ToBI framework. On the basis of data gathered during a fieldtrip in 2012, it is suggested that four/three pitch accents and three boundary tones are needed to describe intonation in four communicative contexts. Additionally, it is shown that for focus marking in Ulyap Kabardian questions, a stress shifting rule dislocates word stress to a prosodically determined position. This shift rule is extraordinary in that it is insensitive to stress clashes. From a cross-linguistic perspective, the intonation system of Ulyap Kabardian bears a higher resemblance to the system of one of the Kabardian dialects spoken in Turkey than to Russian, the principal contact language.

Index Terms: Kabardian, intonation, focus, ToBI, polysynthetism, prominence

1. Introduction

The Kabardian language belongs to the Circassian branch of the North Western Caucasian family. The most characteristic typological features of all Circassian dialects are an abundance of consonant phonemes as opposed to only three vowel phonemes, a lack of lexical tones, a highly polysynthetic verb morphology and an ergative/absolutive case marking. There are about 360,000 speakers of Kabardian in Russia [5] and about 1,000,000 speakers in Turkey [11], all of whom are at least bilingual. It is important to note that the Circassian dialects spoken in Turkey differ substantially from those spoken in Russia due to a higher degree of exposure to the contact language [8]. The village of Ulyap (Улы́п) is located in the eastern part of the Republic of Adygeya in Russia. The Ulyap vernacular is a unique idiom in the Circassian family; its status as Besleney dialect, as postulated by [2], is highly questionable.

The aim of this paper is to provide a description of Ulyap Kabardian intonation used in neutral statements, wh-questions, lists and focus constructions. The study will include a prosodic description in the ToBI framework, phonetic measurements and a comparison to related languages and contact languages.

2. Methodology

Five adult female informants with permanent residence in Ulyap were recorded during multiple sessions using a hama EL-80 headset attached to an Olympus LS-5. The recordings were stored as wav-files (44.1 kHz, stereo, 16 kbit/s) and analysed using Praat [4]. The informants were asked to read out loud the sentences given in (1) – (4). The stimuli served to elicit the intonation patterns of neutral statements (1), wh-questions (2)[1], enumerations (3) and questions with narrow focus (4). There are several versions of (4) due to alternations of the verbal prefixes depending on the grammatical role of the focussed phrase; for the sake of simplicity, only two (A and DO focus) will be discussed here.

(1) *se* *s-jə-dze* *me-wəz*
1SG 1SG-POSS-tooth DYN-hurt
'My tooth hurts.'

(2) *sjə* *we* *q'-w-e-wəzə-r*
what 2SG DIR-2SG-DYN-hurt-ABS
'Where is the pain?' (lit: What is it that hurts you?)

(3) *babəʃʲ-əm* *ŝħe* *dame* *q'amzjə-xe-r* *jə-ʔa-xe*
duck-OBL head wing feather-PL-ABS POSS-have-PL
'Ducks have a head, wings and feathers.'

(4) a. AGENS focus:
fatjəme *asłen* *adəɣa+bze* *mə* *kabjənet-əm*
Fatima Aslan Adyghe+language PROX classroom-OBL
ʃʲ-j-e-z-ʁa-ŝ'e-te-r
LOC-IO-APPL-REL.A-CAUS-know-IPFV-ABS
'Was it Fatima who taught Aslan Adyghe in this classroom?'

b. DO focus:
fatjəme *asłen* ***adəɣa+bze*** *mə* *kabjənet-əm*
Fatima Aslan **Adyghe+language** PROX classroom-OBL
ʃʲə-r-jə-ʁa-ŝ'e-te-r
LOC-IO.APPL-3SG.A-CAUS-know-IPFV-ABS
'Was it Adyghe that Fatima taught Aslan in this classroom?'

3. Results

The model for the description of Ulyap Kabardian intonation follows the ToBI convention [3]. Based on the examples that will be discussed in this section, the intonation model for Ulyap Kabardian consists of the following components.

Tiers:

1. *tone* (T): pitch accents, boundary tones and prosodic boundaries
2. *Ulyap* (U): (phonemic) transcription of the Ulyap Kabardian speech sample on the word level
3. *English* (E): translation on the word level

[1] (1) and (2) were presented separately and did not form a dialogue-like sequence.

4. *break indices* (B): numeric values, perceived breaks
5. *word stress* (S): the syllable that has the lexical word stress; also used to indicate stress movement (applicable only in special focus constructions)
6. *misc* (M): notes regarding voice, timing and other comments

Symbols:

- H*, L* high/low tone on the stressed syllable
- H*+L high rising tone on the stressed syllable, followed by a steep fall
- H+L* falling tone which reaches its low target on the stressed syllable
- - ip-boundary without tonal specification
- H-, L- high/low final boundary tone of an ip
- L% low final boundary tone of an IP
- * stressed (prominent) syllable
- (*) destressed (non-prominent) syllable
- ¬ movement of prominent syllable (*stress shift*)

3.1. Neutral statements

The phrase *se sjədze mewəz* 'my tooth hurts' is a neutral statement in which no constituent is specially marked for focus or emphasis. The contour depicted in fig. 1 shows two H- at the end of both ω *se* 'my' and *sjədze* 'tooth', both lacking pitch accents. The only pitch accent in this IP is H+L* on the stressed syllable of the verb. Since the low target is reached late, H+L* (and not L*) was chosen as label. Note that in this example, the second H- is somewhat obfuscated by the following H+L* accent. It is safe to assume that the break and H- after *se* 'my' is due to elicitation as there is no obvious reason why the pronoun should be prosodically separated from its head and form a single ω and even a φ. Later on, it will be argued that NPs in Ulyap usually form a φ and that φs in non-prominent positions are marked with H-. The phrase-final L% may well be preceded by L-, but as of now no evidence for an additional L- can be furnished, which is why the annotation includes only one boundary tone.

3.2. wh-questions

The intonation pattern for open questions in Ulyap Kabardian is composed of one pitch accent and one boundary tone. As shown in figure 2, the wh-word *sjə* 'what' is accompanied by a high tone which is followed by a steep fall that continues to the left edge of the final verb. One could be led to consider a simple pitch accent H* sufficient to describe the pitch contour because of the obvious deaccentuation of the post-nuclear part of the utterance. However, in order to account for the steep fall that reaches its target well before the final L%, it appears to be more reasonable to assume a complex pitch accent H*+L. Note that the timing of the fall is late, reaching to the end of the second word *we* 'you'. The deaccentuated part is analysed as still belonging to the intonation phrase (but cf. e.g. [16]), therefore L% is placed at the right border of the phrase. Since F0 is stable and level during the whole final ω *q'wewəzər* 'hurts', again no ip-tone L- is set at the end of the phrase.

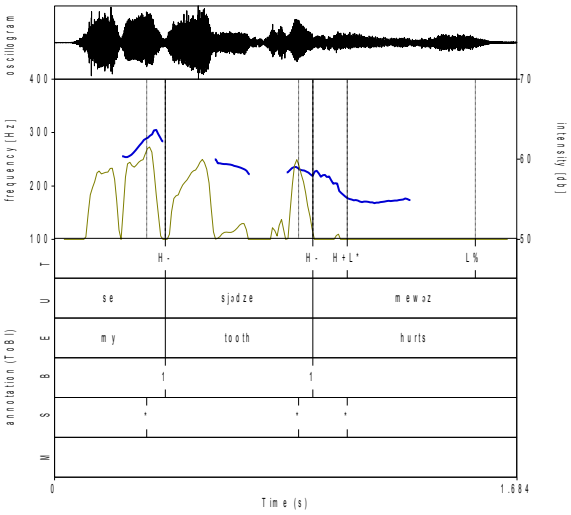


Figure 1: *Neutral statement intonation.*

3.3. Lists and incompleteness

It has been proposed that high F0 universally serves as a prosodic equivalent to incompleteness (e.g. [12], [7]). [1] choose H% to transcribe a final rise in non-final elements of lists in Turkish Kabardian. For Ulyap Kabardian, however, a different tonal analysis is needed, as the contour of the three-item list (fig. 3) suggests. First of all, the lack of resets suggests that the items do not constitute complete intonation phrases but intermediate phrases (the disjuncture between the ips discussed under 3.1 is weaker than here, but for the sake of simplicity, I shall restrict myself to only one phrase type below the IP). Second, the polysyllabic *q'amzjə̂xər* 'feathers' has a high tone in the stressed syllable, but not at the right boundary of their ip. It therefore appears appropriate to use H* followed by an unmarked ip-boundary to account for both the level contour on the monosyllabic *ŝhe* 'head' and the stepped contour on *q'amzjəxər*. A downstep can be observed for the second, but not for the third H*.

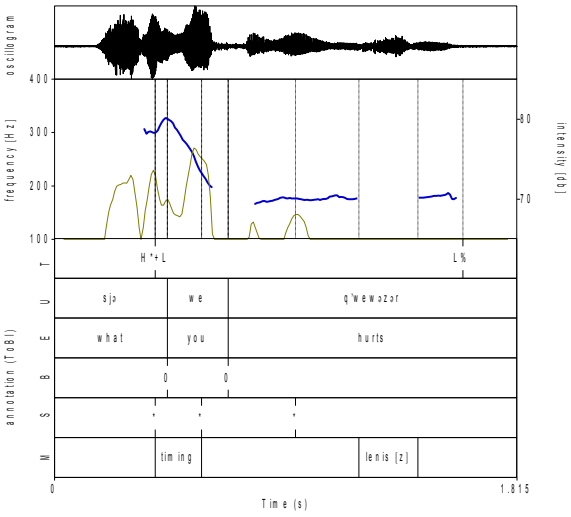


Figure 2: *Wh-question intonation.*

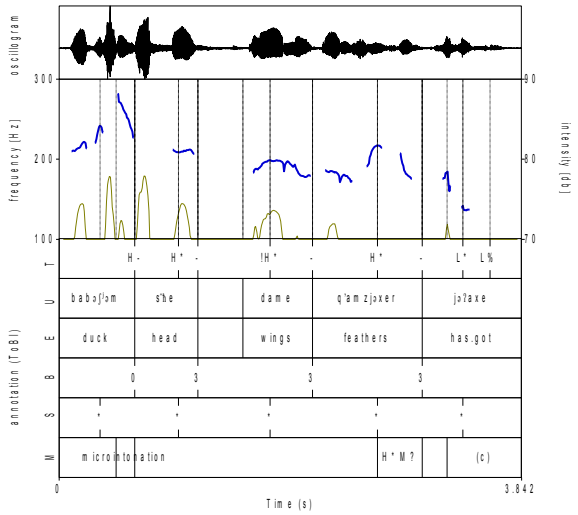


Figure 3: *List intonation.*

F0 on the first word *babə́ʃʲəm* 'duck' is rising steadily, which is best accounted for by a high ip-tone (although no real break can be perceived between the first two words). H- may have to be seen in the context of the topic-like function of *babə́ʃʲəm*, as H- is also used for background elements in questions with narrow focus (see next section). In order to determine the exact function of H- in Ulyap IS, however, further research is necessary.

The final low pitch accent L*, which has already reached its target on the stressed syllable, is followed by a creaky portion that ranges to the end of the phrase. The final *diminuendo* and the voice properties hint at a preceding long breath group, which supports the idea that the list elements constitute only ips, as one would usually expect at least one inhalation in a sequence with three IP-boundaries (H%).

3.4. Focus

In Ulyap Kabardian questions with contrastive focus, non-focussed elements form a φ and are marked with H-, as can be seen in fig. 4 and 5. The most striking feature of those questions is that the position of the most prominent syllable does not coincide with the normal position that word stress predicts. For instance, fig. 4 shows a phrase with focus on the first noun *fatjə́ma* 'Fatima' with obvious stress on the first syllable: not only is it perceptually highly prominent, but it also exceeds the remaining syllables in terms of intensity and spectral clearity. The relevant pitch movement – a high rise followed by a steep fall – should therefore not be ascribed to the left edge of the ip but to the neo-stressed syllable /fa/ (H*+L). Since the post-nuclear fall continues to the right edge of the first word, after which a clear break ensues, an additional low ip-tone L- was added in the annotation.

If one compares this IP to another IP in which a different word (*adəɣábze* 'Adyghe language') is focussed (fig. 5), it becomes evident that it is indeed the first syllable to which stress is (re-)assigned, regardless of its original position. In fig. 5, both intensity and pitch peak are located late on the first syllable, whereas the usually stressed third syllable is now less prominent (though it retains its normal length[2]). The same tonal analysis as in the first example can also be applied to this phrase. Irrespective of the (yet-to-be-defined) prosodic properties of feet in Ulyap Kabardian, the stress shift can be accounted for by the following rule.

(5) *focus stress shift*

$(\sigma_1 \dots \sigma_n)_\omega \rightarrow (\text{'}\sigma_1 \dots \sigma_n)_\omega$ / [__]$_{FOC}$

Similar stress shift rules on the word level have been reported for English in two contexts: focussing bound morphemes (as in *She was included, not éxcluded.*) [14] and when speakers want to avoid stress clash (as in *Japanése mágazines*) [6]. The shift discussed here does not fit either category: there are no bound morphemes in *fatjə́ma*, and the shift can even provoke a stress clash (fig. 5). Curiously, [16] report a similar phenomenon (though not an obligatory rule) for Turkish narrow focus in statements with H*+L, but unfortunately do not elaborate on the matter.

Measuring vowel length revealed that vowels in newly stressed syllables had an average 147.0% duration compared to their counterparts in non-focussed words. However, the deaccentuated vowels were longer (122.9%) as well. In fact, the only vowel that was found to be shorter in the focussed phrase than in unmarked contexts was the first /ə/ of *adəɣabze* 'Adyghe language'.

Intensity was a more reliable indicator for the acoustic measurement of stress. Table 1 and Figure 6 provide an overview of the intensity values of word-stressed vowels in non-focussed constituents (*fatjə́ma*), neo-stressed vowels in focussed constituents (*fátjəma*$_{FOC}$) and their respective counterparts (*fatjə́ma*, *fátjəma*$_{FOC}$). The highly significant differences indicate that the beginning peaks cannot be merely due to an initial rise but have to be the result of prosodic restructuring taking place before the post-lexical level.

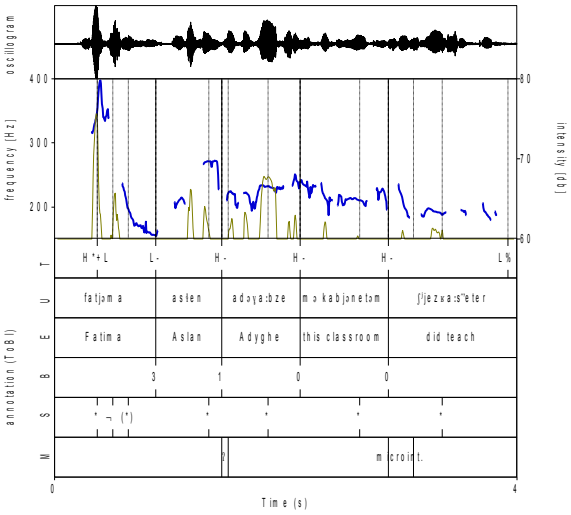


Figure 4: *Question intonation with contrastive focus on the first element.*

[2] The interaction of vowel quality and length in the Circassian dialects is quite complex and cannot be elaborated upon in this paper.

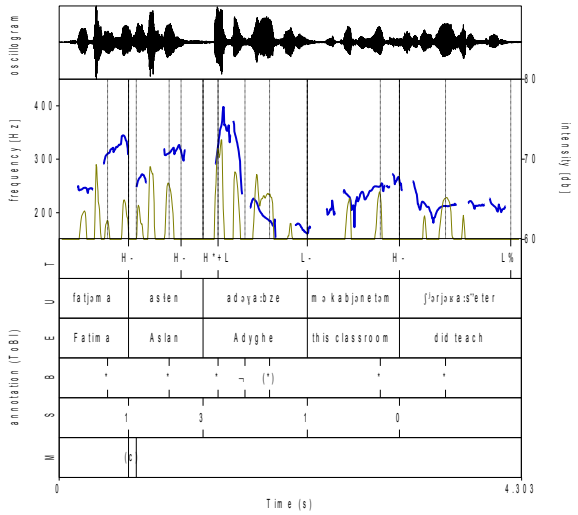


Figure 5: *Question intonation with narrow focus on the third element.*

		word-stress		neo-stress	
		-foc	+foc	-foc	+foc
word-stress	-foc	---	n.s.	*	**
	+foc		---	n.s.	*
neo-stress	-foc			---	***
	+foc				---

Table 1: *Significance levels of intensity values measured in various environments. Vowels in neo-stressed syllables were found to differ significantly (2-tailed t-test, 4 stimuli, 3 speakers) under focus from those in other contexts.*

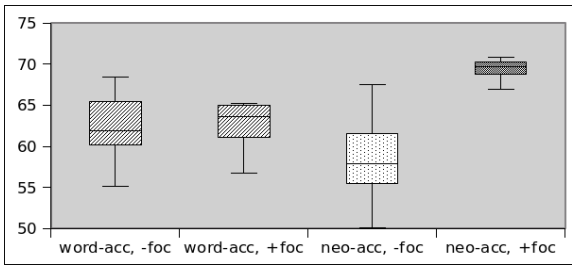


Figure 6: *Boxplot diagrams for intensity values (y-axis, mean energy [dB]) of vowels in different focus and stress environments (cf. Table 1).*

4. Discussion

4.1. Boundary tones

A common feature of all examples discussed so far is the absence of initial and the redundancy of final IP boundary tones. It would be premature to draw any conclusion about IP boundary tones in general at the current state of research. It has become clear, however, that at least for the contexts examined, IP boundary tones – in contrast to ip tones – do not appear to bear any functional weight at all.

4.2. Typological considerations

Table 2 offers a compact synopsis of typical tune patterns in Ulyap Kabardian, Turkish Kabardian [1], Russian [13] and Turkish [9], [16], [10]. The prosodic structure of Ulyap Kabardian shows some remarkable similarities to Turkish Kabardian, the main differences being the tonal interpretations and not the general tune trends. Fewer common features are shared with Russian, the major contact language of Ulyap Kabardian, whereas Turkish Kabardian exhibits remarkable parallels to Standard Turkish. *Stress shift* appears to be present in Turkish [16] but was not reported for Turkish Kabardian [1].

idiom → ↓ context	Ulyap Kabardian	Turkish Kabardian	Russian	Turkish
1 neutral statement	(H-)^{x} H+L* L%	%L H* L%	L* L%	(L H*)^{x} L%
2 wh-question	H*+L L%	%L H* L%	HL* L%	H* {L,H} %
3 list / incompl.	(H-) H*-^{x} L* L%	(H%)$_{x-1}$(H*)$_{x}$ L%	H*M %	?
4 focus question	H*+L L- L%	%L H*HL L%	H*L L%	?

Table 2: *Comparison of intonation patterns for two Kabardian variants and the respective contact languages.*

5. Summary

In this paper, a first model for Ulyap Kabardian intonation in the ToBI-framework was presented. Non-focussed non-verbal constituents in neutral statements are marked with H- only, whereas wh-words have H*+L and list elements have H* and an additional ip-tone. Focus constructions include H*+L and a prosodic rule that shifts word stress of focussed elements to the leftmost ω-position.

As the data analysed in this pilot study cover but a fragment of the Ulyap Kabardian dialect, further studies must take into account a broader data set. It is also beyond doubt that perceptive experiments are necessary to verify the tonal analyses proposed in this paper.

6. Acknowledgements

This research project was funded by the Faculty of Philology at the Ruhr-University Bochum, the Gesellschaft der Freunde der Ruhr-Universität e.V. and the Foundation for Fundamental Linguistic Research fund A-23 (2012). I want to thank G. Moroz for valuable comments during the field trip. Finally, I would like to express my gratitude to all informants for their willingness to contribute to gaining a better understanding of their dialect.

7. References

[1] Applebaum, A. and Gordon, M., "Intonation in Turkish Kabardian", Proceedings of the ICPhS XVI, 2007, 1045-1048.

[2] Balkarov, B. C., "Nekotorye osobennosti beseleneevskogo dialekta kabardinskogo jazyka", in V. Vinogradov [Ed], Trudy Instituta Jazykoznanija 1, 218-230, Moscow, 1952.

[3] Beckmann, M. E., Hirschberg, J. and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in S.-A. Jun [Ed], Prosodic typology, Oxford, University Press, 2005, 9-54.

[4] Boersma, P. and Weenik, D., "Praat: doing phonetics by computer". Online: http://www.praat.org, accessed on 08 Nov 2013.

[5] Colarusso, J., "Kabardian (East Circassian)", Languages of the world, LINCOM, 2006.

[6] Grabe, E. and Warren, P., "Stress shift: do speakers do it or do listeners hear it?", Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV, 1995, 95-110.

[7] Gussenhoven, C., "The phonology of tone and intonation", Cambridge, University Press, 2004.

[8] Höhlig, M., "Kontaktbedingter Sprachwandel in der adygeischen Umgangssprache im Kaukasus und in der Türkei", LINCOM, 1997.

[9] İmer, K. and Çelebi, N., "The intonation of Turkish Cypriot dialect: a contrastive and sociolinguistic interpretation", International Journal of Sociology of Language, 181 (2006), 69-82.

[10] Ipek, C. and Jun, S.-A., "Towards a model of intonational phonology of Turkish: Neutral intonation", ICA Proc., 19, 2013.

[11] Lewis, M. P., Simmons, G. F. and Fenning, C. D., "Ethnologue: Languages of the world". Online: http://www.ethnologue.com, accessed on 11 Nov 2013.

[12] Ohala, J. J., "An ethological perspective on common cross-language utilization of F0 of voice", Phonetica 41, 1-16, 1984.

[13] Odé, C., "Transcription of Russian Intonation. A free interactive research tool and learning module", ACLC, Online: http://www.fon.hum.uva.nl/tori/, accessed on 11 Dec 2013.

[14] Reinhart, T., "Interface Strategies: Optimal and Costly Computation", MIT Press, 2006.

[15] Sumbatova, N. R., "Kommunikativnaja struktura adygejskogo predloženija: Perspektiva i fokus", in J. G. Testelec et al. [Eds], Aspeky polisintetizma. Očerki po grammatike adygejskogo jazyka. Moscow, 559-611.

[16] Özge, U. and Bozsahin, C., "Intonation in the grammar of Turkish", Lingua 120, 132-175, Elsevier, 2010.

Intonation-Based Classification of Language Proficiency Using FDA

Oliver Jokisch[1], *Tristan Langenberg*[1], *Gábor Pintér*[2]

[1]Institute of Communications Engineering, Leipzig University of Telecommunication, Germany
[2]School of Languages and Communication, Kobe University, Japan

jokisch@hftl.de, tristan.langenberg@hftl.de, g-pinter@port.kobe-u.ac.jp

Abstract

State-of-the-art pronunciation tutoring (CAPT) systems are based on ASR technology. Consequently, they can provide a distinguished learning feedback which is focused on phonetic features and the positions of articulation errors. In contrast with the relative success with segmental errors, the acquisition and assessment of second language (L2) prosody is still a challenging problem. Although prosodic parameters like f_0 contour or duration measures are usually displayed, the consequential evaluation components are generally missing. Considering the strong variation in speech data, functional data analysis (FDA) is a useful concept which statistically analyses interrelations between principal components (e.g., given accentuation) and their contribution to superimposed forms (e.g., resulting f_0 contour). This article describes baseline processing and preliminary results of a pilot study on the intonation-based proficiency classification of German by using FDA methods. The experimental part contains the FDA-based classification results compared to a perceptual classification by German natives.

Index Terms: L2 prosody, proficiency, functional data analysis

1. Introduction

Computer-assisted language learning (CALL) and so-called intelligent language tutoring systems (ILTS) have been established components in second language education for more than a decade. Among the proposed methods and systems, automatic pronunciation tutoring (CAPT) plays an increasingly important role. Available CAPT systems offer a wide range of user feedback, such as recorded and reference waveforms, analyzed spectra and underlaying phoneme sequence or animated articulatory organs—often including the intonation contour of uttered phrases. Nevertheless, the pronunciation assessment including the marking of error positions is usually based on segmental (phonetic) features and relies on conventional automatic speech recognition (ASR) modules that rely on hidden Markov models (HMMs) and, for example, use Goodness of Pronunciation (GOP) score as confidence measure [1]. In the system development, elaborate speech databases (originally developed for ASR) can be reused. Although the importance of the prosody acquisition is widely agreed among linguists and teachers, research and development have limited focus on suprasegmental (prosodic) evaluation components. This lack of interest might be surprising, since prosodic core parameters like f_0 contour or rhythmic structures can be easily measured. During the development of the CAPT systems AzAR and Euronounce by TU Dresden and partners [2, 3], effort was invested in suprasegmental databases for the assessment of cross-lingual effects in the acquisition of second (L2) or third language (L3) prosody. The Euronounce database contains 130 speakers of German, Polish, Czech, Slovak and Russian (including 18 language students per L1/L2 pair) and about 200 hours of speech. In further projects the AzAR concept and databases were extended to Mandarin learners of German, to L2 learners of Basque [4, 5, 6], and a baseline method to evaluate intonation contours was suggested.

Considering the strong variation in speech, we found that the functional data analysis (FDA) introduced by Ramsay and Silverman [7, 8] can also provide a powerful approach in speech analysis by statistically exploring interrelations between principal components (e.g., accentuation) and their contribution to forms (e.g., f_0 contour). FDA-based methods in prosodic analysis and synthesis have been already suggested by Gubian et al. [9, 10]. In a recent study, Ward [11] applied principal components analysis (PCA) to several dozen contextual prosodic features in a large set of heterogeneous dialog data. The resulting prosodic components are interpretable as prosodic patterns, including some which involve behaviors of both interlocutors. We intend to apply FDA in different stages of the prosodic assessment—focused on CAPT. In the current article, we describe preliminary results of a pilot study on an automatic intonation-based proficiency classification for German language to test the potential of FDA methods in CAPT environment. In our case the proficiency classification by limited (i.e., only intonational) information is just a working assumption. A detailed prosodic analysis of single speaker utterances or a reliable "overall" proficiency level classification of a speaker is not intended within the scope of this paper. It is clear that the spoken language proficiency is characterized by complex feature sets such as active vocabulary, rules of grammar and phonology, phonetic correctness and so on. Section 2 introduces some previous work on L2 prosody assessment and provides links to proficiency classification. In section 3, we briefly explain the FDA concept. The pilot study on 16 speakers of German is described in section 4—including the L1/L2 database, the baseline processing and the experimental results. The results consist of two parts—addressing FDA-based classification results and a listening test with proficiency classification by native speakers of German.

2. Proficiency and prosody assessment in second language learning

2.1. Proficiency classification

Standardized tests of language proficiency such as the Test of English as a Foreign Language (TOEFL) [12] were already established in the 1960s focusing on reading, listening and writing rather than speaking abilities. In the TOEFL Internet-based Test (iBT) since 2005, the performance evaluation in reading and listening is based on questionnaires. The results of writing and speaking sections are evaluated by three to six human raters which is costly. Consequently, automatic classification methods

are mainly addressing the written proficiency of language learners using algorithms from machine translation [13, 14, 15]. The assessment of spoken language proficiency is focused on phonetic features using GOP or similar measures as already discussed in section 1. Features of non-native prosody, which limit the L2 proficiency, have been studied in different contexts—focused on L2 American or British English. Within the context of this article, some studies dealt with the non-native accent identification [16, 17, 18].

2.2. Prosodic assessment

Beyond the native versus non-natives classification problem, studies of Hönig, Batliner et al. [19, 20] tried to assess L2 productions with respect to intonation and rhythm on a continuous scale, and suggested a suitable set of prosodic features that approximated the decisions of human labelers. In [21], the surveyed feature spaces were extended by acoustic features known from speaker identification tasks (such as short-time spectral features) or general-purpose features from established paralinguistic analysis to indirectly capture complementary prosodic information. The studies consider the perceptual evaluation as a reference—as we do in our work—and describe promising classification strategies. Nevertheless, by fusing different prosodic or even complementary features, the impact of single components (e.g., specific word accentuation or phrase modus variation in the intonation contour) can not be adequately modeled, which results to a less specific user feedback. Our study is targeting on principal intonation components which contribute to the proficiency classification—as an indicator for the influence on L2 prosody—and not on the overall optimal feature representation and classification.

In a previous study [6], we identified Basque as an interesting object of L2 studies in prosody. Basque is an isolated language which does not belong to the Indo-European language group, as one would expect from its geographical location. It has two major neighboring languages, Spanish and French, and the influence of these languages on Basque is noticeable—especially for people studying L2 Basque. In this light it is not surprising that Euskaltzaindia, the Academy of the Basque language, has not yet made a decision about standard prosody due to the variety of accents and intonations across the dialects. L2 students of Basque do not have a clear reference of the preferable pitch pattern and often opt for that of their own L1. This indeterminacy effects the application of conventional quantitative intonation models. The proposition of the Basque study assumes the teacher voice including its intonation pattern as a reference. By providing same text example to the student (cf. "shadowing task"), the intonation quality is simply indicated by the root mean square error (RMSE) between realized f_0 contour of the student and the according reference utterance. For this purpose, f_0 contours need to be normalized on their mean value (gender normalization) and synchronized to the reference by dynamic time warping (DTW) as shown in figure 1.

3. Functional data analysis in prosody

In general, functional data analysis (FDA) names a branch of statistics that analyzes multivariate data which may be treated preferably as curves or surfaces varying over a continuum which is often time, but it can be, among others, spatial location or probability. The data may be subject to measurement errors or even have only an indirect relationship to the curve that they define. It is assumed that the curves are intrinsically smooth.

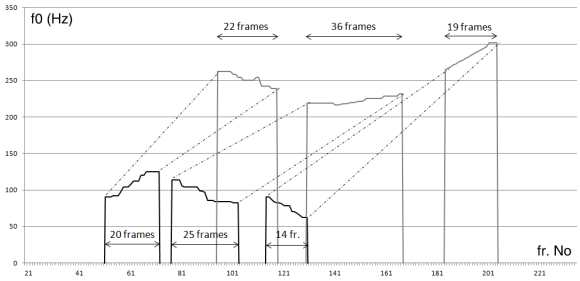


Figure 1: Exemplary f_0 contour mapping: Basque male (reference below) vs. Slovakian female uttering Basque word INDEPENDENTZIA from [6].

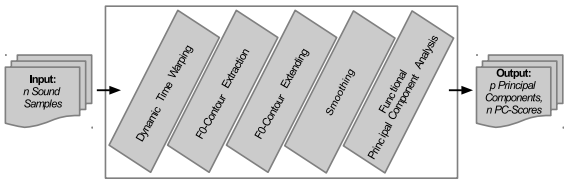


Figure 2: f_0 analysis steps in overview.

Ramsay, Silverman and others developed a set of descriptive and exploratory FDA techniques [7, 8]. Functional data analysis can also use information of slopes and curvatures—reflected in derivatives of the curves—which may reveal aspects of the data generation process. Functional data models and methods resemble those for conventional analyses of multivariate data, including smoothing techniques, regression models, splines and principal components analysis. Gubian et al. introduced the FDA to speech analysis [9] and, in particular, to the f_0 analysis (e.g., for discriminating questions and answers as in [10]). According to Gubian, FDA provides a qualitative/visual description of results and quantitative output in form of statistical values and can be, therefore, interpreted as an "interface between shapes and numbers". We follow this approach and use functional data analysis to abstract away from random f_0 variation in instances of the same utterance. Moreover, the FDA shall extract a maximum of relevant information from our dataset and supports a reliable estimation of the f_0 curves. The analysis process consists of five steps as shown in figure 2. In the first step, all utterances (waveforms) are aligned by dynamic time warping (DTW) to avoid timing mismatch. The subsequent f_0 curve extraction is based on Praat [22]. In the next steps, unvoiced parts are approximated with splines and the resulting contours are smoothed. The final part, the principal component analysis (PCA) for functional data, is the most important analysis step. The PCA reduces the number of functions in the dataset to a lower number of principal components (PCs). The PCs, based on eigenvalues and eigenfunctions by a complex reduction, represent a maximum of information of the whole dataset. For the PCA execution, we need the covariance functions of our data as described in [8],

$$v(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}[x_i(s) - \overline{x}_i(s)] \cdot [x_i(t) - \overline{x}_i(t)] \quad (1)$$

and the definition of the extreme values for the eigenvalues

$$\mu = max_{\xi} \left\{ \sum_{i=1}^{n} [\int \xi(t) x_i(t)\, dt]^2 \right\}. \quad (2)$$

Henceforward one can formulate an eigenvalue problem:

$$\int v(s,t)\xi_j(t)\, dt = \mu_j \xi_j(s). \quad (3)$$

This problem is transformed into the form $\mathbf{V}\vec{\xi} = \mu$ and is solved in the numerical linear algebra by the aid of determinants and the unity matrix which results in PCs—built with their eigenfunctions and eigenvalues (cf. [10]),

$$PC_j(t) = \overline{x}(t) + \sqrt{\mu_j} \cdot \xi_j(t) \quad (4)$$

and principal component scores (PC scores) which are important for the subsequent classification (cf. [8]),

$$\mathbf{C}_{scr}(i,j) = \sum_{j=1}^{p} \sum_{i=1}^{n} \int \xi_j(t)[x_i(t) - \overline{x}(t)]\, dt. \quad (5)$$

By analyzing n utterances (in our case study 16 speaker samples), we calculate p principal components (in our case study 3 PCs) and obtain n (p-dimensional) principal component scores (16 PC scores).

4. Pilot study

4.1. Test design and speaker database

The simple test design aims at the question whether the extracted three principal f_0 components of a short test phrase contain relevant prosodic information to solve two practical targets—as intermediate step for the proficiency assessment:

- Speaker discrimination into native or non-native (L1 vs. L2 speaker of German),
- Proficiency classification on a six-point scale (in the style of the language levels A1, A2, B1, B2, C1 and C2 of the Common European Framework of Reference [23]).

The German test phrase NEIN, SIE KANN ES NICHT. ('No, she can not do it'.) was uttered by 16 speakers (one example per speaker). The test dataset includes eight native (L1 German) speakers and eight non-native speakers with mother tongue (L1) Russian. By disregarding further potential language interferences (e.g., L2/L3 English/German), we consider all non-natives as L2 German speakers. Table 1 summarizes the dataset. The reference classification of the speakers to the mentioned language levels is based on the decision of a language teacher. To simplify the task, we consider the highest L2 speaker level C2 equivalent to the mother tongue level of the L1 speakers (level 1 in the six-point scale).

Table 1: *Test database in overview*

Group	Speaker description	No. of speakers
1	L1 German, male ($L1m_x$)	4
2	L1 German, female ($L1f_x$)	4
3	L2 German, male ($L2m_x$)	4
4	L2 German, female ($L2f_x$)	4

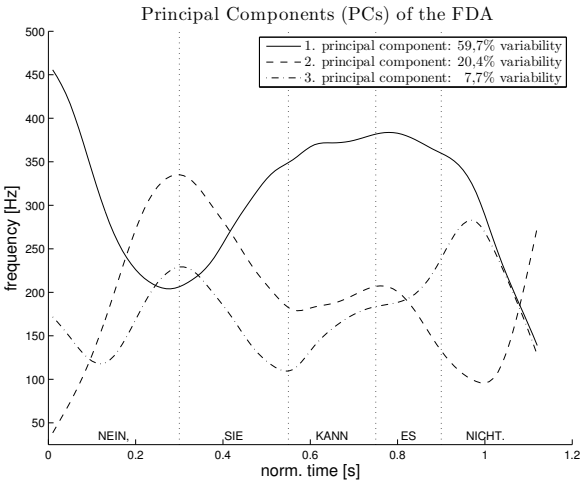


Figure 3: Analyzed principal components of the German phrase: NEIN, SIE KANN ES NICHT. ('No, she can not do it.')

4.2. FDA-based proficiency classification

The first three principal components (PCs) and 16 speaker-based PC scores are calculated as described in section 3. As a reference, the center of mass of each speaker group (L1m, L1f, L2m, L2f) in the PC scores is determined. The root mean square error (RMSE) of each group serves as distance measure. Each speaker distance to the center of mass is divided by the according group RMSE which results to a speaker-specific value of belonging to the certain group—enabling the discrimination into native and nonnatives. By using the PC scores, a geometric distance matrix for the reference speakers is generated (also weighting the PC score dimensions by their information content. The mean deviation of each speaker to its reference value in the matrix is divided by the mean sum of the reference matrix which generates a normalized value of belonging to a certain language level.

4.3. Experimental results

Figure 3 visualizes the first three PCs of the test phrase—representing an information content of 87.8% which we assume as a sufficient accuracy regarding the simple test design. Figure 4 displays all PC scores in two-dimensional PC spaces which leads to three PC combinations with $n = 16$ scores per diagram. The PC scores can be associated with the n f_0 contours. The L1m (male) and L1f (female) speakers create scatter plots which might be associated with feature clusters. Six of the eight L1 speakers can be visually classified. Nevertheless, the L2 speakers do not form visible clusters. Additionally, two L2 speakers seem to cluster with the L1 group.

4.4. Perceptual test

Besides the known proficiency classification provided by language teacher, we performed a perceptual test with 15 naive subjects—ten males and five females with a mean age of 26.3 years. The listeners evaluated 16 utterances (same phrase) in random order and had no prior knowledge about prosodic analysis or the surveyed interrelation to the proficiency assessment.

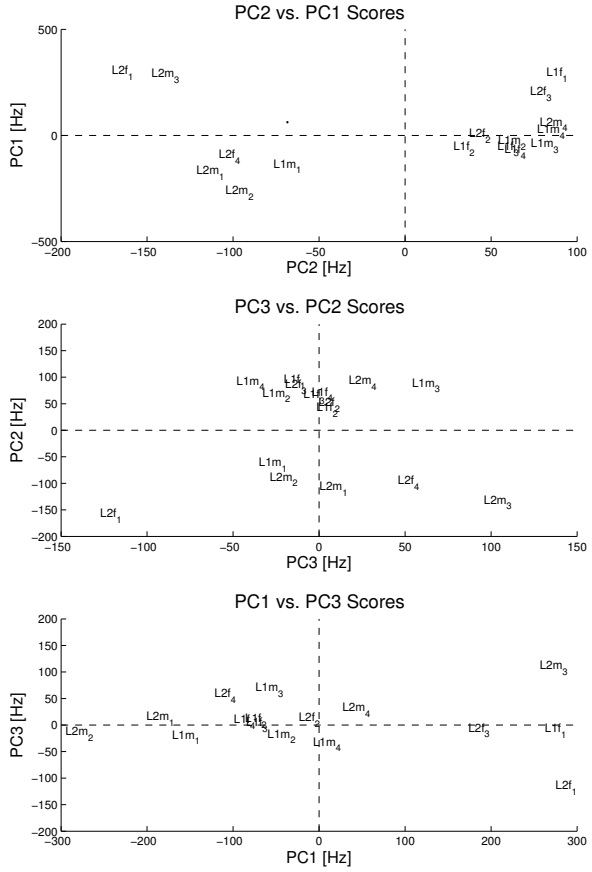


Figure 4: Speaker-specific visualization of the PC scores.

The questionnaire contained the following items:

- Speaker gender (m/f),
- Native speaker of German (y/n),
- Language level on a six-point scale ("1" best).

In L1/L2 classification, the mean decision has a deviation of 0.16 on the normalized scale ("0" L2 . . . "1" L1). In the language level classification, the decisions deviate by 0.98 on the normalized scale ("1" L1 speaker . . . "6" L2 beginner).

4.5. Comparison of FDA-based and perceptual results

By setting the decision threshold to 0.5, the resulting native/non-native classification is shown in figure 5. The "regular classification" is given by the teachers' reference. The listeners' classification ("subject group") is correct for all L1 speakers but includes two errors (speakers 5 and 8) and one border case (speaker 6) in the L2 speaker assessment. The FDA-based classification fails twice in both speaker groups (speakers 1, 5, 12 and 14). Figure 6 visualizes the language level classification of the speakers on the six-point scale. The teachers' classification is assumed as reference. Both, the mean subjective and the FDA-based decisions are correlated to the reference. For nine speakers, the subjective decision closer reflects the teachers' classification. In the remaining seven cases, the FDA-based classification is more accurate.

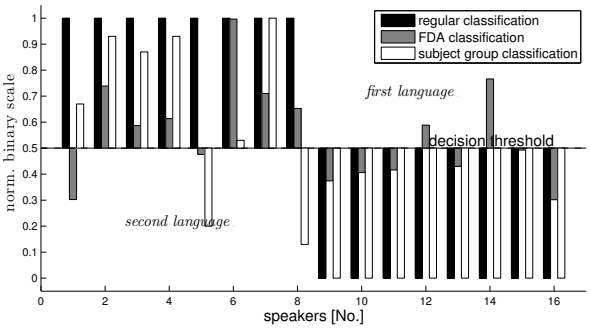


Figure 5: L1/L2 classification by teacher ("regular"), subjects and FDA

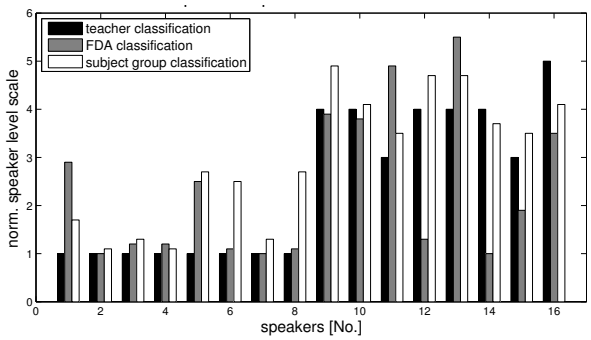


Figure 6: Language level classification by teacher, subjects and intonation-FDA-based ("1" L1 speaker . . . "6" L2 beginner).

5. Discussion

As we expect from our daily-life experience, naive listeners are able to identify non-native speakers and even classify their proficiency on a language level scale with a certain accuracy (widely reflecting the teachers' assessment). In the test design, listeners can use all noticeable problems in a single test phrase such as mispronounced phonemes, wrong segmental durations or accentuation. In contrast, the FDA-based classification is only leaned on the scoring of three principal components extracted from the f_0 contour. It is remarkable that the simple FDA-based classification, which is only using f_0 information and a highly reduced data description, leads to correct classification in the majority of the L1/l2 speaker decisions. The FDA-based language level classification achieves similar results as the perceptual testing, too.

6. Conclusion

The pilot study shows the potential of the functional data analysis in the proficiency assessment but the results are preliminary and need to be consolidated in further experiments with additional data (i.e., more speakers, phrases and variants, further prosodic parameters). In our study we focused on the question how the FDA concept can be utilized for prosodic analysis in the context of pronunciation training. As a preliminary study, two hypothetic classification tasks were carried out with a few simplifications, and only linear classifiers were used. Further extensions of this study will use trainable classifiers and investigate the viability of these methods in real-world assessments in pronunciation tutoring.

7. Acknowledgements

The L1/L2 test data are part of the Euronounce corpus and were recorded by Rainer Jäckel, TU Dresden. We would also like to thank Michael Graf and Ines Rennert, Leipzig University of Telecommunication (HfTL), for their valuable advice.

8. References

[1] Witt, S. M. and Young, S., "Phone-level pronunciation scoring and assessment for interactive language learning," J. Speech Communication, vol. 30, 95–108, 2000.

[2] Jokisch, O., Koloska, U., Hirschfeld, D., Hoffmann, R., "Pronunciation learning and foreign accent reduction by an audiovisual feedback system," in proc. 1st Intern. Conf. on Affective Computing and Intelligent Interaction (ACII), Beijing, 419–425, 2005. Springer LNCS-3784.

[3] Jokisch, O., Jäckel, R., Rusko, M., Demenko, G., Cylwik, N., Ronzhin, A., Hirschfeld, D., Koloska, U., Hanisch, L., Hoffmann, R., "The EURONOUNCE project - An intelligent language tutoring system with multimodal feedback functions: Roadmap and specification," in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 116–123, 2008. Frankfurt.

[4] Ding, H., Jokisch, O., Hoffmann, R., "F0 analysis of Chinese accented German speech," in proc. 5th Intern. Symposium on Chinese Spoken Language Processing (ISCSLP), 49–56, 2006. Singapore.

[5] Hilbert, A., Mixdorff, H., Ding, H., Pfitzinger, H., Jokisch, O., "Prosodic analysis of German produced by Russian and Chinese learners," in proc. 5th Intern. Conf. on Speech Prosody, 2010. Chicago.

[6] Odriozola, I., Jokisch, O., Hernaez, I., Hoffmann, R., "A Pronunciation Tutoring System for Basque - First Development Steps," in proc. Konferenz Elektronische Sprachsignalverarbeitung (ESSV), 2012. Cottbus.

[7] Ramsay, J. O. and Silverman, B. W., "Functional Data Analysis", Springer, New York, 1997.

[8] Ramsay, J. O., Hooker, G. and Graves, S., "Functional Data Analysis with R and MATLAB", Springer, 100–103, New York, 2009.

[9] Gubian, M., Torreira, F., Strik, H., Boves, L., "Functional data analysis as a tool for analyzing speech dynamics – a case study on the French word c'´etait," in proc. Interspeech, 2199–2202, Brighton, 2009.

[10] Gubian, M., Boves, L. and Cangemi, F., "Joint analysis of f_0 and speech rate with Functional Data Analysis," in proc. ICASSP, 4972–4975, Florence, 2011.

[11] Ward, N. G., "Automatic discovery of simply-composable prosodic elements," in proc. 7th Intern. Conf. on Speech Prosody, Dublin, 2014.

[12] Educational Testing Service (ETS), "Test of English as a Foreign Language (TOEFL)". Retrieved December 1, 2013 from https://www.ets.org/toefl

[13] Corsten-Oliver, S., Gamon, M. and Brockett, C., "A machine learning approach to the automatic evaluation of machine translation," in proc. 39th Annual Meeting of the Association for Computational Linguistics (ACL), 148–155, Toulouse, France, 2001.

[14] Lee, J., Zhou, M., Liu, X., "Detection of non-native sentences using machine-translated training data," in proc. Human Language Technology Conference of the North American Chapter of the the Association for Computational Linguistics (NAACL HLT), 93–96, Rochester NY, 2007.

[15] Kotani, K., Yoshimi, T., Kutsumi, T. and Sata, I., "Automatic classification of language learner sentences into native-like or non-native-like based on word alignment distribution," in W. Kouwenhoven (Ed.): Advances in Technology, Education and Development. InTech, Rijeka, 2009.

[16] Tepperman, J., Narayanan, S., "Better non-native intonation scores through prosodic theory," in proc. Interspeech, 1813–1816, Brisbane, 2008.

[17] Piat, M., Fohr, D. and Illina, I., "Foreign accent identification based on prosodic parameters," in proc. Interspeech, 759–762, Brisbane, 2008.

[18] Lopes, J., Trancoso, I. and Abad, A., "A nativeness classifier for ted talks," in proc. ICASSP, 5672–5675, Prague, 2011.

[19] Hönig, F., Batliner, A., Weilhammer, K. and Nöth, E., "Automatic assessment of non-native prosody for English as L2," in proc. Speech Prosody, Chicago, 2010.

[20] Hönig, F., Batliner, A. and Nöth, E., "How many labellers revisited – naives, experts and real experts," in proc. SLATE, Venice, Italy, 2011.

[21] Hönig, F., Bocklet, T., Riedhammer, K., Batliner, A. and Nöth, E., "The automatic assessment of non-native Prosody: combining classical prosodic analysis with acoustic modelling," in proc. Interspeech, Portland, Oregon, 2011.

[22] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer" (version 5.3.05). Retrieved February 24, 2012 from https://www.praat.org

[23] Council of Europe, "Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)". Retrieved December 1, 2013 from https://www.coe.int/lang-cefr

Tonal allophony in Vietnamese: Evidence from task-oriented dialogues

Kieu-Phuong Ha[1], *Martine Grice*[1], *Marc Brunelle*[2]

[1]I*f*L Phonetik, University of Cologne, Germany
[2]Department of Linguistics, University of Ottawa, Canada
hak@uni-koeln.de, martine.grice@uni-koeln.de, mbrunell@uottawa.ca

Abstract

In this paper we investigate the behaviour of the lexical rising tone (SAC) in disyllabic sequences in the Northern variety of Vietnamese. Results from task-oriented dialogues show that this rising tone (SAC), when occurring before the lexical high-level tone (NGANG), can be realised as low level or falling, resembling a different tone in the language (HUYEN). This is the case word-internally and within noun phrases. Two further observations give us an indication that a sandhi process could be developing: (a) this variation is not found in sequences across a larger juncture, and (b) the SAC tone does not undergo this change before other tones.

Index Terms: Northern Vietnamese, tone sandhi, neutralisation, disyllabic sequences, semi-spontaneous speech

1. Introduction

Tone sandhi is a well-known phenomenon in a number of East Asian tone languages such as Taiwanese, Cantonese, Mandarin and other Chinese dialects. It involves 'phonological changes that take place across word boundaries' or 'systematic tone changes, even when they take place word-internally across morpheme boundaries' [1:180]. Examples are *third tone sandhi* in Mandarin Chinese [2], *Min tonal circle* in Taiwanese [2], and *changed tone* in Cantonese [3]. This paper is concerned with the Northern variety of Vietnamese.

Northern Vietnamese has six lexical tones, two of which are level, high (NGANG) and low (or low-falling) (HUYEN), and four of which are contour tones, rising (SAC), falling-rising (HOI), rising-glottalised (NGA), and falling-glottalised (NANG). There is no established tone sandhi in this language, apart from tonal harmony processes in reduplication contexts [4], also referred to as tonal alternation [1]. Moreover, tonal coarticulation has been attested at a phonetic level [5]. However, although sandhi has not been found in controlled experiments, we have informally observed some allophonic variation in conversational speech in younger speakers, more specifically, in disyllabic words, in which the first syllable has a rising tone (SAC) and the second a high level tone (NGANG). In these words the first syllable appears to have a low level or falling pitch, akin to another tone in the language, HUYEN.

Such a change cannot be accounted for by coarticulation, there being no coarticulatory pressure to change a rise before a high-level tone into a fall. However, since this allophonic variation has been observed in spontaneous speech, this study is based on a task that ensures that speakers are formulating their utterances without an orthographic prompt, and that they are interacting with an interlocutor.

We focus in particular on disyllabic sequences within non-reduplicated words and noun phrases. Our specific research questions are whether in conversational speech in the laboratory we obtain 1) a low level or falling allophone of the rising tone (SAC) when produced before the high-level tone (NGANG) and, if this is the case, 2) whether the allophonic variant of SAC resembles the low-level/falling tone HUYEN.

Below we first present an overview of the methodologies used in the elicitation and analysis (section 2). Then we investigate the shape of tone SAC before tone NGANG in sequences with weak junctures against those with larger junctures (section 3.1). In section 3.2 SAC NGANG sequences with weak junctures will be compared with sequences where SAC precedes other tones with comparable junctures. Section 3.3 presents the shape of tone SAC in SAC NGANG sequences against the shape of tone HUYEN in HUYEN NGANG sequences. Conclusions and discussions are provided in section 4.

2. Methodology

2.1. Speech materials and participants

A map task dialogue was recorded using the methodology and design from the HCRC Map Task Corpus [6]. The reason for this choice is that map task dialogues provide real interaction between speakers and assure a certain degree of spontaneity of the data in which the sequences in question are produced.

Two participants have slightly different maps with 11 or 12 landmarks. Some of the landmarks are located differently. While only one map has a route from the start to the end, the other has only the start. The task is for one participant to instruct the other to reproduce the route on the second map without either seeing each other's map. This is done by discussing the route and landmarks. A small screen was placed between the two participants to prevent eye contact. The total duration of the dialogue is approximately 10 minutes. Speaker L is from Thanh Hóa (female, aged 26), speaker M is from Tuyên Quang (female, aged 28). Both provinces are in the Northern Vietnam. The speakers are colleagues and have been living in Hanoi for 4 and 10 years respectively.

2.2. Data transcription and categorisation

The dialogue was orthographically transcribed to enable the selection of the sequences in question. Our target sequences are SAC before NGANG. To investigate the first research question of whether there is an allophonic variant of SAC before NGANG, we first looked at sequences with different syntactic junctures. Then the SAC-NGANG sequences were compared with sequences from a control group, SAC before tones other than NGANG (control group 1). To address the question of whether we are dealing with a sandhi process of SAC-NGANG changed to HUYEN-NGANG (research question 2) we took into account a control group of HUYEN-NGANG sequences (control group 2). Sequences from both control groups had comparable juncture strength.

With respect to the target group, we found in total 75 SAC-NGANG sequences in the dialogue. These sequences were categorised according to their syntactic junctures. We found 4 groups of junctures: i) within disyllabic words, ii) across noun+modifier, classifier+noun, preposition+noun, and verb+object boundaries, iii) across subject+predicate boundaries, and iv) across a sentence boundary. In this paper we analysed only groups i), with the weakest juncture, and iv) with the strongest juncture, as they provide the most cases (in total **36 cases**). Table 1 provides an overview of these two groups of juncture, the frequency of the investigated cases and their syntactic junctures. Sequences that involved a self-repair, a hesitation, or a strong interaction with the intonation at the right edge of the utterances [7] were also excluded.

Table 1. *The frequency of occurrence and examples for each type of syntactic juncture for SAC NGANG sequences (word-for-word glosses in small capitals, corresponding translations in italics)*

Grp	Freq	Syntactic junctures and examples
1	19	disyllabic words
		phía trên SIDE TOP *above*, **phía ngang** SIDE CROSS *across*, **trái tim** *heart*, **ngón tay** *finger*
2	17	sentence+sentence (or fragment) or sentence+particle
		Không phải phía bên phải đâu **nhớ** NEG OUGHT SIDE SIDE RIGHT NEG PARTICLE *It is not on the right hand you know,* **bên** trái điểm-đến SIDE LEFT DESTINATION *on the left of the destination.* mình vẽ một nửa cái bán-cầu **đấy** WE DRAW ONE HALF CLF HEMISPHERE THAT **thôi** ONLY *We draw only half of that hemisphere.*
	∑=36	

Table 2 provides examples for the target group of SAC-NGANG sequences (TG) and two control groups (CG), and the frequency of these sequences; control group 1 (SAC before other tones: SAC-OTHER) and control group 2 (HUYEN-NGANG) had comparable juncture strength.

Table 2. *Examples and frequency of sequences in the target group SAC-NGANG disyllabic words (TG) and 2 control groups (CG1: SAC-OTHER, CG2: HUYEN-NGANG) (word-for-word glosses in small capitals, corresponding translations in italics)*

Grp	Freq	Examples
TG	19	**phía trên** SIDE TOP *above*, **phía ngang** SIDE CROSS *across*, **trái tim** *heart*, **ngón tay** *finger*
CG		
1	44	**uốn lượn** *to curve*, **cánh đồng** *field*, **nối thẳng** *link directly*
2	16	**đầu tiên** *firstly*, **vòng qua** *go round-to*, **rồi xong** *then*
	∑=79	

2.3. Acoustic analysis

To quantify the shape of the contour on SAC syllables, we first extracted the F0 values and calculated the distance between the F0 at the vowel onset and the value at the end of the syllable. The measurement at the vowel onset is chosen to mimimize effects of microprosody due to onset consonants on SAC syllables. If the values of this measure are positive, this indicates that SAC has a rising contour. If the values are negative or zero, it is an indication that SAC is falling or level. This step was conducted, not only for the target sequences SAC-NGANG, but also for the two control groups SAC-OTHER (control group 1) and HUYEN-NGANG (control group 2). All measurements were conducted using Praat.

3. Results

3.1. SAC-NGANG sequences across junctures

We examined whether the juncture strength between the two syllables in SAC-NGANG sequences affects the realisation of the SAC tone by comparing the interval in semitones between the beginning of the vowel and the end of the SAC syllable in the two different juncture conditions. Values for each of the juncture conditions are plotted individually for each speaker in Figure 1.

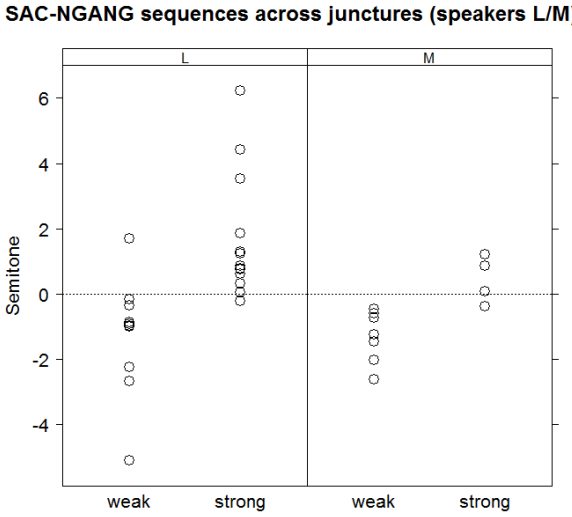


Figure 1: *Distribution of intervals in semitones between vowel onset and end of syllable for the initial syllable with rising tone (SAC) in SAC-NGANG sequences across two juncture strengths. Negative values=falling, positive values=rising; speaker L: weak juncture (N=12), strong juncture (N=13); speaker M: weak juncture (N=7), strong juncture (N=4).*

Here we can see that there is little overlap across the two juncture strengths in the realisation of SAC in SAC-NGANG. Specifically, we can observe a great number of low level or falling realisations (indicated by negative values or values around 0st) and very few rising realisations of the rising (SAC) tone (indicated by positive values) when the juncture is weak. Both speakers show the same tendency. Their roles in

the task (speaker L being instruction giver, speaker M being follower) resulted a greater number of utterances for L than M. The investigated sequences occur in both phrase-medial and phrase-final positions indicating that SAC is changed before NGANG regardless of the position of the sequences in phrases. When the juncture was strong, by contrast, there were predominantly rising realisations, i.e. the tone was realised in its canonical form as in citation contexts.

Figure 2 shows two examples of the contour on SAC in SAC-NGANG sequences produced by speaker L. The first sequence is the word "phía ngang" (Fig. 2a) *across* taken from the utterance:

Không đi lên đâu mà đi **phía ngang**.
NEG GO UP NEG BUT GO SIDE CROSS
Don't go up but go across.

The second sequence (Fig. 2b) has a strong juncture between the two syllables, namely between a sentence and a fragment:

xong vòng lên trên bênh-viện **nhớ,**
THEN CIRCLE UP ABOVE HOSPITAL IMP.PARTICLE
lên trên nóc bệnh-viện ý
UP ABOVE TOP HOSPITAL PARTICLE
Then go up to the hospital, up to the top of that hospital.

We see that in the first case, the contour on the SAC syllable is falling, whereas the contour in the second one has a rise similar to the lexical tone of the particle.

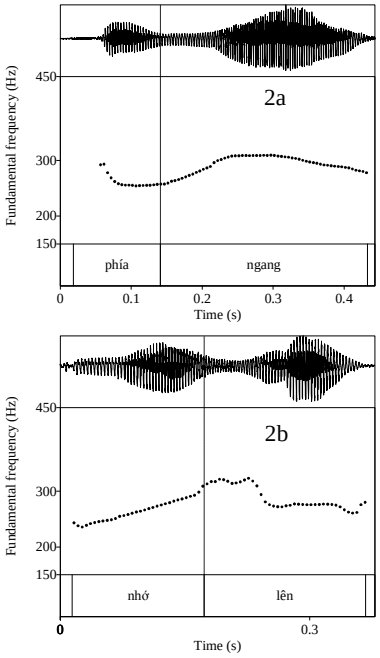


Figure 2: *Examples of the contour of SAC before NGANG in a compound (2a) and in a sequence with a strong juncture between the two syllables (2b). Dotted line shows the boundary between the two syllables.*

3.2. SAC NGANG vs. SAC OTHER sequences

The question that arises is whether, in cases of weak juncture, SAC tends to be produced as falling or level before only the tone NGANG or whether it also has this tendency before other tones. Figure 3 shows the interval in semitones between the vowel onset and end F0 in SAC syllables in SAC-NGANG sequences and SAC-OTHER sequences (all with weak junctures, see Table 2). Almost all values for SAC in SAC-NGANG are negative or at the 0 level, showing that the contours on SAC in these sequences are falling or level. By contrast, the values for SAC in SAC-OTHER sequences are almost all positive, showing that the contours on SAC syllables are rising. Again, both speakers show the same tendency. These results provide evidence of an allophonic variation of SAC when occurring before NGANG: in SAC-NGANG compounds or sequences with minimal junctures, SAC appears to have a falling or level contour, whereas in SAC-OTHER sequences with comparable junctures, tone SAC has a rising contour.

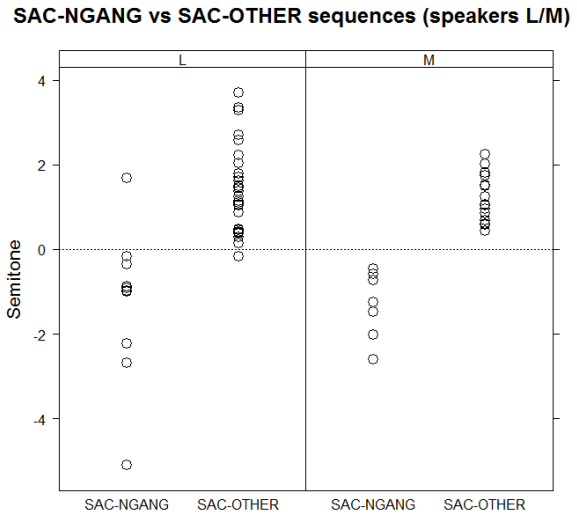


Figure 3: *Distribution of intervals in semitones between vowel onset and end of syllable for the initial syllable with rising (SAC) tone in SAC-NGANG sequences (speaker L, N=12; speaker M, N=7) and SAC-OTHER sequences (speaker L, N=28; speaker M, N=16).*

3.3. SAC NGANG vs. HUYEN NGANG sequences

So far, we have provided evidence of the tone SAC being produced as low level or falling before the tone NGANG. We compared this tone with the lexical low-level/falling tone HUYEN in the sequences HUYEN-NGANG. Figure 4 shows the intervals in semitones between the vowel onset and end F0 in SAC syllables in SAC-NGANG sequences and in HUYEN in HUYEN-NGANG sequences (again, all with weak junctures, see Table 2). Since speaker M only had one realisation of HUYEN-NGANG, we considered only the data of speaker L. The figure shows that a great deal of the intervals have negative values showing that the contours of both tones are falling or level. This indicates that tone SAC produced by speaker L resembles tone HUYEN when they precede tone NGANG.

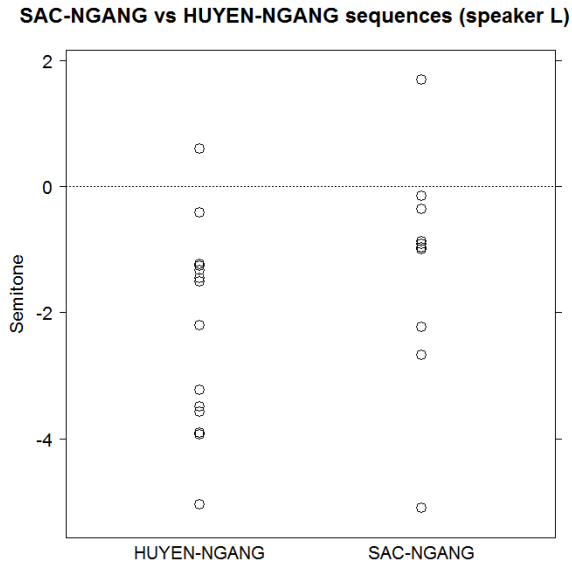


Figure 4: *Distribution of intervals in semitones between vowel onset and end of syllable for the initial syllable with rising (SAC) tone in SAC-NGANG sequences (N=12) and for the initial syllable with low-level/falling (HUYEN) tone in HUYEN-NGANG sequences (N=15).*

4. Discussion and Conclusion

In this study we found that in disyllabic words and within noun phrases the lexical rising tone (SAC) is realised as falling or level when occurring before the lexical high-level tone (NGANG). This tendency does not hold when there is a strong juncture between the two syllables, such as the juncture between two sentences or between phrases and particles. Furthermore, the rising tone (SAC) rarely changes before other tones, suggesting an allophonic variation of SAC before NGANG. The comparison of the rising tone (SAC) and the low-level/falling tone (HUYEN) before NGANG shows that these two tones resemble each other, confirming our informal observation that SAC might be neutralised to HUYEN when preceding NGANG in disyllabic words and within noun phrases. Here we only considered the first syllable in the sequences, but our results suggest that SAC may leave residual cues on NGANG for at least one speaker (L). A corpus controlled in terms of possible microprosodic effects is needed to investigate these possible residual cues. Nonetheless, the type of change on the first syllable of the sequence - from rising to falling before a high tone on the second - cannot be accounted for by coarticulation.

Historically, tone sandhi often originates from allophonic variants of tones [cf. 8]. Thus, there might be an emerging sandhi process of SAC changed to HUYEN when preceding NGANG in certain varieties of Northern Vietnamese. Although our speakers, L and M, come from provinces outside of Hanoi, they have been living in the capital for quite a long time (4 and 10 years respectively). It is so far unclear whether this variation of SAC before NGANG is restricted to Hanoi, and that our speakers have adapted this trend, or whether it is more geographically widespread.

Another issue that needs to be addressed is the question whether the neutralised SAC tone is comparable to the neutral tone in Mandarin Chinese [9]. Our data suggests that this is probably not the case, as we found the change of tone SAC not only in disyllabic sequences involving function words (e.g. classifiers) but also in disyllabic words. Tone SAC appears in these sequences to be comparable to tone HUYEN, rather than seemingly targetless or with a neutral mid target [9].

Further investigations will look at the second syllable (NGANG) and expand the corpus to include speakers of different ages and different speaking styles. Moreover, perception tests are necessary to ascertain whether listeners can actually distinguish between SAC and HUYEN before NGANG. These will help us understand the nature and extent of the allophonic variation and whether we are indeed dealing with an emerging sandhi process.

5. Acknowledgements

We would like to thank our two speakers for taking part in the recordings. Thanks are given to Vũ Kim Bảng and Đinh Hằng, Institute for Linguistics Hanoi, for their help with the recordings and discussions on the data. We thank Bastian Auris for his help with the data processing. This investigation is supported by the German Research Foundation on the project “Tone and Intonation in Vietnamese” at the University of Cologne and by the SSHRCC on the project “Prosodic Typology: Insights from Vietnamese and Eastern Cham” at the Universities of Ottawa and Cologne.

6. References

[1] Yip, M. (2002): Tone. Cambridge: CUP.
[2] Chen, M. (2000): Tone sandhi: patterns across Chinese dialects. Cambridge: CUP.
[3] Yu, Alan C. L. (2007). Understanding near mergers: The case of morphological tone in Cantonese. Phonology 24 (1): 187–214.
[4] Pham, A. H. (2002): Vietnamese Tone. A New Analysis. New York & London: Routledge.
[5] Brunelle, M. (2009): Northern and Southern Vietnamese tone coarticulation: A comparative case study. Journal of the Southeast Asian Linguistics Society 1:49-62.
[6] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. and Weinert, R. (1991). The HCRC Map Task Corpus. Language and Speech, 34, 351-366.
[7] Brunelle, M. & and Hạ, K. P. & Grice, M. (2012): Intonation in Northern Vietnamese. The Linguistic Review, 29(1):3-36.
[8] Gussenhoven, C. (2004): The phonology of tone and intonation. Cambridge: CUP.
[9] Chen, Y., & Xu, Y. (2006). Production of weak elements in speech: Evidence from neutral tone in Standard Chinese. Phonetica 63: 47-75.

Laryngealization or Pitch Accent – the Case of Danish Stød

Nina Grønnum

Department of Scandinavian Studies and Linguistics, University of Copenhagen, Denmark

ninag@hum.ku.dk

Abstract

According to recent proposals Danish stød is the phonetic manifestation of a HL tonal pattern compressed within one syllable, making the stød/non-stød distinction a special case of the more general tonal word accent distinction in Swedish and Norwegian. This review of the relevant aspects of Danish stød and intonation demonstrates that (1) such a tonal representation of stød is contradicted by the phonetic reality. (2) Stød is distributed in words according to roughly the same principles across regional varieties of Danish, but tonal patterns are highly variable. (3) Word accents in Swedish and Norwegian are associated exclusively with stressed syllables, whereas stød occurs also in less than fully stressed syllables, devoid of autonomous pitch movements. (4) A word in Swedish and Norwegian can have one pitch accent only, but Danish words may have more than one stød.

Index Terms: stød, laryngealization, tone, Danish

1. Introduction

The acoustic, perceptual as well as formal properties of standard Danish stød and standard and regional Danish intonation are documented in [1-10]. There is a long tradition in Denmark to describe stød as a kind of creaky voice, explicitly independent of tone, from Høysgaard [11-13] through Martinet [14], Hjelmslev and Andersen [15-16], to Basbøll [17-18]. The phonology and morphology are extensively accounted for and formalized in Basbøll's *Non-Stød Model* [19-20]. Recent years, however, have seen proposals for a different analysis of Danish stød in [21-27], namely as the phonetic by-product of a H and L tone compressed within one syllable, inspired by Kiparsky's analysis of Livonian stød, [28], although Kiparsky himself confines the tonal analysis to Livonian and does not extend it to Danish. There could be several incentives for such a proposal. (1) It is entirely justifiable on physiological grounds, given what is known about the larynx and the vocal folds in F_0 lowering [29-31]. (2) In several South East Asian tone languages, low tone is often accompanied by laryngealization, as in Mandarin and Cantonese [32]. (3) There is an undisputed diachronic relation between the stød/non-stød distinction and the Accent I/Accent II distinction in Swedish and Norwegian, and the idea that stød arises from tonal contours is common in diachronic theories of *stødgenesis* [33-34]. (4) There is also a certain synchronic similarity in the distribution of Accent I/II and stød/non-stød. (5) Stød as pitch accent would unify Danish stød and the Scandinavian word accents in the current autosegmental-metrical framework as, e.g., in [35].

2. Laryngealization

Stød is prototypically a kind of creaky voice: non-modal voice with aperiodic vibrations and irregular amplitude, often but not invariably accompanied by a local fundamental frequency perturbation, an abrupt and brief F_0 dip, typically contained within the second half of long vowels or in the sonorant consonant after short vowels [1-2, 36], consonant with Basbøll's contention in [19] that stød is a property of the second mora of bimoraic syllables. A series of acoustic and perceptual investigations in the early 2000s, summarized in [3-4], showed, however, that the exact acoustic properties, the timing, and the segmental domain of stød are highly variable: vocal fold vibrations are more or less explicitly irregular; the irregularity may onset simultaneously with the stressed vowel or later in the syllable nucleus; it may be contained within the syllable rhyme or it may spill over into a succeeding post-tonic syllable. The considerable acoustic variability does not seem to affect perception: Stød is as clearly audible in the word in the middle as on the left in Figure 1 in spite of the rather stark contrast between the vibratory patterns in the two vowels. A further notable characteristic is its robustness: In fast or non-distinct speech styles, where segments and syllables are freely weakened or lost, stød is faithfully produced and perceived.

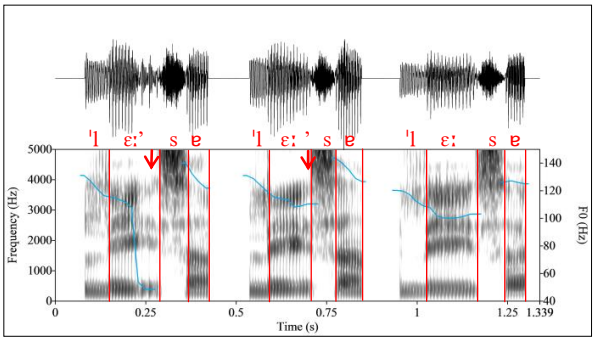
For a speculative account of the neuro-physiological mechanism behind this rather astounding acoustic variability in stød manifestation, see, e.g., [4]. – The variability in Danish stød is similar to what Blankenship reports in [37] about Mazatec and Mpi, and she finds *laryngealization*, i.e. stiffening of the vocal folds, which may or may not result in creaky voice, to be a better concept. So do Garellek and Keating in [38]. As in Danish stød, Gerfen and Baker find in [39] that laryngealization in Coatzospan Mixtec is highly variable within and across speakers and often realized with very subtle F_0 and amplitude cues.

3. Pitch

Figures 1-6 and 8 are Praat pictures ([40]), displaying spectrograms with superposed linearly scaled F_0 tracings. In figures 1 and 6 the microphone signal is included in order to elucidate the vibratory patterns underlying the F_0 perturbations. But otherwise the pertinent facts about stød as they are outlined in the text appear unambiguously from F_0 contour and spectrogram alone. Words depicted in apparent isolation in Figures 1, 4, 5, and 8 are spliced out from within longer utterance contexts. The utterances were scripted and read aloud, except in Figure 4 which presents non-scripted speech.

3.1 Standard Danish in Copenhagen is not HL

Figure 1 demonstrates that (1) the F_0 perturbation, the deep and steep fall, in the word on the left is contained within a very narrow time frame, not to be mistaken for a falling tone distributed over the whole syllable; and (2) there is an evident overall similarity, from lower stressed to higher unstressed syllable, in all three words – the two with stød left and mid and the one without on the right. This is the F_0 pattern associated with stressed and post-tonic syllables in Copenhagen Danish, cf. [8-10], apparent also in Figures 2 and 4-6. It would be characterized as L*H in current notational practice. A maximally developed F_0 pattern describes a very modest initial F_0 fall in the stressed vowel, succeeded by a steep and considerably larger rise, typically 3-4 semitones, to the first post-tonic syllable, and then a fall through succeeding post-tonics, if any. The shorter the stress group, the less extensive the F_0 pattern:



Figur 1: *Three words:* læser *'reads' with explicit stød (left), less explicit stød (mid), and* læser *'reader' without stød (right). Arrows point to the laryngealized part of the vowel. Male Copenhagen speaker.*

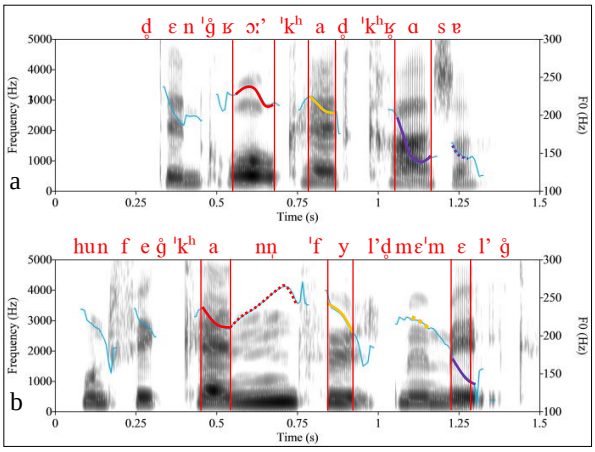


Figure 2: Den grå kat kradser *'The grey cat scratches' (a), and* Hun fik kanden fyldt med mælk *'She had the jug filled with milk' (b). Stressed vowels are touched up in full lines: red (first), yellow (second), and purple (third); unstressed vowels and the syllabic consonant in corrresponding dotted lines. Female Copenhagen speaker.*

compare *kanden* in Figure 2b with the two long words in Figure 4. In the absence of any post-tonic syllable, all that remains is the slight initial fall as in *grå* and *kat* in Figure 2a. Note specifically the pairwise similarity between the F_0 contours within the stressed vowels of (i) stødless *kan(den)* and *grå* with its nearly invisible stød (red lines); (ii) stødless *kat* and *fyldt* with stød and an explicit local F_0 drop in the [l] (yellow lines); (c) stødless *krad(ser)* and *mælk* with stød and a likewise explicit local F_0 drop in the [l] (purple lines). The red, yellow and purple stressed vowel lines are also situated, pairwise, at the same F_0 level in utterance a and b, respectively. In other words, the presence or absence of post-tonic syllables does not affect the scaling of the stressed syllables in the speaker's range. That precludes a suggestion in [35:224-5] that the pitch pattern in Danish is in fact HL, only the H is delayed (i.e.: to the post-tonic). But in the absence of any post-tonic syllable to carry the H tone, one would expect the stressed syllable to move upwards in the range and take up its rightful H position. That does not happen.

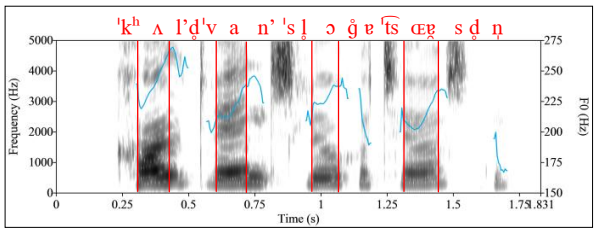


Figure 3: Koldt vand slukker tørsten *'Cold water slakes one's thirst.' Female Aalborg speaker.*

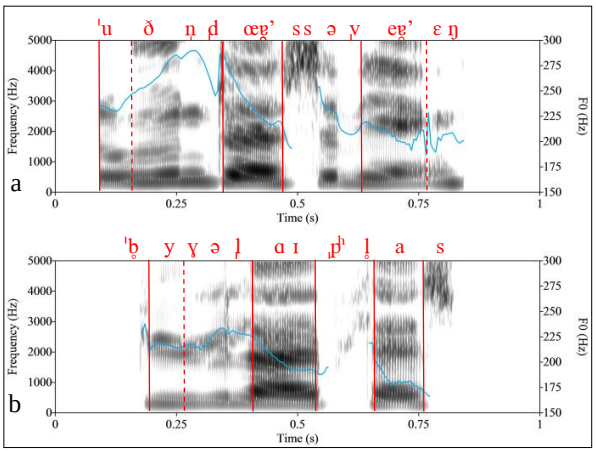


Figure 4*: Two compound words:* udendørsservering *'open air serving' with stød in the two syllables with secondary stress (a), and* byggelegeplads *'adventure playground' without stød (b). Female Copenhagen speaker.*

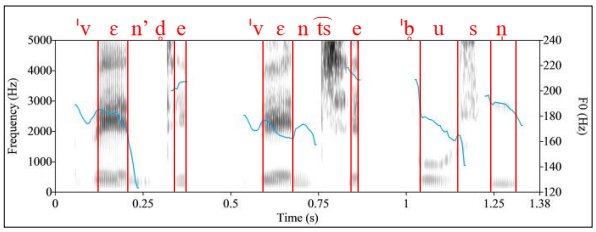


Figure 5: ... vend det ... *'turn it' (left),* ... ven til ... *'friend to' (mid), and* ... bussen ... *'the bus' (right). Same speaker as in Figure 2.*

Stød as a local F_0 perturbation, a brief and more or less explicit lowering of F_0, is independent of its location on the F_0 pattern. It may occur at the low turning point, prior to the rise to the post-tonic, as in *læser* in Figure 1 left, or *fyldt med* in Figure 2b. It may occur at the top of the F_0 pattern as in *koldt* and *vand* in Figure 3, and low on the falling flank in a long series of post-tonics as in *udendørsservering* in Figure 4a. There is nothing to distinguish stød in one position from stød in any other position. The fundamental F_0 similarity between stød and non-stød is also evident in the figures. Note particularly that there is nothing reminiscent of two HL pitch accents in the smoothly falling F_0 contour associated with the

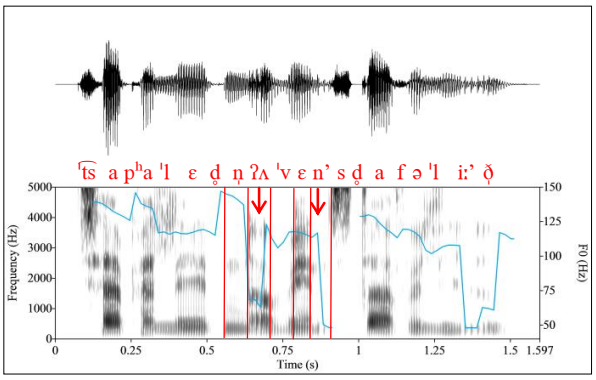


Figure 6: Tag paletten og vend staffeliet *‘Take the palette and turn the easel’. Same speaker as in Figure 1.*

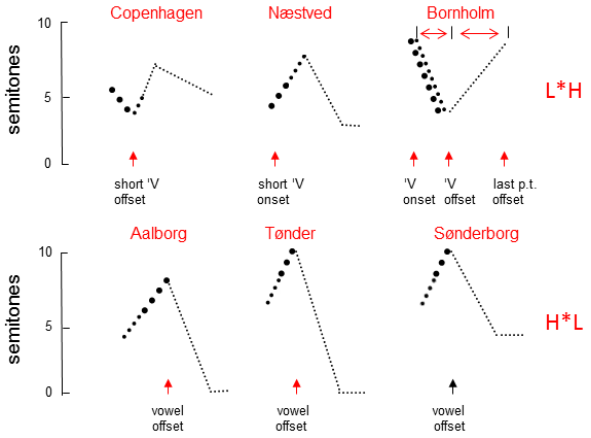


Figure 7: *Prosodic stress group patterns in six varieties of Standard Danish. Vertical arrows point to the segmental anchor point. See further the text.*

second and third parts of the compound *udendørsservering* in Figure 4a. In this respect it resembles the stødless word below in 4b, *byggelegeplads*, though the latter is produced within a narrower frequency range.

Figure 5 depicts two stressed monosyllables succeeded by an unstressed word, *vend det* with stød, *ven til* without stød, and a stødless disyllable *bussen*. Again, the overall F_0 patterns in the three disyllabic sequences are identical: a movement from lower stressed syllable to the higher post-tonic, three L*Hs.

Below the two arrows in Figure 6 are two very similar F_0 events: very steep and very local falls framing the sequence *og vend*. The first is associated with the glottal onset at the juncture before the vowel in *og*, the second accompanies the stød in *vend*. Presumably, no one would suggest that glottal attack at vowel onset be associated with a phonological pitch accent. Under a tonal analysis, then, the leftmost F_0 perturbation is the result of a glottal onset which is not quite a complete glottal closure here, but comes out as creaky voice, whereas the same F_0 perturbation on the right would be an autonomous tonal gesture with a laryngealized side effect.

Note also how microprosodic effects may induce extensive F_0 movements in a vowel. An uninitiated observer might ascribe a falling pitch accent to the first syllable of *bussen* in Figure 5: the extent of its F_0 fall is nearly as comprehensive as the fall in *vend*. But the first steep part of the fall in *bus-* is due to the transition from the unvoiced stop consonant to the vowel, the final steep movement is due to the transition from vowel to obstruent, and in fact the vowel is perceived as a non-dynamic level pitch. The pertinent, perceptible movement in the three disyllabic sequences – whether or not they contain a word boundary – is the movement from the lower stressed syllable to the higher post-tonic. None of these patterns contain a HL tonal sequence, they are all perceptually LH.

To sum up: The pitch pattern in Copenhagen Danish is essentially the same in words with and without stød, and it is LH, not HL. Furthermore, under temporal constraint the pattern is *truncated* rather than *compressed* into the stressed syllable. A representation in terms of a HL tone compressed into one syllable is as far removed from the phonetic reality as can be.

3.2 Regional varieties of Standard Danish

Figure 7 depicts, in six different locations in Denmark, stylized tracings of the F_0 patterns associated with the *prosodic stress group*: a stressed syllable plus all succeeding unstressed syllables, if any, irrespective of intervening word boundaries. They are not impressionistic drawings or educated guesses but based on acoustic data from a considerable number of recordings of four speakers from each region, recording the same scripted material under identical conditions, cf. [8]. Heavy dots depict the location of short stressed vowels on the F_0 pattern, medium dots depict the extension in long stressed vowels, and fine dots depict the course of unstressed syllables in the prosodic stress group. In Bornholm, the vertical strokes enclose movements which may be expanded or compressed in time, as suggested by the horizontal arrows, in concordance with the duration of the stressed vowel (long or short, the falling part), and the number of post-tonic syllables in the stress group (the rising part).

These F_0 patterns would adequately represent trisyllabic prosodic stress groups. This is of no particular concern for Næstved, Aalborg, Tønder and Sønderborg, because F_0 will generally reach its low minimum already in the second post-tonic and continue low and level. But in Copenhagen, where the fall from the high turning point in the first post-tonic is less steep, it would typically continue to fall further, beyond what is depicted in Figure 7, only to level out around a fourth or fifth post-tonic, as in *udendørsservering* in Figure 4a. In other words, there is no fixed pitch relation between a L* and the termination of a *preceding* stress group pattern: the L* will be approached from above after short stress groups, cf. *fyldt* in Figure 2b, and from below after long stress groups, cf. ... *med ru-* ... in Figure 8.

The variation in shape and range of F_0 patterns is considerable. Differences among them are easier to capture when the governing principle in their execution is made explicit: In all of them, except Bornholm, there is one point only in the melodic fragment which is constrained in terms of its alignment with a segment in the prosodic stress group, as indicated by the vertical arrows in the figure. This point is low relative to the first post-tonic in Copenhagen and Næstved and high in Aalborg, Tønder and Sønderborg. The former two may accordingly be characterized as L*H, the latter three as H*L.

The anchor is invariably associated with the stressed vowel, namely with its offset across the board in the H*L

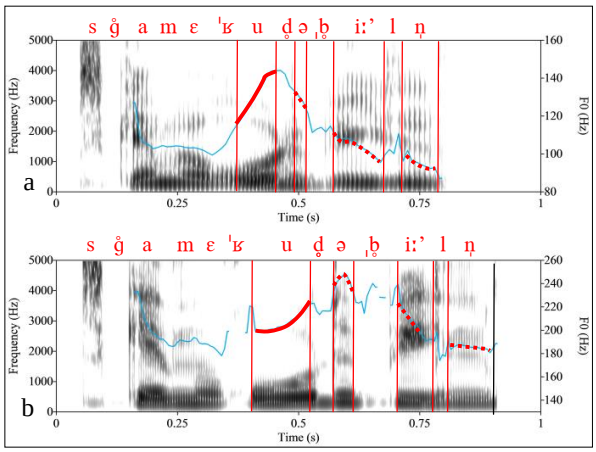


Figure 8: ...skal med rutebilen... *'are going by bus', male Aalborg speaker (a), female Næstved speaker (b).*

varieties, but with its onset in Næstved. In Copenhagen the low turning point coincides with the offset of the stressed vowel if that vowel is short, and occurs about halfway through a long stressed vowel. Leaving out Bornholm for the moment, *neither segments nor syllables in the prosodic stress group on either side of the anchor have separate tonal representations*. They are simply strung out on the melody like pearls of varying length on an undulating string. Voiceless segments interrupt the melody and may have very local microprosodic effects but do not otherwise interfere with the score. A complete pattern only materializes in so far as there are syllables to carry it, otherwise it is truncated from the end. Accordingly, monosyllabic stress groups with a short vowel surrounded by voiceless consonants exhibit only the part of the melody indicated by the heavy dotted line part in the five patterns in Figure 7.

Figure 8 shows two actual passages, ... *skal med rutebilen*... by a male speaker from Aalborg (a) and a female speaker from Næstved (b). The frequency scale spans exactly an octave in both cases, so the ranges covered by the two patterns are immediately comparable. The offset of the stressed vowel (full red line) coincides exactly with the peak of the F_0 pattern in Aalborg, as stipulated in the stylization in Figure 7, and the fall from the peak is rapid and extensive. Since the vowel is short, it does not make it to the peak of the pattern in Næstved, and the peak therefore coincides with the first post-tonic syllable, also as stipulated in the stylization in Figure 7. The fall is not as extensive as in Aalborg. The stød in the second part of the compound is acoustically rather weak in both instances, introducing a moderate perturbation only, although it is clearly audible in both cases.

Bornholm stands out from the other varieties by the very elastic relation between segments and F_0: the duration of the falling part of the tonal pattern varies – albeit slightly – with the duration of the stressed vowel so that that the low turning point coincides with the offset of the vowel, be it short or long. The duration of the rise varies with the duration of the unstressed part of the stress group so that the rise terminates on the last post-tonic syllable. This latter variation is considerable, between one and many post-tonics, and perceptually the rise is more conspicuous than the fall. In other words, the pattern has three targets, a high onset, a low turning point and a high offset. There is, however, a limit to how fast the fall-rise may be executed, a compressibility maximum, and when this limit is reached, as in a monosyllabic stress group with a short vowel surrounded by voiceless consonants, the fall disappears, and what remains is a steep rise from a low onset to a high offset. Given that the fall may be deleted and that otherwise the rise in the F_0 pattern is perceptually more salient than the fall, Bornholm may also be adequately represented as L*H.

Among the L*H varieties, two have stød, Copenhagen and Næstved, whereas Bornholm does not. Likewise, among the H*L varieties Aalborg has stød, but Tønder and Sønderborg do not. So much for an insoluble correlation between tonal movement and stød in Danish.

3.3 Stød under non-primary stress

Pitch accents in Swedish and Norwegian are associated with primary stressed syllables, and hence words – also compounds – can each have one pitch accent only, [41]. But we have stød in syllables with non-primary stress, cf. *udendørsservering* in Figure 4a and *rutebilen* in Figure 8. And, contrary to expectations in a tonal representation, these syllables are not associated with independent, autonomous F_0 movements. And of course, when syllables with secondary stress can have stød, words will have more than one when the stressed syllable also has stød as in, for example, [ˈˌgʌlˈfˌbʌlˈˌd] *golfbold* 'golf ball', [ˈlanˈsman'] *landsmand* 'fellow countryman'.

4. Discussion and conclusions

Despite all the pros listed in the introduction, stød as the manifestation of a compressed HL tone faces insurmountable empirical obstacles. There is nothing in the phonetic reality to support it. The overall F_0 pattern is similar in words with and without stød, and when stød is not accompanied by any F_0 perturbation at all, stød and non-stød F_0 patterns are identical. Likewise, the principles which govern the presence or absence of stød in words of different structure, are roughly the same across regional varieties of Danish, in so far as they have stød at all, but F_0 patterns are not. In other words, the laryngealization typically present in Danish stød is not a phonetic accompaniment to a compressed HL pitch accent. On the contrary: the F_0 perturbation is a by-product of the laryngealization and not invariably present. Laryngealization is the articulatory, acoustic, and perceptual constant in stød production, F_0 perturbation is not. Such a state of affairs is not exclusive to Danish either. Voice quality differences in some of the South East Asian languages are not merely the synchronic phonetic accompaniment to tonal differences. On the contrary, tones developed from phonation types, not the other way around [42-44]. Furthermore, in so-called *laryngeally complex* languages, like Mpi [45], Jalapa Mazatec [38], and Comaltepec Chinantec [46], tonal and phonatory contrasts coexist and cross-classify. Thus, laryngealization may accompany any tone, whether high or low.

Space does not permit a review of the similarly serious obstacles in the phonology and grammar of Danish stød, if represented as a H and L tone compressed in one syllable. The reader is referred to [6].

In conclusion: laryngealization as an autonomous syllable prosody, orthogonal to pitch and intonation, suffer none of the shortcomings attached to a tonal representation and is altogether more satisfactory than any representation of Danish stød as underlyingly tonal in nature.

5. References

[1] Fischer-Jørgensen, E., Phonetic analysis of the stød in Danish, University of Turku, 1989a.

[2] Fischer-Jørgensen, E., "Phonetic analysis of the stød in standard Danish", Phonetica 46:1-59, 1989b.

[3] Grønnum, N. and Basbøll, H., "Consonant length, stød and morae in standard Danish", Phonetica 58:230-253, 2001.

[4] Grønnum, N. and Basbøll, H., "Danish Stød: Phonological and Cognitive Issues", in M.-J. Solé, P.S. Beddor and M. Ohala, Experimental Approaches to Phonology, 192-206, OUP, 2007.

[5] Grønnum, N. and Basbøll, H., "Danish Stød – Towards simpler structural principles?" in O. Niebuhr [Ed], Understanding prosody – The Role of Context, Function, and Communication, 27-46, Walter de Gruyter, 2012.

[6] Grønnum, N., Vazquez-Larruscaín, M. and Basbøll, H., "Danish Stød: Laryngealization or Tone?", Phonetica 70:66-92, 2013.

[7] Grønnum, N., "Two issues in the prosody of Standard Danish", in A. Cutler and D.R. Ladd [Eds], Prosody – Models and Measurements, 27-38, Springer Verlag, 1983.

[8] Grønnum, N., "Prosodic parameters in a variety of regional Danish standard languages, with a view towards Swedish and German", Phonetica 47:182-214, 1990.

[9] Grønnum, N., The Groundworks of Danish Intonation. An Introduction, Museum Tusculanum Press, 1992.

[10] Grønnum, N., "Superposition and subordination in intonation - a non-linear approach", Proc. 13th Int. Cong. Phonetic Sc., Stockholm, vol. II:124-131, 1995.

[11] Høysgaard, J.P., Concordia res parvæ crescunt, eller Anden Prøve af Dansk Orthographie, København, 1743. Reprinted in [47], 219-248.

[12] Høysgaard, J. P., Accentuered og Raisonnered Grammatica, København, 1747. Reprinted in [47], 249-488.

[13] Høysgaard, J. P., Første Anhang til den Accentuerede Grammatika, København, 1769. Reprinted in [48], 507-550.

[14] Martinet, A., La phonologie du mot en danois, Libraire C. Klincksieck, 1937. Also in Bulletin de la Société Linguistique de Paris 38: 169-266.

[15] Hjemslev, L., "Grundtræk af det danske udtrykssystem med særligt henblik på stødet", Selskab for Nordisk Filologi. Årsberetning for 1948-49-50, 14-24, 1951. English translation in Essais Linguistiques II, Travaux du Cercle Linguistique de Copenhague 14:247-266, Nordisk Sprog- og Kulturforlag, 1973.

[16] Andersen, P. and Hjemslev, L., Fonetik, Rosenkilde og Bagger, 1967. Originally chapter XIV and XV of Nordisk Lærebog for Talepædagoger, 233-354, Københavns Universitets Fond til Tilvejebringelse af Læremidler.

[17] Basbøll, H., "Some remarks concerning the stød in a generative grammar of Danish", in F. Kiefer [Ed], Derivational Processes, 5-30, KVAL, Stockholm, 1972.

[18] Basbøll, H., "Stød in Modern Danish", Folia Linguistica XIX.1-2:1-50, 1985.

[19] Basbøll, H., The Phonology of Danish, OUP, 2005.

[20] Basbøll, H., "Stød, diachrony and the Non-stød Model," North-Western European Language Evolution 54/55:147-189, 2008.

[21] Riad, T., Curl, stød, and generalized accent 2. Proc. Fonetik 98 Stockholm, 8-11, 1998.

[22] Riad, T., "The origin of Danish stød", in A. Lahiri, Analogy, levelling, markedness: principles of change in phonology and morphology, 261-300, Mouton de Gruyter, 2000.

[23] Riad, T., "Eskilstuna as the tonal key to Danish", Proc. Fonetik 2009 Stockholm, June 10-12, 12-17, 2009. Accessed June 2013 at http://www2.ling.su.se/fon/fonetik_2009/proceedings_fonetik2009.pdf.

[24] Itô, J. and Mester, A., "Stø^{2}det i dansk", handout, Scandinavian Summer School in Generative Phonology, Hvalfjarðarströnd, Iceland, June 16–28, 1997.

[25] Itô, J. and Mester, A., "(Un)accentedness and the Perfect Prosodic Word", manuscript 42 pp. Submitted to Linguistic Inquiry.

[26] Morén, B., "Danish Stød and Eastern Norwegian Pitch Accent: The Myth of Lexical Tones", The 13th Manchester Phonology Meeting, 2005a. Accessed June 2013 at http://www.hum.uit.no/a/moren/cv_files/morenMFM13.pdf.

[27] Morén, B., "Danish Stød and Eastern Norwegian Pitch Accent: Prosody, Morphology and Non-tonal Lexical Specification", The 11th Meeting on the Norwegian Language, Bergen, Norway, 2005b. Accessed June 2013 at http://www.hum.uit.no/a/moren/cv_files/ morenMons11.pdf.

[28] Kiparsky, P., "Livonian Stød", manuscript 16 pp, 1995, updated 2006. To appear in Boersma, P., van Oostendorp, M., Hermans, B. and Kehrein, W. [Eds], Segments and Tone, Niemeyer. Accessed June 2013 at http://www.stanford.edu/~kiparsky/Papers/livonian. pdf.

[29] Lindblom, B., "F_0 lowering, creaky voice, and glottal stop: Jan Gauffin's account of how the larynx works in speech", Proc. Fonetik 2009, Stockholm, June 10-12, 8-11. Accessed June 2013 at http://www2.ling.su.se/fon/fonetik_2009/proceedings_fonetik 2009.pdf.

[30] Lindqvist-Gauffin, J., "Laryngeal mechanisms in speech", STL-QPSR 2-3:26-32, 1969. Accessed June 2013 at http://www.speech.kth.se/prod/publications/files/qpsr/1969/1969_10_2-3_026-032.pdf.

[31] Lindqvist-Gauffin, J., "A descriptive model of laryngeal articulation in speech", STL-QPSR 13(2-3), 1-9, 1972. Accessed June 2013 at http://www.speech.kth.se/prod/publications/files/qpsr/1969/ 1969_10_2-3_026-032.pdf.

[32] Yu, K. M., "Laryngealization and features for Chinese tonal recognition", Proc. Interspeech 2010. Accessed June 2013 at http://www.linguistics.ucla.edu/people/grads/krisyu/output/p41721.pdf.

[33] Verner, K., "Review of Kock: Språkhistoriska undersökningar om svensk accent", Anzeiger für deutscher Altertum VII:1-13, 1878. Reprinted in Afhandlinger og Breve. Edited by Selskab for Germansk Filologi, 84-104, Frimodt, 1878.

[34] Storm, J., „Om tonefaldet (tonelaget) i de skandinaviske sprog", Forhandlinger i Videnskabsselskabet i Christiania 1874, 286-29, 1874.

[35] Gussenhoven, C., The phonology of tone and intonation, CUP, 2004.

[36] Smith, S., Bidrag til Løsning af Problemer vedrørende Stødet i Dansk Rigssprog, Kaifer, 1944.

[37] Blankenship, B., "The timing of nonmodal phonation in vowels", J. Phonetics 30:163-191, 2002.

[38] Garellek, M. and Keating, P., "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec", J. Int. Phonetic Ass. 41/2:185-205, 2011.

[39] Gerfen, C., and Baker, K., "The production and perception of laryngealized vowels in Coatzospan Mixtec", J. Phonetics 33: 311-334, 2005.

[40] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer [Computer program]". Accessed 11/2010 at http://www.praat.org/ 2006.

[41] Bruce, G. and Hermans, B., "Word tones in Germanic languages" in H. van der Hulst [Ed] Word Prosodic Systems in the Languages of Europe, 605-658, Mouton de Gruter, 1999.

[42] Egerod, S., "Phonation Types in Chinese and South East Asian Languages", Acta Linguistica Hafniensia 13:159-171, 1971.

[43] Abramson, A.S., Thongkum, T.L. and Nye, P.W., "Voice Register in Suai (Kuai): An Analysis of Perceptual and Acoustic Data", Phonetica 61:147-171, 2004.

[44] Michaud, A., "Final Consonants and Glottalization: New Perspectives from Hanoi Vietnamese", Phonetica 61:119-146, 2004.

[45] Ladefoged, P. and Maddieson, I., The sounds of the world's languages, Blackwells, 1996.

[46] Silverman, D., "Laryngeal complexity in Otomanguean vowels", Phonology 14:235-261, 1997.

[47] Bertelsen, H., Danske Grammatikere IV, Gyldendal, 1920. Reprinted by Det Danske Sprog og Litteraturselskab, C.A. Reitzels Boghandel A/S, 1979.

[48] Bertelsen, H., Danske Grammatikere V, Gyldendal 1923. Reprinted by Det Danske Sprog og Litteraturselskab, C.A. Reitzels Boghandel A/S, 1979.

Intonational Phonology of Cuban Spanish: A preliminary model

Ann Aly Bailey

Department of Linguistics, University of California, Los Angeles, United States

aabailey@ucla.edu

Abstract

The present study proposes a preliminary model of intonational phonology for Cuban Spanish in the framework of Autosegmental-Metrical phonology. Data from controlled and semi-spontaneous speech were used to establish the boundary tones and pitch accents which are contrastive in this variety of Spanish. It was found that Cuban Spanish shares various tonal categories with both the Pan Spanish ToBI (Tones and Break Indices) [1] and other Caribbean Island Spanish dialects (Puerto Rican [2] and Dominican [3]), but differ from these dialects in how those pitch accents and boundary tones are used to convey meaning. Cuban Spanish shares its primary prenuclear pitch accents and nuclear contours for imperative statement and narrow focus with the Pan Sp_ToBI, but shares the nuclear contours for broad focus, vocative, and wh-questions with Puerto Rican Spanish. Similar to the other Caribbean Island Spanish varieties, the Cuban Spanish boundary tone inventory consists of a subset of the attested boundary tones found in the Pan Sp_ToBI, and all three Caribbean varieties share low boundary tones in non-wh questions, a marker of Caribbean Spanish speech.

Index Terms: Intonation, Cuban-Spanish, Sp-ToBI

1. Introduction

This study study proposes a preliminary model of intonational phonology of Cuban Spanish within the framework of the Autosegmental-Metrical (AM) model, and the conventions of Cuban Spanish ToBI (Tones and Break Indices), adopting the labeling conventions of the current Pan Spanish Tones and Break Indices (Pan Sp-ToBI). The first description and proposal for Pan Sp_ToBI was developed in [4] and later revised and expanded in [1] to include additional phrase accents and boundary tones. Although the aforementioned studies have developed a transcription system that accounts for distinctive intonational features of the Spanish language in general, not all dialects and varieties of Spanish use these features in the same way. Therefore, more recently, descriptions of Spanish intonation and the transcription systems from ten different varieties have been developed (described in volumes such as [5]), including Puerto Rican and Dominican Spanish, two of the three Caribbean Island Spanish varieties. However, as established by [2] and [3], the intonation of the Caribbean dialects also differs in various ways, despite their geographic proximity. The investigation of the intonational phonology of Cuban Spanish, a Caribbean island dialect that is currently undocumented within the AM framework, will allow for additional comparisons amongst the Caribbean dialects and other varieties of Spanish in general.

2. Methodology

The strained political relations that Cuba has with other countries have made linguistic data on this dialect difficult to obtain. However, the large Cuban community in South Florida provides a unique linguistic context in which data collection is possible. This study presents both controlled speech from a reading task and semi-spontaneous speech from a Discourse Completion Task (DCT) modeled from [5] to allow for comparison. The speech of eight speakers (four Cuban-born, four Miami-born) was consulted for the present analysis. All consultants were either born in or currently live in South Florida and were a median age of 42.3 years old. The data were collected on an LS-11 portable recording device at a 16-bit rate at 44.1 kHz and analyzed in *Praat*, version 5.3.60 [6].

3. Intonational Phonology of Cuban Spanish

This section will present the prosodic structure of Cuban Spanish, defined by intonation (section 3.1), and the tonal inventory, with boundary tones and pitch accents in sections 3.2 and 3.3, respectively.

3.1 Prosodic Structure of Spanish

Like other varieties of Spanish, Cuban Spanish has a lexical stress system with stress usually occurring on the penultimate syllable of content words. Above the word level, there is evidence for Intermediate Phrases (ip) and Intonation Phrases (IP), although no evidence for tonal stacking at the end of an Intonation Phrases has been found [1]. An IP is defined by a boundary tone at its right edge, with lengthening of the IP-final syllable, and an optional pause after the IP. An ip is defined by a phrase accent at its right edge, with slight lengthening of the ip-final syllable, and a pitch reset starting on the word after the ip. An ip has at least one pitch accent which is realized on the stressed syllable of most content words. As in English, the last pitch accent in an ip is the most prominent, called a nuclear pitch accent.

3.2 Boundary tones

Similar to the boundary tone inventories established in the Pan Sp_ToBI, Cuban Spanish also contains both monotonal (H (High), M (Mid), L (Low) and bitonal (LH (rising), HL (falling)) boundary tones. Similar to the other Caribbean Island varieties [2, 3], Cuban Spanish contains only a subset of the monotonal and bitonal boundary tones established in the Pan Sp_ToBI and does not contain tritonal boundary tones.

3.2.1 Monotonal boundary tones

Three monotonal boundary tones were found to mark the right edge of both ip and IP: L, M and H. The low boundary tone marking the end of IP (L%) is found primarily in declarative statements, imperatives, exclamatives, wh-questions, and requests (for Cuban-born speakers only). The L% can be realized either as a falling or low plateau tone, depending on the nuclear pitch accent. A low boundary tone marking an ip (L-) is also used in this dialect, usually to connect sections of a longer utterance, such as a declarative with a tag phrase or utterances with relative clauses, suggesting a syntactic dependency that is sensitive to these phrase edges.

High boundary tones marking the end of an IP (H%) were used in various types of questions, such as yes-no questions, echo questions, requests, tag questions, and less frequently, for a subset of wh-questions. Similar to the L% boundary tone, H% can be realized either with a rising F0 (from the nuclear pitch accent) or as a high plateau continued from a high nuclear pitch accent. High ip boundary tones (H-) were used by speakers in listing contexts as well as to signal continuation in prolonged speech.

Mid boundary tones marking the end of IP (M%) were less frequent than L% or H%, and were seen in vocatives, polite questions, and more emphatic exclamative statements. M% boundary tones in this dialect are typically realized with a slightly falling F0 from a high nuclear pitch accent, or less commonly, a slightly rising F0 from a low nuclear pitch accent. Mid ip boundary tones (M-) were also used to signal continuation in prolonged speech, but was typically seen in faster, more disfluent speech as opposed to H-, which signaled continuation in more careful speech.

Figure 1 shows L%, M%, and H% boundary tones, respectively. The figures in this paper contain the following text grid tiers, from top to bottom: (1) orthography, (2) stressed syllable of content words, (3) tones, (4) breaks, and (5) English gloss.

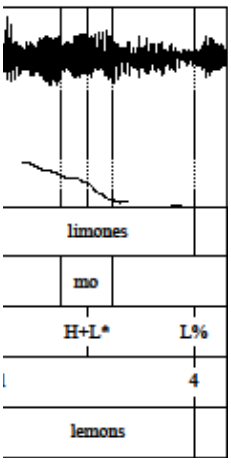


Figure 1a. L% boundary tone.

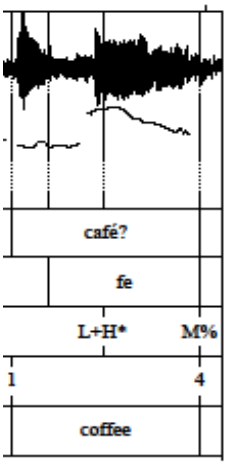


Figure 1b. M% boundary tone

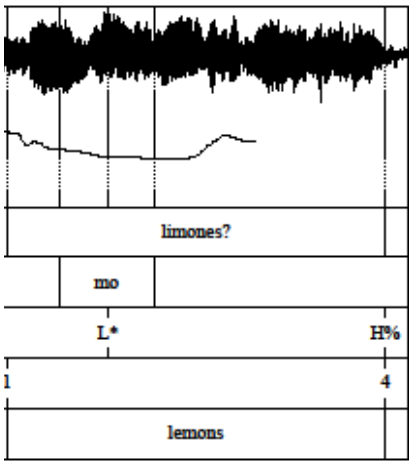
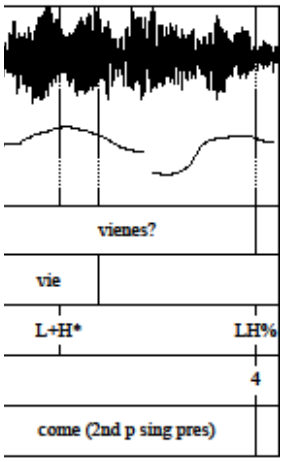


Figure 1c. H% boundary tone

3.2.2 Bitonal boundary tones

Two bitonal boundary tones were found in this dialect, LH and HL. These occurred less frequently than their monotonal counterparts. LH% IP-boundary tones were found in disjunctive wh-questions, sarcastic/rhetorical contexts, and non-wh exclamative statements. LH is realized with a falling (or sustained) F0 after the nuclear pitch accent and followed with a rising F0 in the same syllable that may or may not be as high as a previous H tone in the same ip.

The falling bitonal boundary tone (HL) is seen at the end of an IP (HL%) in exhortative imperatives (both in statement and question form). This boundary tone is not seen at ip boundaries in the data collected. HL% is realized with a rising or sustained F0 after the nuclear pitch accent which falls within the same syllable. Figure 2 shows examples of both bitonal boundary tones at the end of IP, following L+H* nuclear pitch accents. Figure 2a shows a LH% boundary tone occurring on a word with penultimate stress and 2b shows a HL% realized on the only, thus stressed, syllable of the final word (*qué*). Here, in order to accommodate two H targets (H of L+H* pitch accent and H of HL% boundary tone), the syllable is substantially lengthened in Fig.2b.



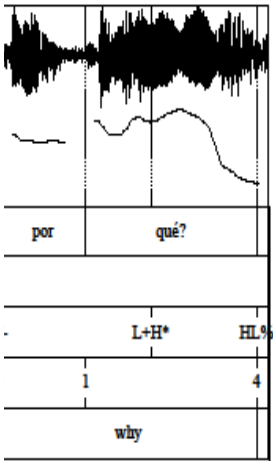


Figure 2. LH% boundary tone (2a, left) and HL% boundary tone (2b, right).

3.3 Pitch accents

This section will introduce the tonal inventory of pre-nuclear pitch accents (sec.3.3.1) and nuclear pitch accents (sec.3.3.2). Section 3.3.3 will discuss how narrow focus and stress clash affect the realization of pitch accents in Cuban Spanish.

3.3.1 Pre-nuclear pitch accents

As in Pan Spanish, three rising, bitonal nuclear pitch accents were seen in the data collected: one in which the stressed syllable is aligned with the pitch trough before rising (L*+H); one in which the stressed syllable is aligned with the peak of the pitch accent after a low F0 on the preceding syllable (L+H*); and lastly, a pitch accent in which rising to a delayed peak (realized on the next syllable) is seen in the stressed syllable (L+<H*) after a low F0 on the preceding syllable. The most frequent prenuclear pitch accent differed by task types, suggesting an effect for speech formality or style. In the

controlled task, L+<H* accounted for 53% of all pre-nuclear pitch accents and was the most common pitch accent before ip boundaries, occurring 78% of time. However, this pitch accent was found more often in semi-spontaneous speech: L*+H occurred in over 90% of the semi-spontaneous data but only 30% of the controlled data. L+<H* pre-nuclear pitch accents are mainly used in emphatic semi-spontaneous speech, further supporting the different styles associated with these pitch accents. The least common pre-nuclear pitch accent was L+H*, which only occurred in 17% of the data. Although uncommon, L+H* was the pitch accent used in over 90% of words with final stress in both tasks.

3.3.2 Nuclear configurations

Only two possible nuclear pitch accents were found in declaratives: H+L* and L+H*. H+L* was the most common nuclear pitch accent in declarative utterances, as L+H* was typically used only by Cuban-born speakers. An example is shown in Figure 3.

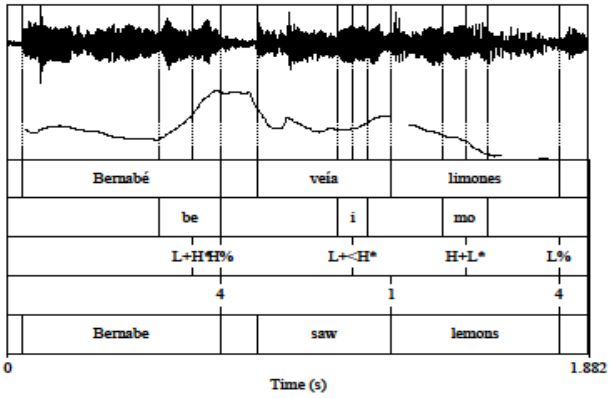


Figure 3. Declarative statement with H+L* L% nuclear configuration.

Unlike H+L*, which was the most common nuclear pitch accent only for declaratives (66% of analyzed cases) in Cuban Spanish, L+H* was the most common nuclear pitch accent and was used in various types of utterances. Besides a possible nuclear pitch accent for declaratives, L+H* was also used as the most common nuclear pitch accent for exclamatives (81%), imperatives, vocatives (79%), yes/no questions and requests (70%), and echo questions (76%). Although these discourse categories share the same nuclear pitch accent, they differ from each other by boundary tone or pitch scale. L+H* L% nuclear configurations are used for exclamatives (81%), commands (51%), and as an option for yes/no questions, mostly by Cuban-born speakers (45%)**;** L+H* H% is used in the tag phrase of tag questions and as the more common variant for yes/no questions, used by both speaker groups (55%), echo questions (62%), and a variant of wh-questions used more frequently by Cuban-American speakers (45%); finally, L+H* M% is the preferred nuclear configuration for vocatives (80%) and is used as an option for polite yes/no questions and requests, mostly by Cuban-born speakers (10%). Questions, especially polite questions, tend to involve upstepped high tones, creating a larger pitch range than a similar (L+)H* in non-questions. This is shown in Figure 4.

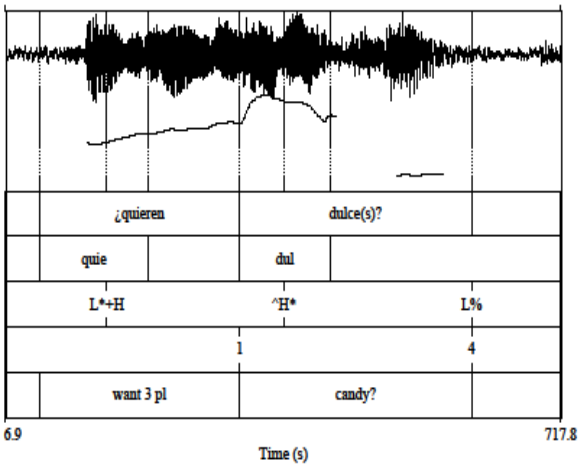


Figure 4: Yes/No Question with upstepped (^) (L+) H* L% nuclear contour

Monotonal nuclear pitch accents are less common in Cuban Spanish and are typically variants of rising pitch accents that are undershot due to stress location and position (e.g. L+H* H% may be realized as L* H% when stress is final). Figure 5 shows a tag question in which a monosyllabic tag phrase is realized as L* H% instead of L+H* H%.

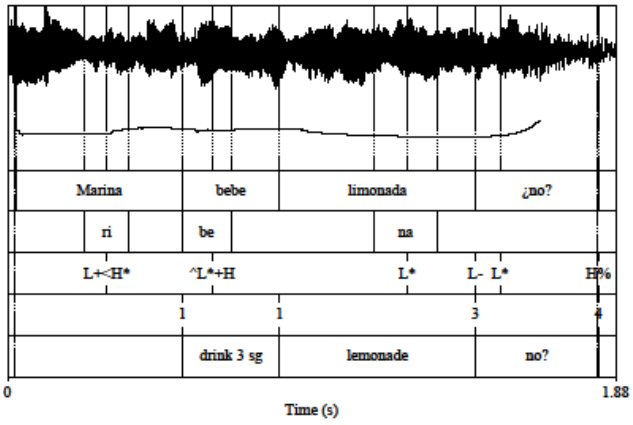


Figure 5.Tag question with L(+H*) L% tag phrase

3.3.3 Stress clash and focus

Data from both the controlled and semi-spontaneous tasks revealed that when two stressed syllables are adjacent and cause a stress clash, part of a pitch accent (in the case of a bitonal pitch accent) or an entire pitch accent may not be realized. In the case of adjacent rising prenuclear pitch accents, the most common stress clash is the result of a word with penultimate or final stress followed by a word whose first syllable is stressed. These clashes are resolved in one of two ways: either a phrase break is inserted between the two words and the pitch is reset on the second word (typical in slower, more formal speech) or, when the two are in the same ip, a portion of the second (bitonal) pitch accent is not realized (more common in faster, less formal speech). An example of this latter resolution is shown in Figure 5.

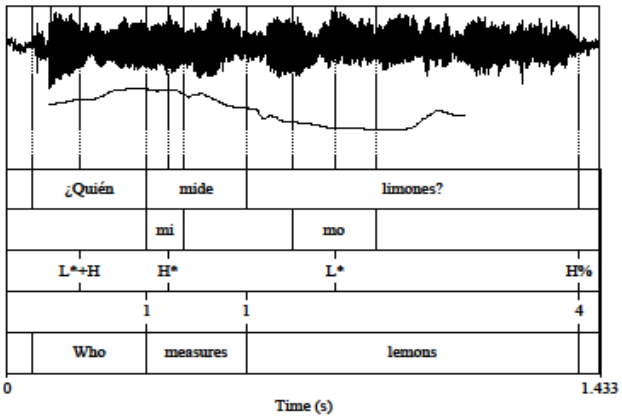


Figure 5. Stress clash between *quién* and *mid*e

When the L*+H nuclear pitch accent on a penultimate syllable is followed by a H% boundary tone, it was not clear whether the H part of the pitch accent is realized or not because both the H trailing tone and the H boundary tone occur on the IP-final syllable. In this case, only L* is labeled on the stressed syllable.

The basic tonal pattern of sentences found in the neutral focus context changed in several ways when speakers were asked to focus lexical items. The most typical focus realization included an overall expanded pitch range, an L+H* pitch accent on the focused item in nearly 100% of cases, and an IP boundary before or after the focused word (depending on the location of the focused items). When longer words with penultimate stress were focused, such as *limonada*, 'lemonade', two kinds of secondary stress effects were found: either primary stress was shifted to the first syllable instead of being realized on the penult, or both the initial and penultimate syllables were both realized with L+H* pitch accents. This latter pattern is shown in Figure 6.

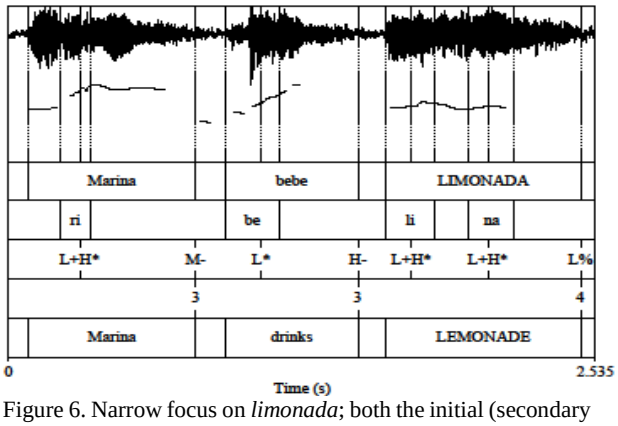


Figure 6. Narrow focus on *limonada*; both the initial (secondary stress) and penultimate syllable (primary stress) are prominent.

4. Discussion

The data analyzed revealed several within-dialect and cross-dialectal patterns with respect to intonation. Within the Cuban Spanish speakers, Cuban-born speakers produced more nuclear configurations per discourse category than Cuban-American speakers, whose realizations represented a subset of the Cuban-born speakers. These differences could have various sources; as the Cuban-American speakers are heritage speakers of Spanish who are also native speakers of American English, their contact with English could have an influence on their Spanish. Additionally, as South Florida has a diverse Hispanic community apart from the Cuban majority, dialect leveling (as seen with Puerto Rican Spanish in New York [7]) with other varieties of Spanish may have also influenced the non-Cuban-born Spanish speakers toward certain prosodic realizations. Data from other dialects of Spanish present in South Florida as well as English data from these participants is needed in order to investigate the source of this variation.

When the intonation data from the Cuban Spanish speakers is compared with that of the Pan Sp_ToBI and other Caribbean Spanish dialects, few similarities emerge. Cuban Spanish shares its prenuclear pitch accent and the tonal configurations for exclamatives and narrow focus with the Pan Sp_ToBI, but shares the tonal contours of categories such as broad focus, vocatives, and polite questions with Puerto Rican Spanish. The boundary tone inventory found in Cuban Spanish resembles the other Caribbean Island varieties more than the Pan Sp_ToBI; the Caribbean Island dialects (including Cuban) contain only a subset of the monotonal and bitonal boundary tones specified in the Pan Sp_ToBI and do not contain any tritonal boundary tones.

The use of L% in questions found in Cuban Spanish was not reported for Pan Spanish, but occurs in the other two Caribbean Island Spanish dialects [2, 3]. That is, all three Caribbean island varieties share L% boundary tones for non-wh questions, which is a marker of Caribbean Spanish dialects. Among the Caribbean dialects mentioned, however, Cuban Spanish was closer to Puerto Rican Spanish than to Dominican Spanish by employing M% boundary tones in vocatives, HL% for exhortative imperatives, and H% boundary tones in tag questions and some requests (mostly by Cuban-American speakers), but Cuban and Puerto Rican varieties still differ in the usage of their tonal inventory across discourse categories. As previously mentioned, the relative political isolation of Cuba as well as the dominance of the Cuban dialect in South Florida may be a source of this divergence with respect to the other dialects of geographical proximity.

5. Conclusion

The current study presents a preliminary model of intonational phonology of Cuban Spanish in the AM framework based on the intonation data from eight Cuban Spanish speakers living in South Florida. Using the conventions of the Pan Sp_ToBI, the tonal inventory of this dialect was proposed as well as the tonal contours that define various discourse categories and sentence types. The proposed tonal inventory was then compared with that of Pan Sp-ToBI and other varieties of Caribeean Spanish. Further analysis will consider additional discourse categories present in the semi-spontaneous data of the speakers as well as sources for within dialect variation.

6. Acknowledgments

I am thankful to the UCLA Graduate Summer Research Mentorship, which made data collection possible, and to my advisor, Sun-Ah Jun, for her help and support, my Research Assistant Ulysses Cázares for his invaluable help, and my colleague Mariska Bolyanatz for her help and input.

7. References

[1] Estebas Vilaplana, E. and Prieto, P., "La notación prosódica del español: Una revisión del Sp_ToBI". Estudios de Fonética Experimental, 17: 265-283, 2008.

[2] Armstrong, M., "Puerto Rican Spanish intonation" in P. Prieto and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, 155-189, Lincom Europa: München, 2010.

[3] Willis, E., "Dominican Spanish Intonation", in P. Prieto and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, 125-153, Lincom Europa: München, 2010.

[4] Beckman, M., Díaz-Campos, M., McGory, J., and Morgan, T., "Intonation across Spanish, in the Tones and Breaks Indices Framework, Probus 14: 9-36, 2002.

[5] Prieto, P. and Roseano, P. [Eds], "Transcription of Intonation of the Spanish Language", Lincom Europa: München, 2010.

[6] Boersma, P. and Weenink, D., "Praat: Doing phonetics by computer" [computer program], version 5.3.60, retrieved from www.praat.org, 2013.

[7] Otheguy, R. and Zentella, A., "Spanish in New York". Oxford University Press: New York, 2012.

Modeling of a rise-fall intonation pattern in the language of young Paris speakers

Roberto Paternostro [1], *Jean-Philippe Goldman*[2]

[1] Université Paris Ouest – MoDyCo, France & Università di Brescia, Italy
[2] University of Geneva, Switzerland
paternostro.roberto@gmail.com Jean-Philippe.Goldman@unige.ch

Abstract

Intonation seems to be one of the major cues for identifying youth language in the Paris region. As part of a large-scale corpus-based analysis, this paper attempts to model a rise-fall final prosodic pattern, considered to be representative of a Paris working-class suburbs accent. Comparison with the emphatic rise-fall prosodic pattern, well-known in general French, will provide the opportunity for sociolinguistic insights. The ethnic hypothesis is dismissed in favor of a context-bound and interaction-sensitive interpretation.

Index Terms: socio-phonetics, prosodic patterns, linguistic variation and change, acoustic modeling.

1. Introduction

Among other linguistic elements, such as lexical items and segmental features[1], intonation is pointed out as one of the major cues for the identification of Paris youth language. Various studies suggest that the realization of a rise-fall final prosodic pattern, characterized by a strong pitch rise and a sharp fall in the fundamental frequency (henceforth F_0) that is particularly ample and with a possible lengthening of the penultimate syllable, is the main factor in the perception of a *banlieue*[2] accent [1]. Furthermore, such an accent is often associated with young speakers living in working-class areas of the city and coming from multicultural environments, especially of North or Sub-Saharan African origin [2], [20]. However, Paternostro [3] has shown that even though speakers in direct contact with multicultural environments seem to be the leaders of the spread of rise-fall final prosodic patterns and that their peers in indirect contact with multicultural environments seem rather to adopt and diffuse these phonetic variants, the two groups cannot be considered as two different populations. In fact they represent a methodological "artifact" which tends to separate into different populations speakers who are part of the same community of practice, sharing the same social experience and the same identity values. The ethnic hypothesis is thus dismissed and the dynamics of variation must be rather sought in a set of processes of adaptation to a feeling of communicational proximity [4]. Comparison with emphatic rise-fall prosodic patterns, well-known in general French, shows that *banlieue* intonation contours seem to convey emphasis[3] as well and are likely to express speakers' involvement in interaction. This paper reports on a large-scale study attempting to model the acoustic features of this intonation contour, on the basis of the 'Multicultural Paris French' (MPF) corpus currently being collected in the Paris region [5].

2. Background

Conein & Gadet [6] noticed the spread of a particular 'strong' prosodic pattern in the speech of young Paris speakers, characterized by a large melodic movement and a lengthening of the penultimate syllable. However, this pattern seems to be more hereditary than innovative, since its features overlap with those of 'popular' French. Fagyal [1] focused on the prosodic realizations of middle-school students living in the northern working-class suburbs of Paris. She noticed significant lengthening of the penultimate syllable, marked by a high tonal target. Lehka & Le Gac [7] observed similar intonation contours among adolescents in the working-class suburbs of Rouen, particularly characterized by a sharp fall on the last prosodic unit. However, no significant penultimate lengthening was observed. Le Gac *et al.* [8] compared Paris and Rouen realizations and found that Paris speakers tended to produce rise-fall patterns with a rather reduced pitch range whereas the informants from Rouen produced both reduced and expanded patterns.

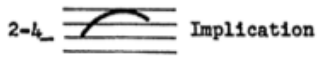


Figure 1: *Implication contour by Delattre*

Rise-fall intonation contours, such as the implication[4] contour described by Delattre [9] or the emphatic contour described by Di Cristo [10], are well-known in general French[5]. Lehka-Lemarchand [11] already noticed similar features between *banlieue* and emphatic contours. Even though, her assumption is not backed up by acoustic and statistical analyses, she concludes that the difference between the two contours is greater than their similarity and prefers to consider *banlieue* contours as a specific phenomenon linked to non-standard language.

[1] We refer, for example, to back-slang words such as *meuf* (back-slang for *femme*, 'woman') or *keuf* (back-slang for *flic*, 'policeman'). Affrication of dental stops before high closed vowels /i/ and /y/ (such as in *voiture* [vwatʃyʁ], 'car' or *dire* [dʒiʁ], 'to say') and a strong unvoiced fricative /ʁ/ seem to be among the most important segmental cues [4].
[2] Working-class suburbs of Paris.
[3] By "emphasis" we mean "involvement" of speakers in interaction, according to Selting [14].
[4] The link between implication and emphasis (or involvement) is not fortuitous. Just as emphasis, implication presupposes implicitness and complicity between speakers.
[5] Stewart [18] compares *banlieue* intonation contours to Delattre's commandment and exclamation ones [9], which are clearly falling contours. However, Lekha & Le Gac [7] and Lekha-Lemarchand [11] argue that in the perception of *banlieue* contours the rise matters more than fall.

Banlieue contours	Emphatic contours
Ample movements of F_0	Ample movements of F_0
Steep slope and sharp fall of F_0	Normal rise, sharp fall of F_0
Shorter	Longer
Early alignment (penultimate)	Late alignment (last syllable)

Table 1: *Comparison between 'banlieue' and emphatic contours in the literature*

The link between emphatic contours and *banlieue* accent is not fortuitous. On the one hand, Stewart & Fagyal [12] suggest that *banlieue* intonation contours often convey aggressivity and are collectively perceived as quarrelsome, even when occurring in a neutral context. On the other hand, Paternostro [13] shows that the *banlieue* accent is highly related to the expression of communicational proximity and speakers' involvement in interaction and can be seen as the actualization of an emphatic speech style[1]. According to Selting [14], an emphatic speech style results from communication that is strongly marked by 'emotion' and displays a shared 'involvement' of the interactants. Prosodic cues seem to play a major role, since they are used as a "device to evoke context-sensitive interpretations of emphatic 'peaks of involvement'". Ample melodic movements of F_0, associated with increased duration and intensity, are the prosodic markers mainly used by an emphatic speech style in French [15].

3. Modeling *banlieue* contours

Our corpus consists of 593 occurrences of rise-fall intonation contours, taken from 3h05m of 6 teenagers' sample speech. The informants are 2 girls (Ana and Juline) and 4 boys (Koffi , Aziz, Hakim and Walid), living in the Paris *banlieue*. Interviews were carried out within the framework of the MPF project[2], with respect to the ecology of interactions. Informants were not selected according to a socio-demographic categorization, but rather according to an acquaintance network. Since investigators and informants were not strangers to one another, speech circulates freely, long stories emerge, and mutual speaker involvement takes place [16]. 391 contours were labeled "emphatic" and 202 "*banlieue*" by 72 Paris students, during a perception test, details of which are given below.

3.1 Perception test

The perception test consisted of two parts[3]. The first part aimed at determining whether rise-fall intonation contours are associated with the expression of emphasis. Our 72 judges were asked to listen to 50 filtered speech samples containing standard rise and/or fall contours (continuation and final intonation contours) or rise-fall contours considered to convey emphasis. Judges were asked to qualify each stimulus according to a grid of bipolar adjectives. They also had to evaluate the level of emphasis conveyed on a four-point scale (0, 1, 2, 3). Results show that standard continuation and/or final intonation contours were referred to as "neutral" and "calm" whereas rise-fall contours were referred to as "emphatic" and "angry". Rise-fall contours also conveyed more emphasis than standard ones: 1.40 (out of 3) vs. 0.84. Results are statistically significant[4].

The second part of the test aimed at determining whether judges[5] were able to distinguish between emphatic contours (henceforth *EM*) and *banlieue* contours (henceforth *BA*) and at evaluating the level of emphasis conveyed. It will also help us for corpus annotation and show whether *banlieue* contours convey emphasis or not. Judges listened to 100[6] short speech samples and ticked the answer in a reply form. The level of emphasis was evaluated on a four-point scale (0-3). Results show that judges were able to distinguish EM contours from BA contours in 59% of cases[7]. BA contours seem to convey more emphasis than EM: 1.67 vs. 1.48 (out of 3)[8]. These issues are discussed below.

3.2 Methods and materials

Rise-fall intonation contours were annotated manually under *Praat* according to 3 points: (1) beginning of the rise; (2) the top; (3) end of the fall (see: Fig. 2). Pitch values were verified and corrected manually, in order to avoid octave leaps and other detection errors.

Pitch values were extracted with their temporal position for each point. They were then converted to semi-tones (henceforth *st*), for the purpose of comparison between male and female voices. The last 3 syllables were also segmented to find out whether EM and BA contours align on the same syllable or not. The length for each syllable as well as for the complete contour (from point 1 to point 3, see: Fig. 2) was also calculated. Rise slope and fall slope were also estimated through the formula 'semi-tones per second' (henceforth st/s). Annotations and values were cross-checked several times, both manually and automatically.

[1] See also Fagyal & Stewart [19], as far as interactivity is concerned.

[2] MPF is the French part of the research project 'Multicultural London English – Multicultural Paris French' (www.mle-mpf.fr). It aims at investigating language variation and change occurring in Western European cities, in a situation of linguistic contact.

[3] Only details that are relevant for this paper will be discussed here.

[4] $F(1.46) = 10.04$ $p < 0.002$.

[5] The 72 judges were the same for both the first and the second part of the perception test.

[6] We were only able to test the judges' perception for 100 intonation contours, randomly selected out of 593. Testing the whole amount of data would have been impossible. The remaining intonation contours were labeled according to the annotator's perception.

[7] Results are statistically significant. Statistical significance was calculated using a chi-squared test for each stimulus. The limited space available here does not allow the results to be shown individually.

[8] Results are statistically significant: t 15.31 $p = 0.0001$ for EM, and t 15.54 $p = 0.0001$ for BA. Otherwise, interaction does not reach the significance level: $F(1.58) = 2.637$ $p < 0.10$.

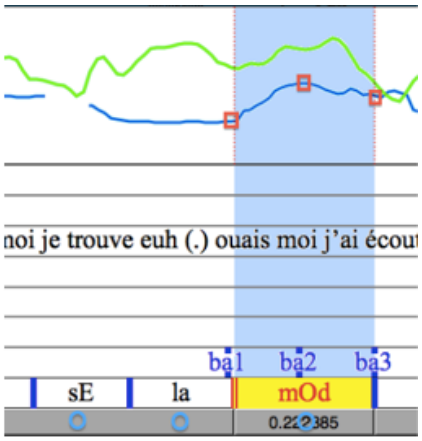


Figure 2: *Sample of 3-point annotation for a BA contour. Red squares, from left to right, correspond to the beginning of the rise (point 1 = ba1), the top (point 2 = ba2) and the end of the fall (point 3 = ba3). Blue circles, from right to left, correspond to the last syllable (syll 0, 'mOd'), the penultimate (syll 1, 'la') and the ante-penultimate (syll 2, 'sE').*

3.3 Analyses and results

3.3.1 Rise and fall

Results concerning rise and fall range show that BA contours rise and fall slightly more than EM contours (fig. 4). Results are statistically significant[1]. We will discuss below whether such a tonal difference is likely to be perceived or not.

Rise / Fall		S. dev.	Difference
EM rise	4.621	2.306	0.947
BA rise	5.568	2.592	
EM fall	-4.030	-1.955	-0.711
BA fall	-4.741	-2.420	

Table 2: *EM and BA contours' rise/fall range*

3.3.2 Duration and penultimate lengthening

As far as duration is concerned, there is no significant difference between the global duration of EM and BA contours (point 3 - point 1, on fig. 2)[2]. However, the fall of the EM contours is significantly longer (0.036 s) than that of BA.

Global duration		S. dev.	Difference
EM	0.240 s	0.075	0.002
BA	0.233 s	0.077	

Table 3: *EM and BA contours' global duration*

We also tried to determine whether the penultimate syllable is significantly lengthened compared to the last one. The French language usually has a fixed final stress on the last syllable of the prosodic group, which is mainly expressed by duration. Final stressed syllables are twice as long as unstressed syllables. Our results show that while the EM penultimate seems to be slightly shorter (-18%) than that of BA, there is no significant lengthening of the penultimate syllable.

3.3.3 Pitch alignment according to syllable

Results concerning intonation contour alignment show that 51% of EM contours align on the last syllable (syll 0) whereas 64% of BA contours align on the penultimate (syll 1). Only a small percentage of both contours align on the antepenultimate (syll 2), as can be seen below (fig. 3). Results are statistically significant ($p < 0.01$).

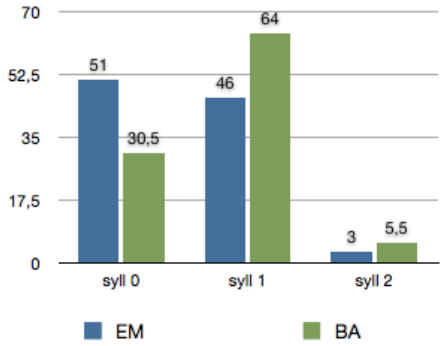


Figure 3: *EM and BA contour alignment according to syllables*

Nevertheless, interpretation of the results is not straightforward. BA contours show a general tendency to align earlier than EM ones. EM contours start to rise 18% later[3] than BA. EM contours also reach the high pitch target 5.8% later[4] and fall 3% later[5] than BA contours (see fig. 4 below). Results are statistically significant, $p < 0.01$.

Finally, we found that BA contours rise faster (10.5 st/s) and fall faster (12.7 st/s) than EM contours ($p < 0.001$).

3.3.4 Summary

To summarize, BA intonation contours seem to rise higher and fall lower than EM. BA contours are slightly shorter than EM, especially as far as final syllables are concerned. BA penultimate syllables are actually a bit longer than EM. However, BA and EM penultimate syllables are never significantly longer than final ones. BA contours align earlier and rise faster than EM.

These results may indicate that there is a difference between BA and EM intonation contours. The question now is whether this difference is marked enough to be perceived.

4. Discussion

Our results concerning the BA contours' rise and fall range show lower values than those found by Lehka-Lemarchand [11] in Rouen suburbs (9.6 st for rise and -8.6 st for fall). No quantitative values can be found in the literature for Paris, apart from the study comparing Paris and Rouen [8]. Paris

[1] EM vs. BA rise: $F(1.592) = 22.21$ $p < 0.0001$. EM vs. BA fall: $F(1.592) = 15.71$ $p < 0.0001$.
[2] Given the large amount of data, values have not been normalized.
[3] Point 1, on fig. 2.
[4] Point 2, on fig. 2.
[5] Point 3, on fig. 2.

speakers do not seem to realize very strong BA contours, compared to the rise-fall configurations found in Rouen.
As far as alignment is concerned, the beginning of the rising slope always aligns on the last syllable for BA contours in the Rouen corpus, while Fagyal [1] found that they always align on the penultimate in Paris. Our study showed that BA mainly aligns on the penultimate whereas EM mainly aligns on the last syllable. However, the percentage difference is not that high and does not reach the significance threshold. BA and EM do not differ much as far as alignment is concerned: 46% of EM contours actually align on the penultimate syllable, as do BA contours. In fact, further analyses have shown that globally both BA and EM intonation contours mainly align before the beginning of the last syllable (see fig. 7 below)
With regard to rise and fall speed, not only do BA rise higher and fall lower than EM, they also rise and fall faster. This probably induces the perception of stronger intonation contours, which is also strengthened by early alignment on the penultimate. Intensity has not been measured, but we hypothesize that it can play a significant role in intensifying the perception of 'stronger' high-low prosodic patterns.

In order to answer our question as to whether the acoustic difference between BA and EM contours can be perceived or not, the *glissando* threshold put forward by t'Hart *et al.* [17] can be taken into account. *Glissando* threshold refers to the variation of F_0 and defines when this range is perceived as a rise and/or fall tone. It is usually expressed in semi-tones per second and indicates the speed needed for the rise-fall pattern to be perceived. The authors indicate that it should be equal to or greater than 0.16 st/t^2. Our results show that the *glissando* threshold of rising BA contours is 0.83 st/t^2, which is 0.16 higher than the *glissando* threshold of rising EM ones (0.67 st/t^2). This result is significant enough for the BA contour rise to be perceived as different from the EM one ($p < 0.05$). On the contrary, the results are not significant as far as falling BA and EM contours are concerned.

Finally, amongst our numerous results, early alignment and the *glissando* threshold of BA intonation contours seem to be the two key elements leading to the perception of 'stronger' BA intonation contours.

5. Conclusions

Acoustic analysis has shown that there is a slight difference between *banlieue* and emphatic contours. BA intonation contours start to rise earlier, rise higher, faster, and fall lower than EM intonation contours. Nevertheless, their similarities seems to be greater than their differences, since the features of BA contours are defined in comparative terms.

Closer analyses of judges' perception showed that the stimuli evaluated at a 100% rate of agreement were all BA contours whereas stimuli evaluated at a < 30% rate of agreement were EM contours. The higher the rate of agreement, the 'stronger' the contour. The lower the rate of agreement, the 'weaker' the contour.

As can be seen below (fig. 4), the modeling of BA and EM does not in fact show two different kinds of rise-fall intonation contours. This suggests that BA contours are similar to EM but 'stronger'. Comparison with a control group of 119 standard continuation contours (CONT), annotated in the same corpus (fig. 4), reinforces this assumption in that BA and EM intonation contours are quite different from other contours in general French, but they are not substantially different from each other.

Acoustic differences alone are not sufficient in themselves to assert the existence of two different high-low prosodic patterns. Extra phonetic, linguistic and semiotic details are required for a *banlieue* accent to be perceived.

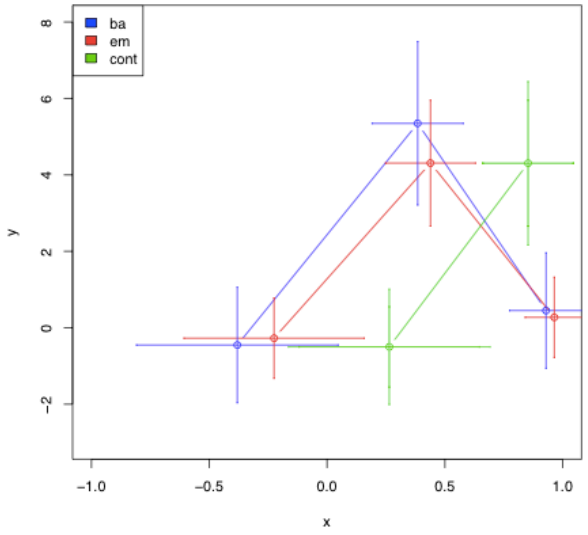


Figure 4: *Comparison between EM (red), BA (blue) and CONT (green) alignment with regard to final syllable (normalized between 0 to 1 on x-axis) and penultimate syllables (prior to 0.0 on x-axis). The y-axis represents the tonal difference (in semi-tones). The reference point of the pitch is set on the beginning of the rise. Crosses represent mean and standard deviation in temporal (x-axis) and tonal (y-axis) dimensions.*

The acoustic data seem then to match sociolinguistic insights: the high-low prosodic pattern cannot be considered a new phenomenon and is far from being specific to young people coming from working-class suburbs and/or from immigration. The novelty resides in the fact that EM/BA contours are likely to be used in context for pragmatic purposes, according to speakers' mutual involvement in interaction and communicational proximity. "*La vie dans la cite, c'est chaleureux*"[1], said one of our informants. That is why we suggest that the two kinds of intonation contours, emphatic and *banlieue*, can be considered as a single phenomenon, situated on a continuum, and that they are context-bound and interaction-sensitive. Thus, the realization of such a prosodic pattern ranges from one extreme to the other, depending on the speakers' shared understanding of the communicative situation and the speakers' level of interactivity.

6. References

[1] Fagyal, Z., Accent de banlieue. Aspects prosodiques du français populaire en contact avec les langues de l'immigration, L'Harmattan, 2010.

[2] Fagyal, Z., "Prosodic consequences of being a *Beur*: French in Contact with Immigrant Languages in Paris", Selected papers from NWAV 32, Philadelphia, 2004, Working Papers in Linguistics 10(2): 91-104, 2005.

[1] Life in working-class suburbs is warm and friendly.

[3] Paternostro, R., “Aspects phonétiques de l’« accent parisien multiculturel » : innovation, créativité, métissage(s)”, Cahiers de l’AFLS 17(2): 32-54, 2012.
[4] Guerin, E. and Paternostro, R., “Characterising the ‘langue des jeunes’: Paris youth language in question”, in H., Tyne, V., André, A., Boulton, C., Benzitoun, C. and Y., Greub [Eds], Ecological and Data-Driven Perspectives in French Language Studies, Cambridge Scholars Publishing, in press.
[5] Gadet, F., “Collecting a new corpus in the Paris area: intertwining methodological and sociolinguistic reflexions”, in M. Jones and D. Hornsby [Eds], Language and Social Structure in Urban France, Legenda, in press.
[6] Conein, B. and Gadet, F., “Le ‘Français populaire’ des jeunes de la banlieue parisienne entre permanence et innovation”, in J., Androutsopoulos and A., Scholz [Eds] Actes du colloque de Heidelberg. Jugendsprache / Langue des jeunes / Youth language, Peter Lang, 105-123, 1998.
[7] Lehka, I. and Le Gac, D., “Identification d'un marqueur prosodique de l'accent de banlieue : le cas d'une banlieue rouennaise”, Actes du colloque MIDL 2004, Paris, 29-30 November 2004, 145-150.
[8] Le Gac, D., Jamin, M. and Lehka I., “A preliminary study of prosodic patterns in two varieties of suburban youth speech in France”, Proceedings of 3rd International Conference on Speech Prosody SP2006, Dresden May 2006.
[9] Delattre, P., “Les dix intonations de base du français”. The French Review, 40(1): 1-14, 1966.
[10] Di Cristo, A., “Intonation in French”, in A., Di Cristo and D. J., Hirst [Eds] Intonation systems : a survey of twenty languages, Cambridge University Press, 88-103, 1998.
[11] Lehka-Lemarchand, I., Accent de banlieue. Approche phonétique et sociolinguistique de la prosodie des jeunes d’une banlieue rouennaise. PhD dissertation, Université de Rouen, 2007.
[12] Stewart, C. and Fagyal, Z., “Engueulade ou énumération ? Attitudes envers quelques énoncés enregistrés dans les ‘banlieues’”, in M.-M., Bertucci and V., Houdart-Merot [Eds] Situations de banlieues : enseignement, langues, cultures, Institut National de Recherche, 241-252, 2005.
[13] Paternostro, R., “La « langue des jeunes » Parisiens : une forme actualisée dans la « proximité » ? Aspects phonétiques et questions méthodologiques”, Cahiers de Recherche de l’Ecole Doctorale en Linguistique Française 7: 9-19, 2013.
[14] Selting, M., “Emphatic speech style : with special focus on the prosodic signalling of heightened emotive involvement in conversation”, Journal of Pragmatics 22: 375-408, 1994.
[15] Bagou, O., “Validation perceptive et réalisations acoustiques de l’implication emphatique dans la narration orale spontanée”, Cahiers de linguistique française 23: 39-59, 2001.
[16] Gadet, F. and Guerin, E., “Des données pour étudier la variation : petits gestes méthodologiques, gros effets”, Cahiers de linguistique 38(1): 41-65, 2012.
[17] t’Hart, J., Collier, R. and Cohen, A., “A perceptual study of intonation. An experimental-phonetic approach to speech melody”, Cambridge University Press, 1990.
[18] Stewart, C., “On the Socio-Indexicality of a Parisian French Intonation Contour”, Journal of French Language Studies 22(02): 251 – 271, 2012.
[19] Fagyal, Z. and Stewart, C., "Prosodic style-shifting in preadolescent peer-group interactions in a working-class suburb of Paris", in F., Kern and M., Selting [Eds], Ethnic Styles of Speaking in European Metropolitan Areas. John Benjamins, 2011, 75–99.
[20] P. Boula de Mareüil, P. and Lehka-Lemarchand, I., “Can a prosodic pattern induce/reduce the perception of a lower-class suburban accent in French?”, 17th International Congress of Phonetic Sciences, 348–351, 2011.

Topic and Focus Intonation in Argentinean *Porteño*

David Le Gac[1]

[1] Laboratoire DySoLa, Université de Rouen, Normandie Univ, France
david.legac@univ-rouen.fr

Abstract

This paper investigates the intonation of topics and focus in Argentinean *Porteño*. We have found that whereas tonal alignment is phonetically conditioned, pitch height and duration constitute the main cues to express various types of focus in declarative and interrogative sentences; at least four intonational categories seem to be used by our speakers and the relevance of a register feature is discussed. As for topics, they are marked by special tunes that depend on the type of sentence; in particular, the topic tune in questions is the opposite of those found in declaratives.

Index Terms: *Porteño* Spanish, focus, topic, question, tonal alignment, pitch height, register feature, duration

1. Introduction

It is now well established that, in many languages, intonation plays a major role in the expression of the two main components of the information structure: topic and focus ([1], [2]). In English ([3]), French ([4]–[8]), Modern Greek ([8], [9]) and Spanish ([10]–[13]), among other languages, focus affects the intonation of the whole sentence: it assigns the nuclear pitch accent to the focus word and all the post-focal words are de-accented. Previous studies have also shown that 'informative' focus and contrastive focus are given different prominence ([1], [5], [14]). Furthermore, focusing has been found to lengthen the duration of the focus word and to shorten pre- and post-focal parts ([5], [9]). As for topics, they generally appear at the beginning of a sentence and express the 'aboutness' of a sentence, that is, "what is being talked about", and serve as the "the point of departure of the message" (see [15]). They are often described as implying a continuation rise at the end of the topicalized phrase in declarative sentences as in French ([4], [7], [6], [8]) and Modern Greek ([8], [9], [16]); in Spanish, [12] proposed that a phrasal H- indicates the end of the constituent conveying old information. In Greek, though, topics in questions have an inverted intonation pattern compared to that of the topics in declaratives: they display a L* instead of the H* and are right bounded by a low tone ([8], [9], [16]). This tonal inversion may also be produced in French ([4], [6], [8]), and has not yet been adequately explained.

Turning to the Argentinian Spanish, the intonation of the variety spoken in Buenos Aires, called *Porteño*, notably diverges from the Castilian and Latin American standards of Spanish. Pitch peak in pre-nuclear accents is described as being regularly located within the stressed syllable ([11], [17]–[19]) instead of being aligned with the post-tonic syllable as in the other varieties of Spanish. Broad focus declaratives differ from narrow focus ones in the alignment and height of the non-final pitch peaks ([19]–[21]), which happens earlier for the latter. However, in a pilot study investigating the perception of narrow focus in *Porteño*, [22] found that pitch alignment patterns cannot account for the distinction of narrow from broad focus in this dialect. Increasing pitch values and/or vowel duration, as well as de-accenting of material after a word in narrow focus, as opposed to the pitch accents present in broad focus contribute to a narrow focus perception. Focus thus deserves further study in order to determine its intonational properties in *Porteño*. As for topics, to the best of our knowledge, no study has yet explored their prosodic expression in this variety of Spanish.

This paper investigates the prosodic cues of topic and focus in both declarative and interrogative sentences, adopting the autosegmental-metrical theory of intonation ([23]). Regarding focus, we examine the melodic pattern and the duration of three types of narrow focused words: i) *'informative'* focused word (henceforth "IFoc") – i.e. the answering part of a WH-word; ii) *contrastive* focused word ("CFoc"); and iii) a focused word in a *yes-no question* ("QFoc"). These focused words are all located after the topics, in the middle of the sentence, and are compared to a pre-nuclear ("PNucl") non-focused word in the same position. We aim at determining which pitch accents and edge tones these words are associated with and to what extent tonal alignment, pitch height and duration contribute to distinguish them. Concerning topics, we will examine what type of pitch accent and what type of boundary tone they are associated with; in particular, we will determine whether their tonal pattern varies according to the type of sentence (declarative *vs*. question).

2. Methodology

In order to trigger the various kinds of focus, we used a question-response paradigm with a general lead-in context about two friends talking about a party. (1) and (2) below are two examples of these dialogues. Topics began the sentences and consisted of either one NP (a head noun and an adjective; cf. (2)) or two XPs (cf. (1)). The different focus words ("Q/C/IFoc" words) were proper nouns (*Penélope*, *Verónica*) in subject position followed by a verb (*hablaba*) and its complements. Except for the verb and the functional words, all words were stressed on the antepenultimate syllable. We obtained eight dialogues. We also used a broad focus sentence with the same structure as the other ones to study the prosody of PNucl word. These dialogues were written on cards and randomized with other sentences designed for other experiments. Each sentence was repeated four times by two female speakers of *Porteño* (one in her fifties, speaker PH, the other one in her thirties, speaker LD). The recordings took place in quiet rooms, and were done using the computer's sound card (44kHz 16bits) directly.

(1) Qu.: *¿*[*El miércoles*]$_{\text{TOP}}$ [*en Córdoba*]$_{\text{TOP}}$ [*quién*]$_{\text{WH-wd}}$ *hablaba con Nélida de mecánica?*
'Wednesday, in C., who was talking with N. about mecánica?'

Ans.: [*El miércoles*]$_{\text{TOP}}$ [*en Córdoba*]$_{\text{TOP}}$ [*Penélope*]$_{\text{IFoc}}$ *hablaba con Nélida de mecánica.*

(2) Q: ¿[*El miércoles último*]$_{\text{TOP}}$ [*Verónica*]$_{\text{QFoc}}$ *hablaba…* ?
'last Wednesday was Verónica talking…'

A: *No,* [*El miércoles último*]$_{\text{TOP}}$ [*Penélope*]$_{\text{CFoc}}$ *hablaba…*

The acoustic analysis of duration and fundamental frequency (F_0) were done using the Praat software package [24]. To compare the melodic realizations of the speakers and the various types of focus and topics, we converted the

original F_0 values in Hertz into semi-tones ("ST") by taking the frequency minimum of the utterances as the reference value. Moreover, in the vein of [25] and [26], we also generated time-normalized contours using the syllable as the domain of normalization and a 10-point time resolution per syllable.

3. Results and discussion: focus

Figure 1 below shows the time-normalized F_0 curves in ST of the PNucl and focus words in the different types of sentences (questions and declaratives) averaged across repetitions of the same type of focus.

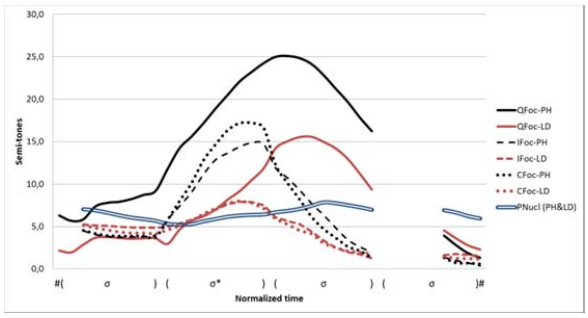


Figure 1: *time-normalized curves in semi-tones of PNucl- C/I/QFoc words for speakers PH and LD*

3.1. Pre-nuclear noun in broad focus sentences

Figure 1 shows a somewhat unexpected result concerning the pitch accent of the PNucl noun. As seen in the introduction, *Porteño* is said to differ from the other varieties of Spanish by realizing pre-nuclear accents as H* tones located *within* the stressed syllable. As can be seen in Figure 1, however, the accentual peak appears in the center of the post-tonic syllable, preceded by a low tone aligned with the beginning of the stressed syllable. This accentual 'late' rise is actually similar to the one found in the other varieties of Spanish (see [12]).

This accentual peak displacement is likely to be related to the proparoxytonic nature of the PNucl noun *Penélope*. Indeed, the other pre-nuclear proparoxytonic nouns in our corpus (*Verónica, Nélida, mecánica*) display the same tonal alignment; on the other hand, the paroxytonic verb *hablaba* occurs with the expected H* within the stressed syllable. A closer look at the literature about the prosody of *Porteño* reveals that most words taken into account in the previous studies are paroxytonic (or oxytonic); when proparoxytones are under investigation as in [27]'s study, these nouns undergo the same tonal alignment of H* as the one found in our recordings (see [27]'s figures).

3.2. Narrow focused nouns

As shown in Figure 1, all speakers realize a clear melodic contrast between the PNucl noun and the narrow focused words in terms of pitch height and tonal alignment. In declarative sentences, C/IFoc words are marked by a F_0 peak located *within* the stressed syllable, generally in the middle of the stressed vowel; we didn't find any evidence in our data for an early peak marking narrow focalization as reported by [19] and [21]. The peaks of C/IFoc words are preceded by a low tone located in the pre-tonic vowel. Figure 1 also shows that C/IFoc words are right bounded by another low tone, which reaches the low level of the speaker's range. Moreover, like in other languages and varieties of Spanish, all words after the focused element in declaratives are realized as a low plateau up to the end of the sentence or, at least, with very reduced pitch accents.

As can be seen in Figure 1, Sp.PH further distinguishes the C/IFoc words from the PNucl word by strongly increasing the height of the focal peaks in comparison with that of the PNucl word (+7.9 ST). Consequently, the words in focus are usually pronounced higher than the other pitch maxima in the sentence (those of the topics; see section 4.), whereas the PNucl high tone appears as downstepped compared to the preceding pitch accents ([11]–[13], [20], [21]). The melodic configurations of Sp.PH are thus along the lines of [19], [20]'s observations that focal peaks in *Porteño* are substantially higher than the non-focal ones. Notice though that Sp.LD does not focus words in this way; PNucl noun and I/CFoc words culminate at the same height.

As for questions, the stressed syllable of the QFoc word is realized by both speakers with a large rising contour (12/15 ST of range). This rise differs from those of the C/IFoc words in the alignment of the pitch peak: starting from the pre-tonic syllable, it ends with a peak anchored to the onset consonant of the post-stressed syllable, and not to the center of the stressed vowel as observed in C/IFoc words. In addition, for both speakers, the tonal peak of the QFoc words reaches a much higher F_0 value than that of the peaks of C/IFoc words; the difference is about 9 ST.

Nevertheless, Figure 1 reveals differences between the speakers as far as the tonal height is concerned. Sp.PH presents the highest F_0 values for the QFoc (25.1 ST), which reaches the top values of her pitch range (*falsetto* voice); she also realizes a pitch height contrast between the IFoc (14.7 ST) and CFoc (17.5 ST) words. On the other hand, although Sp.LD distinguishes between a QFoc word and C/IFoc words, she realizes the QFoc peak at a similar height to that of Sp.PH's C/IFoc words in declarative sentences.

As far as the tonal pattern of WH-word *quién* is concerned, it seems, *mutatis mutandis*, to be similar to that of the QFoc word. They are both characterized by a rise culminating at the same height. Their peak is aligned with the consonant immediately to the right and is followed by a steep fall. But the intonation of the rest of the sentence differs according to the type of word. In WH questions, all pitch accents rightward from the WH-word are de-accented; there is sometimes an optional final rise such as the one described by [11], [21] and [28]. In QFoc sentences, on the other hand, both speakers systematically produce a rising-falling contour at the end of the sentence; this contour is a kind of copy of the QFoc tone pattern and may be downstepped or not.

3.3. Duration of PNucl and focused words

The duration of the PNucl and focused words is given in Figure 2 below. As in other languages, there is an effect of focus on word duration: PNucl word is shorter than the focused words for both speakers. However, there are notable differences between the speakers. While Sp.PH exhibits a similar duration for all types of focus words, Sp.LD produces the CFoc noun with a much longer duration than the IFoc one; in addition, she greatly increases the length of the QFoc word as well. In fact, she is likely to lengthen the QFoc and CFoc words in order to compensate her lower F_0 values compared to

those of Sp.PH. Rising pitch or increasing duration thus seems to be two complementary strategies used by *Porteño* speakers to signal question and focus.

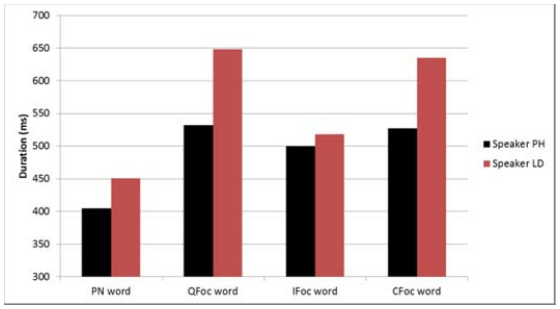


Figure 2: *duration of PNucl word and focus words*

3.4. Phonological representation

Let us first consider the tonal alignment of the various pitch peaks we have observed so far. It appears that the alignment of these peaks is actually *predictable* from the phonetic context. As seen above, the words in focus ends with a very low tone. Following [12], [29] and [20], we will interpret this tone as a phrasal tone L- and assume that it triggers the displacement of the peak toward the center of the stressed vowel in C/IFoc words. As for the delayed peak on QFoc and WH-words, it may be viewed as the result from an additional effect of the extra-high target to be reached. When there is no phrasal tone, as in PNucl words, the peak surfaces on the post-tonic syllable.

In the spirit of [12] and [20], we will therefore assume a unique underlying /(L+H)*/ pitch accent. The L tone accounts for the low valley that begins the PNucl and focus words and which contrasts with the high onset found at the beginning of the topics in questions (see section 4). The L tone is linked to the onset of the stressed syllable, and the H tone is more loosely associated. Like in standard Spanish, we will suppose that the post-tonic syllable is the default surface location of the H tone if there is 'enough room' ([12]) within the word, specifically, within proparoxytones in the case of *Porteño*. However, a gradual phenomenon conditioned by such factors as the proximity of other tones and word boundary moves the peak toward the left. Our proposal is, moreover, consistent with the results of [22] suggesting that, in this dialect, the perception of narrow focus is not achieved by tonal alignment but rather by duration and pitch height. Nevertheless, it remains to clarify whether the early peak in *Porteño*, reported by previous authors, is due to the (par)oxytonic nature of words.

The second issue we need to address is about the representation of the *tonal height* of the proposed pitch accent (L+H)*. Previous authors argued for a binary choice between two categories of high tones contrasting in pitch height: [11] claims a H+H* pitch accent to account for the pitch range increasing found in questions; [19]–[21] use an upstepped high tone they note "^H" or "¡H" for capturing the contrast between "neutral" high tones "H" and those marking questions or emphasis. These two categories of high tones allow us to adequately describe the melodic patterns of Sp.LD, using the label /(L+H)*/ for the PNucl and C/IFoc words and /(L+^H)*/ pitch accent for the QFoc and WH-words.

However, the tonal patterns produced by Sp.PH suggest a phonological contrast between three heights for the pitch peaks: indeed, the difference in pitch height between the PNucl noun, the C/IFoc words and the QFoc/WH-words varies in the same proportion (8 to 10 ST); and remember that for Sp.PH, pitch height is the only feature that distinguishes the C/IFoc words from QFoc and WH-words. Moreover, Sp.LD dramatically increases duration to distinguish between these different types of words. In other words, we think that there actually is a three-way categorical intonational contrast between the PNucl word, the focused words in declarative sentences and those in questions. But, the speakers adopt a different strategy to achieve this contrast: Sp.LD increases both pitch height and duration whereas for Sp.PH, contrasting pitch height appears to be the main distinctive feature. We will therefore hypothesize the following three pitch accents for Sp.PH: /(L+H^1)*/ *vs.* /(L+H^2)*/ *vs.* /(L+H^3)*/, H^3 indicating the top levels of the pitch range. Further investigations have to be carried out to answer whether Sp.LD has the H^3 tone in her intonational system; this tone might signal questions with surprise or incredulity for instance ([28]).

4. Topics

Let us begin with topicalization in the declarative sentences. In sentences with two topics (Figure 3 top), the speakers employ the (L+H)* pitch accent on each topic. The accentual H tone undergoes the gradual moving toward the left we've seen on the focus, that is, when the post-tonic syllable bears a low target, the H tone aligns with the middle of the stressed syllable (see F_0 track "D-2TOP-1"); if there is no subsequent low tone ("D-2TOP-2/3"), then the H tone appears on the post-tonic syllable. Both topics may end with a high tone that we analyze as a phrasal H-; this H- either surfaces as a rising contour LH- ("D-2TOP-1") or as a level high tone from the accentual H ("D-2TOP-2"). This is along the lines of [12] who proposed a phrasal H- signaling old information. In some recordings (cf. "D-2TOP-3"), however, there is a gradual fall from the accentual H up to the following accentual L belonging to the subsequent topic or focus; in other words, topics may end with no tonal target. Finally, note that there is no evidence for downstepped or upstepped topics.

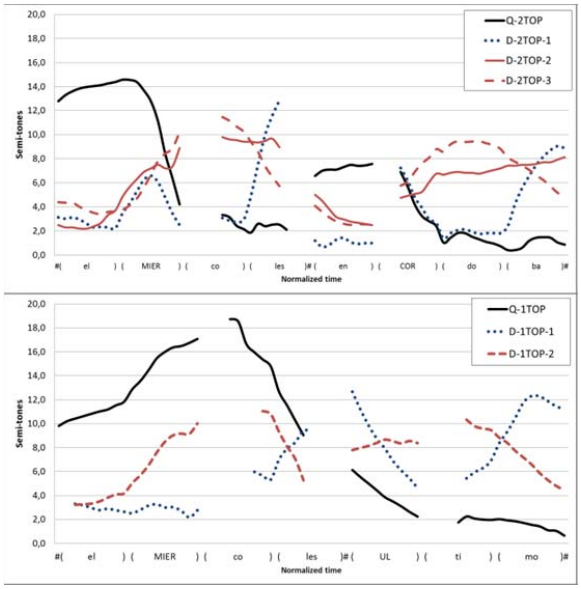


Figure 3: *time-normalized F_0 tracks in ST for the topics in questions ("Q-") and in declaratives ("D-"). Top: two topics (2TOP); bottom: one topic (1TOP).*

As for the topics comprising a single NP (Figure 3 bottom), two strategies are employed by the speakers. Sp.LD generally uses sequences of two [(L+H)*Ø] tunes associated with each lexical word (cf. "D-1TOP-2") and where the high peak is delayed because of the lack of phrasal tone. Sp.PH (cf. "D-1TOP-1") usually realizes the stressed syllable of the head noun *miércoles* with an L* tone, which is followed by a gradual rise to the beginning of the accented syllable of the adjective *último*. The adjective bears a single H* tone, which appears early within the stressed syllable because of the rising contour LH- ending the topic and the absence of an accentual leading L; we will label this topic tune as [L*..H* LH-].

Turning to questions, the tonal patterns of the topics are remarkably stable across both speakers; Figure 3 thus displays a unique averaged F_0 curve for each type of topic. In sentences with two topics, the tonal pattern of each topic is characterized by a steep fall on each stressed syllable, whose onset is associated with the peak and the end with the low target. We will represent this pitch accent as (H*L*), with two starred tones to indicate that they both associate within the stressed syllable. As can be seen in Figure 3, there is no evidence for a low tone before the H*: the melodic onset of the topics is almost at the same height as that of the H* and this clearly contrasts with the low onset at the beginning of the topics in declaratives or of the PNucl / focus words. The presence of an L* within the stressed syllable is further demonstrated by the very early alignment of the H* in comparison with those we have seen so far. Each topic ends with a low tone pronounced in the lowest part of the speakers' pitch range like the one right bounding the focus; we will therefore analyze this low tone as a phrasal L-.

As Figure 3 shows, the H* tone of the second topic is a lot lower than the first topic H* tone (about 7 ST) and accordingly seems to have been downstepped by the preceding H*L*. However, the H* tone of the second topic reaches similar values (around 7 ST) as those of the peaks found on the second topic in declarative sentences. Moreover, an intermediate phrase boundary (L/H-) arguably blocks any downstepping rule. On the other hand, there is a markedly increase in pitch height of the first topic H* tone in questions compared to the height of those found in declaratives (around 7/8 ST). In fact, the height in ST of the first topic H* tone corresponds to that of the focus in declarative sentences pronounced by Sp.PH or that of the QFoc uttered by Sp.LD (14-17 ST); we will thus claim that this H* tone belongs to the same tone category as the focus one, namely the H^2 category.

As regards the questions with a single topic, we also observed a unique tune for both speakers, we will represent as follows: [H^2*..L* L-]. This tune is characterized by an H^2* tone occurring at the end of the first stressed syllable. Again, this H^2* tone distinguishes itself from those occurring in declarative sentences in terms of pitch height (17 ST *vs*. 9 ST/ 3 ST respectively). It is connected to an L* tone appearing on the stressed syllable of the subsequent adjective by a gradual fall; the L* tone is then followed by a phrasal L-. Like two-topic sentences, no phonological tonal target is inserted at the onset of this tune because melodic curve confirms that the F_0 values at the beginning of the topics depend on the vicinity of the H^2* tone: the closer to the beginning the latter is, the higher the F_0 is produced.

The similarity between the tunes of both types of topicalization leads us to propose a more general template that we will note [(H^2*..L*) L-]. (H^2*..L*) refers to a 'splitting' pitch accent where both starred tones *must* be realized *within* a stressed syllable: inside a topic phrase, either they associate to separate stressed syllables as in the case of the single topic or, if there is only one stressed syllable, they both associate to the latter as we saw it in sentences with two topics.

We will conclude this paper with two final considerations. First, like in Modern Greek and French (cf. [7]–[9], [16]), it appears that, in *Porteño*, topic phrases in declaratives and questions exhibit opposite tonal patterns: in questions, topics have a (H^2*..L*) pitch accent followed by a L-, whereas, in declaratives, they have a (L+H)* or a L*..H* pitch accent followed by a H-. This type of inversion is thus likely to be a common intonational phenomenon that would deserve to be investigated and understood in more depth. In particular, one may wonder whether it is due to some phonological rule of tone inversion ([8]) or to a perception-oriented process that would help the listener determine whether the utterance she is processing is a declarative or a question. The tone inversion could also be induced by a difference in the activation state of the discourse referents in the sense of [15] (see also [16], [30], [31]): in our corpora, the topics of the questions are actually *inactive* ('new') whereas those of the declarative are *active* ('given').

The final point we wish to address is about the phonological structure of pitch height in *Porteño*. We saw that one can distinguish three (Sp.LD) or four (Sp.PH) pitch levels, which were given scalar values (L, H^1, H^2 and H^3). Like in tonal languages ([32]–[34]), however, it may be more accurate to use binary features such as the [±Upper] – a 'register' feature – and [±high] of [33] to account for the various pitch levels of *Porteño*. Within this system, we obtain the following equivalences: H^3=[+U,+h], H^2=[+U,-h], H^1=[-U,+h] and L=[-U,-h].

The main advantage of using these features is that it allows making generalizations across the speakers and the tonal patterns of the sentences. Thus, instead of merely seeing that Sp.LD uses a H^2 to signal questions and Sp.PH a H^3, it may now be stated that all speakers mark questions with the [+U] feature. Furthermore, being an autosegment, this feature may associate with other [high] features within the sentence. This would account for the increased pitch appearing both at the beginning and the end of questions. Let's say that *all* accentual tones in questions, except the one linked to the QFoc and WH-words ([+U,±h]) are actually underlying *low* tones, i.e. [-U,-h]. One may now propose that questions are further marked by the association of the [+U] borne by QFoc and WH-words with the *first* pitch accent of the sentence, generating a [+U,-h]* tone (i.e. H^2*), and, optionally, to the *final* sentence accent or boundary tone, giving the final rising or falling contours. Using a unique feature [+U] allows to generalize [11]'s proposal of a sentence initial H% and a H+H* pitch accent that express questions in Spanish: [11]'s H% and a H+H* are two distinct objects but they both have the same effect of upstepping pitch; on the other hand, a single [+U] feature provides a unified explanation to the increased pitch characterizing questions in Spanish.

5. Acknowledgements

We would like to thank Patricia Hernandez and Lucia Dorín from the *Instituto de Enseñanza Superior en Lenguas Vivas* "Juan Ramon Fernandez" in Buenos Aires for helping us to carry out this study and for participating in the recordings.

6. References

[1] D. Hirst and A. Di Cristo, Eds., *Intonation Systems: A Survey of Twenty Languages*. Cambridge, UK: Cambridge University Press, 1998.

[2] C. Lee, M. Gordon, and D. Buring, Eds., *Topic and Focus: Cross-Linguistic Perspective on Meaning and Intonation*, vol. 82. New York, NY: Springer-Verlag, 2007.

[3] M. E. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonol. Yearb.*, vol. 3, pp. 255–309, 1986.

[4] M. Rossi, *et al.*, Eds., *L'intonation: de l'acoustique à la sémantique*. Paris: Klincksieck, 1981.

[5] P. Touati, *Structures prosodiques du suédois et du français. Profils temporels et configurations tonales.* Lund, Sweden: Lund University Press, 1987.

[6] M. Rossi, *L'intonation, le système du français. Description et modélisation*. Ophrys, 1999.

[7] A. Di Cristo, "Intonation in French," in *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds. Cambridge, UK: Cambridge University Press, pp. 195–218, 1998.

[8] D. Le Gac and H.-Y. Yoo, "Intonative structure of focalization in French and Greek," in *Romance Languages and Linguistic Theory 2000: Selected papers from "Going Romance" 2000, Utrecht, 30 November–2 December*, C. Beyssade, R. Bok-Bennema, F. Drijkoningen, and P. Monachesi, Eds. Amsterdam / Philadelphia, PA: John Benjamins Publishing, pp. 213–231, 2002.

[9] M. Baltazani and S.-A. Jun, "Topic and focus intonation in Greek," in *Proc. 14th Int. Congr. Phonetics Sciences*, San Francisco, 1999.

[10] S. Alcoba and J. Murillo, "Intonation in Spanish," in *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds. Cambridge, UK: Cambridge University Press, pp. 152–178, 1998.

[11] J. Sosa, *La entonación del español: Su estructura fónica, variabilidad y dialectología*. Madrid: Catedra, 1999.

[12] J. I. Hualde, "Intonation in Spanish and the other Ibero-Romance Languages," in *Romance Phonology and Variation*, C. R. Wiltshire, Ed. Amsterdam: Benjamins, pp. 101–115, 2000.

[13] M. E. Beckman *et al.*, "Intonation across Spanish, in the Tones and Break Indices framework," *Probus*, vol. 14, no. 1, pp. 9–36, 2002.

[14] T. L. Face, "Local intonational marking of Spanish contrastive focus," *Probus*, vol. 14, no. 1, pp. 71–92, 2002.

[15] K. Lambrecht, *Information Structure and Sentence Form*. Cambridge, UK: Cambridge University Press, 1994.

[16] D. Le Gac and H. Yoo, "Intonation of left dislocated topics in Modern Greek," in *INTERSPEECH-2011*, pp. 1361–1364, 2011.

[17] G. A. Toledo, "H en el Español de Buenos Aires," *Lang. Linguist.*, no. 26, pp. 107–127, 2000.

[18] L. Colantoni and J. Gurlekian, "Convergence and intonation: Historical evidence from Buenos Aires Spanish," *Biling. Lang. Cogn.*, vol. 7, pp. 107–119, 2004.

[19] C. Gabriel, "Focal pitch accents and subject positions in Spanish: Comparing close-to-standard varieties and Argentinean Porteño." in *Speech Prosody 2006*, Dresden, Germany. Online: http://sprosig.isle.illinois.edu/sp2006/contents/papers/PS4-03_0028.pdf, accessed on 2006.

[20] P. Barjam, "The intonational phonology of Porteño Spanish", MA thesis, Univ. of California at Los Angeles, Los Angeles, CA, 2004.

[21] C. Gabriel *et al.*, "Argentinean Spanish Intonation," in *Transcription of Intonation of the Spanish Language*, P. Prieto and P. Roseano, Eds. LINCOM Publishers, pp. 285–317, 2010.

[22] J. Lang-Rigal, "Perception of Narrow Focus Prosody in Buenos Aires Spanish," in *Selected Proc. 5th Conf. Laboratory Approaches to Romance Phonology*, Provo, UT, Brigham Young University, pp. 118–126, 2011.

[23] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge, UK: Cambridge University Press, 2008.

[24] P. Boersma and D. Weenink, *PRAAT*. Online: http://www.fon.hum.uva.nl/praat/.

[25] Y. Xu, "Contextual tonal variations in Mandarin," *J. Phon.*, vol. 25, pp. 61–83, 1997.

[26] Y. Xu, "ProsodyPro - A tool for large-scale systematic prosody analysis," in *Proc. Tools and Resources Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France, pp. 7–10, 2013.

[27] I. Feldhausen, C. Gabriel, and A. Pešková, "Prosodic phrasing in Argentinean Spanish: Buenos Aires and Neuquén," in *Speech Prosody 2010*, Chicago, Illinois, DOI: 100111, 2010.

[28] S.-A. Lee, F. Martinez-Gil, and M. E. Beckman, "The intonational expression of incredulity in absolute interrogatives in Buenos Aires Spanish," in *Selected Proc. 4th Conf. on Laboratory Approaches to Spanish Phonology*, M. Ortega-Llebaria, Ed. Somerville, MA, USA: Cascadilla Proceedings Project, pp. 47–56, 2010.

[29] H. Nibert, "Phonetic and phonological evidence for intermediate phrasing in Spanish intonation," PhD Dissertation, University of Illinois at Urbana Champaign, Champaign, IL, 2000.

[30] S. Baumann, "Information structure and prosody: Linguistic categories for spoken language annotation," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, and J. Schliesser, Eds. Berlin, DE: Walter de Gruyter, pp. 153–180, 2006

[31] J. Doetjes, E. Delais-Roussarie, and P. Sleeman, "The prosody of left detached constituents in French," in *Speech Prosody 2002*, Aix-en-Provence, France, pp. 247-250, 2002.

[32] Z. Bao, *The Structure of Tone*. New York, NY/Oxford, UK: Oxford University Press, 1999.

[33] M. Yip, *Tone*. Cambridge, UK: Cambridge University Press, 2002.

[34] J. A. Goldsmith, E. Hume, and L. Wetzels, Eds., *Tones and Features: Phonetic and Phonological Perspectives*. Berlin, DE: Walter de Gruyter, 2011.

Analysis of Prosodic and Rhetorical Structural Influence on Pause Duration in Chinese Reading Texts

Liang Zhang[1], Yuan Jia[2], Aijun Li[3]

Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China

liangzhang1025@gmail.com, summeryuan_2003@126.com, liaj@cass.org.cn

Abstract

This paper investigates factors that influence pause duration in Chinese reading texts through examining the stress degree in pre-pausal and post-pausal positions and the rhetorical structure in discourse as a whole. The RSTTool is used in diagramming the rhetorical structures of the texts. The recordings, extracted from the ASCCD corpus, are further analyzed acoustically and statistically by applying Praat and R. The statistical analysis results show that the stress degree in both pre- and post-pausal positions has a significant impact on pause duration. Moreover, the nuclearity in both positions have also been shown to have a remarkable influence. Specifically, the nucleus in pre-pausal and satellite in post-pausal positions can significantly lengthen the pause duration.

1. Introduction

The Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) is a theory of text organization that has led to areas of application beyond discourse analysis and text generation. It has been applied to several linguistic areas, i.e. theoretical linguistics, psycholinguistics, and computational linguistics. RST defines a set of relations to identify the specific relationship that holds between two Elementary Discourse Units (EDUs) of a text. These relations are categorized into two types: mononuclear and multinuclear relations. After the elaborate description and categorization on the relations in large amount of real discourse spans, 30 rhetorical relations were studied in Mann (2005). A novel feature of RST is the concept of nuclearity. As well as presenting the relationship between two text spans, rhetorical relations also convey the information about which span is more central to the writer's purposes. Relations such as background and circumstance are of nucleus-satellite in that the EDUs linked by these relations are distinguished by their centrality: one is called the nucleus (N) and the other is called the satellite (S), with satellite subordinate to the nucleus.

Among various discourse structure theories, RST has been used widely in recent years to diagram the rhetorical hierarchy annotation. The popularity of RST has led to the development of an RST Treebank of manually annotated English texts, which is available for training and testing purposes (Carlson et al. 2003). It consists of 385 Wall Street Journal articles from the Penn Treebank (Marcus et al. 1993) with a total of 176,383 words. Another well-annotated RST corpus is Potsdam Commentary Corpus (PCC), which consists of 172 commentaries from Mearkische Allgemeine Zeitung, a German regional daily (Stede, 2004). Some other researchers (Stent, 2000; Taboada, 2004) also tried to apply RST to annotate spoken dialogues in Task dimension.

The development of RST in analyzing Chinese discourse was mainly in the areas of syntactic analysis (Chen, 2008), prosodic analysis (Tseng, 2006), Systemic Functional Grammar (SFG) (Wang & Dong, 1995), and Second Language Acquisition (SLA) (Wang and Xia, 2005). The other application field has been Computer Sciences and Language Processing, aiming principally at auto-annotation of rhetorical structures by training the model with hundreds of essays or news articles (T'sou et al., 1992; T'sou et al., 1996; Skoufaki, 2009).

Yue (2006) enriched Chinese language resources through building up a Chinese news commentary Treebank, using the RST as the theoretical framework. The corpus, consisting of 400 news texts of about 780,000 characters, has been applied to computation of a priori scores, needed in Chinese summarizers and be used as a platform for training and testing statistics-based discourse parsers. Their annotation efforts have proved, on a fairly large scale, the cross-language transferability of RST and its formalization.

The research on interfacing RST and prosody has attracted much attention. However, little work has been done in Mandarin Chinese discourse. Yang and Yang (2012) examined how rhetorical structures were reflected by boundary prosodic parameters in Mandarin Chinese discourses, through investigating recordings of ten paragraphs of news commentaries, with the prosodic parameters (pause duration, pitch reset, and final lengthening). The results were in line with previous studies (Noordman et al., 1999; Ouden et al., 2009). Nonetheless, no further detailed analysis has been done on stress degree, nuclearity and any other possible influential factors.

The above overview of previous studies shows that RST, which has been adopted in interdisciplinary research, were mainly restricted in the discourse dimension. The empirical study on the interface of prosody and RST, from the perspective of stress degree and nuclearity, is of fundamental importance but is largely under-explored. The work reported in this paper aims to fill this gap. Particularly, this paper is concerned with the duration of rhetorical pause in reading speech of Mandarin. We investigate the effect of the stress degree and nuclearity of preceding and upcoming EDUs, and its rhetorical structure on pause duration at different hierarchy within a RST diagram of the text.

2. Data

The materials selected in this current study are three reading texts chosen from the Annotated Speech Corpus of Chinese Discourse (ASCCD), which was built by the Phonetic Lab, Institute of Linguistics, Chinese Academy of Social Sciences (CASS). The data were collected from ten Mandarin speakers (5 males and 5 females) in Beijing. The C-ToBI system was used for annotation and four tiers were labeled. In this research, the stress tier (ST) is used, in which each prosodic unit were annotated with one of the four degrees (0: weak; 1: normal; 2: secondary stress; 3: primary stress). The stress degrees of preceding and succeeding rhetorical pauses are extracted for further analysis.

The RSTTool provided by O' Donell (1997) was used in this study to create the RST diagram. One disadvantage of RST, as mentioned in previous study, is the comparative subjectivity in labeling relations. The diagrams used in this study were double-checked by another researcher in the same field, in order to avoid inaccuracy as much as possible. The resultant RST diagram is showed in Figure 1.

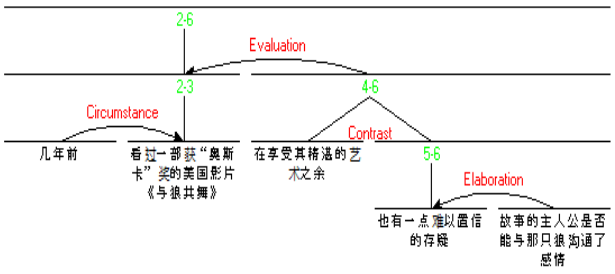


Figure 1. *An example of RST diagram*

(A few years ago/ I watched an Oscar-winning film: Dances with Wolves/ besides the enjoyment of exquisite art/ I found one point incredible/ whether the hero of the story really communicated with the wolf.)

In all multi-span diagrams, each rhetorical pause is marked with the sum of spans of adjacent EDUs. For instance, “Circumstance” in Figure 1 would be labeled as 4 (2+2, sum of first and second EDU span, with the highest span excluded). Similarly, we get 6 for “Contrast” and 6 for “Elaboration”. For three texts in total, the spans are ranked from 3 to 14. To facilitate statistical analysis, the spans are re-categorized into four hierarchies with 3 consecutive spans in each category.

Pause duration is defined as the silence interval between the ending of one segment and the beginning of the next segment. Based on the rhetorical structures of each discourse, the durations of rhetorical pauses are further annotated and extracted from the corpus.

In the three texts, 156 rhetorical relations are diagramed, and accordingly 1560 pause durations are extracted.

3. Methodology, Results and Analysis

In this section, the study systematically examines three factors that could influence the pause duration: rhetorical hierarchy, stress degree and nuclearity in the preceding and succeeding pause positions. These three factors fall into two categories: prosody and rhetorical structure. Through R for statistical computing, t-test and ANOVA analysis are adopted to investigate the inner-relations and correlations of these variables.

3.1. Rhetorical Relations and Hierarchy

To study the effect of hierarchical position with four levels on pause duration, an ANOVA analysis is conducted. The results show that the depth of the hierarchies significantly affects the duration of the pauses ($p<0.001$), indicating that pause duration gets longer when the depth of hierarchies increases. This is consistent with the findings from previous studies on Chinese and other languages such as Spanish, English.

The result shows that the schema *Title* takes the longest pause in all, which reflects the widest semantic distance between adjacent segments, while *Summary* takes the second longest. No significant difference is found ($p=0.166$) between causal and non-causal relationships.

Despite the accordance with the previous results in Yang and Yang (2012), this paper, by involving stress degree and nuclearity, further analyzed their effects on pause duration, and their inner relationship with rhetorical hierarchy.

3.2. Stress Degree

For the stress degree in the pre-pausal position, statistical analysis result (Figure 2) shows a significant difference ($p<0.001$) between ST1 and ST3. As the stress degree goes higher (from 1 to 3), the pause duration decreases.

The same significant difference is observed in the post-pausal position between ST1 and ST2 ($p<0.05$) and between ST1 and ST3 ($p<0.001$). However, with stress degree goes higher, the pause duration increases.

Based on the above results, it is clear that the stress degrees at both preceding and succeeding pause EDUs have a significant influence on pause duration, while the stress degree in the post-pausal position shows a higher influence on each stress degree, as demonstrated in Figure 2. The X-axis is the stress degree in previous EDU, while the boxplot is colored according to stress degree in post-EDU. The post-pausal stress degree, from red to blue, shows an increasing trend without the restriction of stress degree in pre-pausal position. While for pause duration in preceding pause position, there is no visually obvious decrease from normal to primary stress.

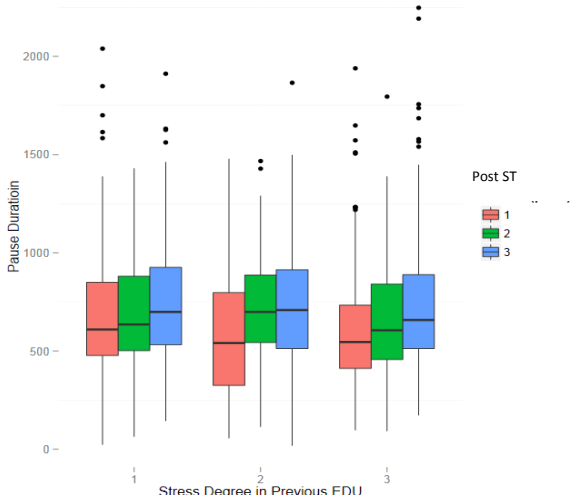


Figure 2. *The significant influence of upcoming stress degree on pause duration*

In the next step, we examine the distribution of duration in combined stress type in four hierarchies. With reference to the results above, as showed in Figure 3, comparatively speaking, the ST combination “3-1” (up right square) and “1-3” (down left square) generates the shortest and longest duration respectively.

Another noticeable point is that the pause duration and rhetorical hierarchy theory generally but does not necessarily apply to every subordinate category (see circles in Figure 3). This inspires us to make a novel assumption: there is a ranking of the influential factors of pause duration, which leads to the question: what is the ranking of the factors, such as hierarchy,

stress degree, nuclearity, sentence complexity that influence pause duration in reading texts? Further attention should be paid to this interesting point and more experiments need to be done. It is however currently beyond the scope of this paper.

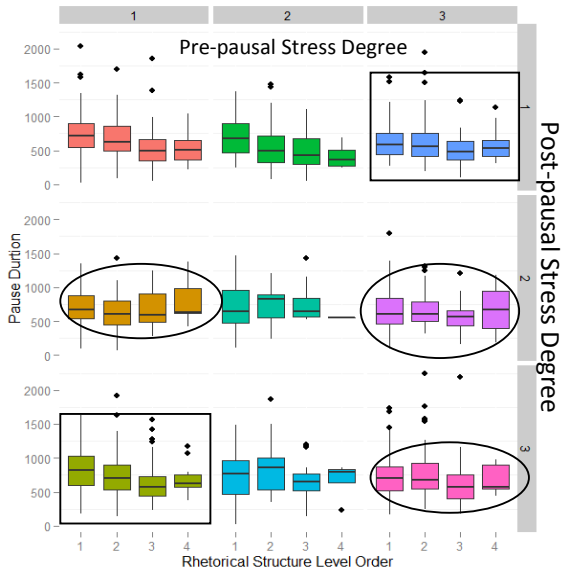


Figure 3. *Pause duration in different stress degree combinations in four rhetorical hierarchies*

3.3. Nuclearity

This part examines the nuclearity of EDUs in proceeding and succeeding rhetorical pause position, aiming at investigating possible influence of nuclearity on pause duration.

3.3.1. Nucleus and Satellite EDU

In the nuclearity of the pre-pausal EDU, statistic result shows a significant difference ($p<0.001$) between N and S. The result indicates that if the EDU is a nucleus in the rhetorical relation, the duration of its succeeding pause would be significantly longer than that after a satellite EDU.

The same significant difference is found in the post-pausal EDU between N and S ($p<0.001$). However, the result is opposite to the previous one: if the EDU is a nucleus one in the rhetorical relation, the duration of its preceding pause would be significantly shorter than that before a satellite EDU.

Therefore, when the test came to pause duration in mononuclear relations as a whole, we assume and has confirmed that the influence of nuclearity on duration is neutralized with no significant difference appears in t.test between N-S and S-N combination, even though practically the N-S combination produces slightly longer pause duration than that in the S-N combination.

3.3.2. Nuclearity and Rhetorical Relations

A closer examination is performed on the components of rhetorical relations in both combinations. Among the 345 tokens of N-S relations, nearly one third are *Evaluation* relations, which, unlike other relations, carry particular emotions (Hou, 2012). Moreover, from the discourse analysis point of view, *Evaluation* indicates the change of footing registration of the writer, with the purpose of jumping out of the narration and gaining interaction and solidarity with readers. The change in registration were clearly obtained by the reader, and thus in the recordings, they may produce a longer pause to indicate the register change.

With the purpose of further analysis of the distribution of mononuclear relation in four hierarchies, Figure 4 is plotted out with the nuclearity combination and pause duration.

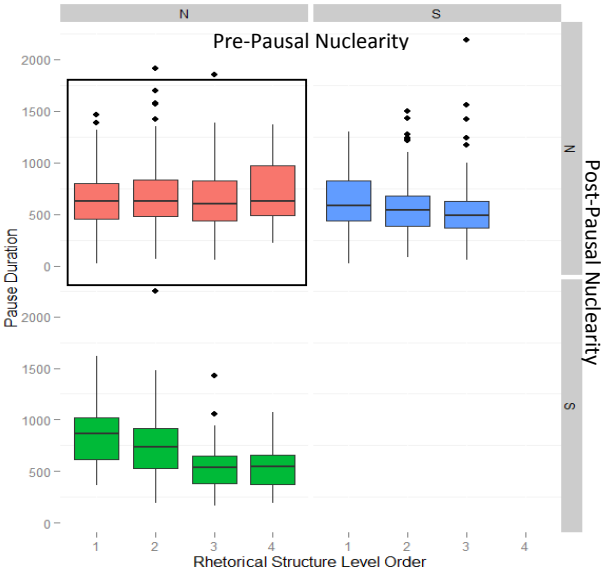


Figure 4. *Pause duration in different nuclearity combinations in four rhetorical hierarchies*

Two points are notable in the boxplot above.

1. The neutralization of pre-and post-pausal nuclearity influence could be resulted in by the uneven distribution of S-N combination in four levels.

2. The pause duration in multinuclear relations (in the square) does not follow the hierarchy pattern show a decrease with the increase of hierarchy. There was no significant difference at all between each rhetorical hierarchy. This leads back to the assumption that mentioned in the stress degree section. Is the duration of multinuclear relations not restricted by the hierarchy theory? Could nuclearity rank first in the restraints? These questions definitely require more attentions and further detailed experiments.

4. Discussion

The study, with the adoption of RST and stress degree on three reading texts may suffer from several biases in the following aspects. First, it is a comparatively small corpus, which may result in skewness of the data. To be specific, there is no S-N combination in forth rhetorical hierarchy. Secondly, though double-check was applied during the process of labeling rhetorical relations, there could still be few mistakes since the

inevitability of subjectivity. Third, the accuracy and consistency of stress degree annotation, which was done based on pure perception of experienced linguistics.

The results also show a gendered variability, in which the male speakers tend to have longer pause duration than the female speakers in every dimension.

5. Conclusion

Through RST labeling on Chinese reading texts and statistical analysis on stress degree and nuclearity in pre- and post-pausal position, the present study explores the influential factors of pause duration. The following observations can be inferred based on the results: stress degree in both pre- and post-pausal positions has significant influence on pause duration, in which "1-3" indicates longest duration while "3-1" the shortest. Nuclearity in both positions separately showed remarkable effect on pause duration: nucleus in pre-pausal and satellite in post-pausal position significantly lengthened pause duration.

This is a pioneer research between RST and prosody with the parameter hierarchy, stress degree combination and nuclearity, which not only jumped out of the hierarchy restriction, but also casts new light on the interface research.

Further experiments are worth investigation given the assumed sequence of influential factors of pause duration, such as stress degree combination, multinuclear relation, and rhetorical hierarchy. Larger corpus is needed for a more thorough research. It may also involve logical speech discourses, such as presentations or story-telling, which are more complicated in topics and less structured in construction.

6. Acknowledgements

This research is supported by National Program on Key Basic Research Project (973 Program) under Grant 2013 CB329301, as well as the Innovation Program of Chinese Academy of Social Sciences "Key Laboratory of Phonetics and Speech Science".

References

[1] Carlson, L., D. Marcu, & M.E. Okurowski (2003), Building a discourse-tagged corpus in the framework of rhetorical structure theory, in van Kuppevelt, J. and R. Smith, editors, *Current Directions in Discourse and Dialogue*, Kluwer Academic Publishers, Dordrecht, the Netherlands, pp. 85–112.

[2] Chen, Liping, (2008), Research on Rhetorical Structure Theory and Sentence group. *Journal of Suzhou University (Philosophy & Social Science)*. Jul. No. 4.

[3] Dai, Weihua. & Xue, Yan. (2004). Rhetorical Genre Studies and Rhetorical Structure Theory. Foreign Language Education. Vol.25. No. 3. pp. 35-39.

[4] Mann, W. & Thompson . (1988). S. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243-281

[5] Mann, W. (2005), RST Web Site, now http://www.sfu.ca/rst

[6] Marcus, P.M., M.A. Marcinkiewicz, and B. Santorini (1993), Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics*19 (2), pp. 313–330.

[7] Noordman, L., Dassen, I., Swerts, M., Terken, J. (1999). Prosodic markers of text structure. In: van Hoek, K., Kibrik, A., Noordman, L. (Eds.), *Discourse Studies in Cognitive Linguistics, Selected Papers 5th International Cognitive Linguistics Conference.* John Benjamins, Amsterdam, pp. 133–148.

[8] O' Donnell, Michael. (1997). RST-Tool: An RST analysis tool. In Proceedings of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisburg, Germany.

[9] Ouden, J.N. den, Noordman, L. & Terken, J.M.B. (2002). The Prosodic Realization of Organizational Features of Texts. In B. Bel & I. Marlien (Eds.), *Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence, France,* (*pp. 543-546*). Aix-en-Provence, France: Laboratoire Parole et Language.

[10] Rao, Rajiv. (2010), Final Lengthening and Pause Duration in Three Dialects of Spanish. *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology.* Somerville, MA, USA.

[11] Skoufaki, Sophia. (2009). An Exploratory Application of Rhetorical Structure Theory to Detect Coherence Errors in L2 English Writing: Possible Implications for Automated Writing Evaluation Software. *Computational Linguistics and Chinese Language Processing*. Vol. 14, No.2, pp. 181-204

[12] Stede, Manfred. (2004). The Potsdam Commentary Corpus. *In Preceedings of the ACL 2004 Workshop 'Discourse Annotation'*, Barcelona. 2004

[13] Taboada, Maite. (2004). Rhetorical Relations in Dialogue: A contrastive study . In C L. Moder & A. Marinovic-Zic (Eds.), Discourse Across Languages and Cultures. Amsterdam and Philadelphia: John Benjamins.

[14] T'sou, Benjamin, Ho, H., Lai, B., Lun, C. & Lin, H. (1992). A Knowledge-based Machine-aided Syetem for Chinese Text Abstraction. *In Proceedings of International Conference on Computational Linguistics*, France, pp. 1039-1042.

[15] T'sou, Benjamin. , Lai, Tom., Chan, Samuel., Gao, Weijun. & Zhan, Xuegang (2000). Enhancement of Chinese Discourse Marker Tagger with C4.5. In Proceedings of the Second Chinese Language Processing Workshop. Hong Kong. 38-45.

[16] Wang, Wei & Dong, Jiping. (1995) Rhetorical Structure Theory and Systemic Functional Grammar. Shandong Foreign Language Teaching. Vol.2.

[17] Wang , Yanping, & Xia, Zhen. (2005). A Contrastive Study of Rhetorical Structure in Argumentative Texts: An Empirical Study in Eneligh-Chinese contrastive Rhetoric. *Journal of Henan University (Social Sciences)*. Vol. 45. No. 5.

[18] Yang, Xiaohong & Yang, Yufang. (2012). Prosodic Realization of Rhetorical Structure in Chinese Discourse. *IEEE Transactions on Audio, Speech , and Language Processing.* Vol 20. No.4

[19] Yue, Ming. (2006). Annotation and Analysis of Chinese Financial News Commentaries in Terms of Rhetorical Structure. Ph.D. Thesis. Communication University of China. China

[20] Zvonik, Elena. & Cummins, Fred. (2002). Pause Duration and Variability in Reading Texts. *In preceeding of 7th International Conference on Spoken Language Processing, Interspeech*, Denver, Colorado, USA.

[21] Tseng, Chiuyu. (2006). Higher Level Organization and Discourse Prosody. *The Second International Symposium on Tonal Aspects of Language*. pp.23-34

Statistical and temporal properties of prosodic units in French conversational speech

Irina Nesterenko

INALCO, France

irina.nesterenko@inalco.fr

Abstract

Our study investigates statistical and temporal properties of prosodic units, which were previously identified within laboratory phonology paradigm, in a corpus of French conversational speech. Prosodic annotation of our corpus implements two-level hierarchical model distinguishing major prosodic units (Intonational Phrases, *IPs*) and minor prosodic units (Accentual Phrases, *AP*). Both temporal data and distribution of the number of APs in an IP evidence the global tendency to produce shorter units in conversation. Moreover, Intonational phrases containing no more than two Accentual phrases cover 80% of the data. We discuss the implication of these results for both phonological studies on the constraints on prosodic phrasing and oral document tagging.

Index Terms: French, prosodic phrasing, conversational speech, temporal organization

1. Introduction

Our study focuses on the issue of prosodic phrasing in conversational speech. Prosodic phrasing refers to the structuring of speech material in terms of boundaries and groupings. These boundaries vary as to their relative strength thus defining a number of levels in prosodic constituency. Recently, prosodic phrasing has been the subject of advanced formal modeling within the framework of prosodic phonology [1-2] though the extensive phonetic and phonological studies dealt with laboratory speech.

In the models of prosodic phrasing proposed for French it is common to distinguish two levels of phrasing above the word (cf. [3]): Intonational phrases (IP) and Accentual phrases (AP, though the label can differ between the authors, cf. rhythmic units of [4], prosodic words of [5] and phonological phrases of [6]). Both units receive in speech a specific phonetic marking. Thus, there is an obligatory F0 rise on the last syllable of the accentual phrase (labeled LH* in autosegmental-metrical model of Jun & Fougeron) accompanied by pre-boundary lengthening [7] and, optionally, by an initial rise (Hi) on the first syllable. On the other hand, the intonational phrase is the domain of a major intonation contour and is characterized by greater degree of pre-boundary lengthening [3, 8]. There is as well semantic-pragmatic and syntactic constraints on Intonational phrase boundaries distributions: for example, several syntactic constructions, such as root clauses, vocatives and parenthetical expressions, form *IPs* of their own; cf. as well Sense unit constraint as defined in [9] and tested in [10] .

In our study we rely on the axiom that prosodic units previously identified within laboratory phonology paradigm could be identified in conversational speech [11]. We consider that such corpus studies are about how the language is spoken and how available prosodic means are exploited by the speakers, prosodic units in conversation being endowed with organizing function. We addressed the issue of interaction between two levels of prosodic hierarchy in conversational speech and we focused on temporal properties of prosodic constituents.

2. Corpus and methodology

Our study is based on an excerpt from the *Corpus of Interactional Data* [12] (http://crdo.up.univaix.fr/corpus.php?langue=fr). We focused on one dialogue between two familiar female speakers who conversed on humorous situations in which they may have found themselves involved. From the interactional point of view, the corpus is not homogenous, combining negotiation sequences, question-answer exchanges and substantial monologues. The total size of the corpus was 12681 words.

The corpus was manually transcribed using an enriched orthography: in order to facilitate further processing of the corpus, our transcription conventions include special notations to signal a number of reduction phenomena (i.e. elisions, word truncations). Next, this transcription was automatically converted to a phonemic transcription of speech material and then automatically aligned to the speech signal. Subsequently, the corpus was enriched with various linguistic annotations (manual or (semi-)automatic) as a means to study interfaces between phonetics, phonology, prosody, morphology, syntax, pragmatics, discourse and gesture as they operate in conversational speech. In the following paragraphs we detail the syntactic and prosodic annotation underlying our study.

2.1. Prosodic annotation

The general prosodic annotation scheme for the corpus includes

- metrical structure in terms of perceived prominences;
- tonal structure: we distinguish the level of underlying tones and the level of surface tones (INTSINT);
- prosodic constituency.

The corpus was manually annotated in terms of IP and AP boundaries. This annotation was guided by perception, based on a distinction between strong and weak prosodic breaks. Acoustic cues to prosodic phrasing, salient in perception, were also taken into account. Thus acoustic and perceptual cues to an IP boundary are: i) an intonation unit is associated with a specific melodic contour; ii) there is a high (H) or a low (L) boundary; iii) there is *pitch reset*; and iv) there is pre-boundary lengthening. Acoustic and perceptual cues to an AP boundary are: i) specific pitch movement (final rise); ii) relative scaling of adjacent APs; iii) slight degree of preboundary lengthening.

Before proceeding with the phonetic and statistical analyses, we ascertained that our annotations were sufficiently reliable (cf. [13, 14] attesting mean kappa Cohen values of 0.79).

3. Results

3.1. Two levels of hierarchy

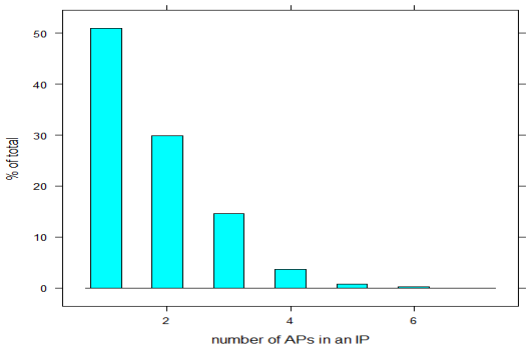


Figure 1: Distribution of the number of Accentual Phrases per Intonational Phrase

First, we looked at the distribution of number of APs in an IP (Fig.1). On average, an IP groups together 1,74 APs. Note that almost half of the identified IP units contain only one AP. In [15] studying radio speech in English, the author communicates that the IPs with no more than two accents cover 80% of the corpus, though almost equal proportions of IPs with one (39%) and two (41%) prosodic prominences was observed in this study. In our data, IPs containing only one AP, i.e. with only one intonational prominence, dominate. We consider such a distribution as quite typical for conversational data: in fact, conversation abounds in short replies and interaction words, speakers need to provide context settings for their interlocutors; as a consequence, quite often, but not always, such units tend to form a prosodic unit on their own and are quite often cumulated at the beginning of a phrase (cf. Fig 2 which illustrate the waveform, the F0 contour and phrasing annotation of the utterance $[(Demain)_{AP}]_{IP}[(c'est)_{AP}(le\ repas)_{AP}(de\ Noël)_{AP}]_{IP}$, 'Tomorrow, there is a Christmas meal': time setting adverbe *demain* 'tomorrow' being set apart in a separate IP (of 0.517s); *we should specify that there is no error in Praat's pitch detection at the end of first IP, the gap being induced by creaky voice*).

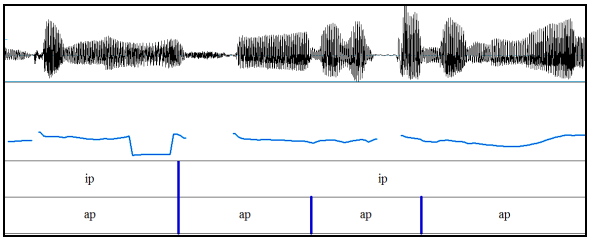


Figure 2: Waveform, F0 contour and phrasing annotation of the utterance $[(Demain)_{AP}]_{IP}[(c'est)_{AP}(le\ repas)_{AP}(de\ Noël)_{AP}]_{IP}$, 'Tomorrow, there is a Christmas meal'

Our data attest that there is a relatively small number of units with more than 3 APs: such a finding has its implications for studies of prosody-syntaxe mapping and theoretical issue of prosodic levels in French. In fact, laboratory studies showed that IPs with more than three APs tend to be restructured and such a restructuring is accompanied by the emergence of an additional level of phrasing, that of Intermediate Phrases (*ips*) (such restructuring applies to specially identified syntactic structures [17,18],; cf. as well the constraint on prosodic phrasing authorising only 2 prosodic prominences per unit). Our data indicate that such a restructuring is probable in only 20% of units, all the syntactic structures put together. We suggest that Intermediate phrase level in French should not be described not within Strict Layer Hypothesis on prosodic phrasing [9], but within a probabilistic framework for which corpus studies are an important source of data: the level of Intermediate Phrases emerges whenever the required conditions are met, as to the distribution of prosodic prominences, focusing and speech rate.

3.2. Temporal organisation of perceived units

Next we looked at temporal organisation of prosodic units in spontaneous speech and analysed the distributions of durations.

3.2.1. Intonational phrases

Mean duration of IPs in our data is of 0.841s. (75% of units are shorter than 1.14s., though they exhibit a rather important variation, coefficient of variation C_v = 0.589). In our previous study based on Russian spontaneous speech [16] we obtain similar results for major prosodic units (mean duration of 0.86s.). Both contrast with the data on reading in which the authors communicate the mean durations of prosodic units as of 1.2-1.6 s. Consequently, mean IP duration seems to be a property, which allows distinguishing conversation and reading and consequently, tagging oral documents from these two styles.

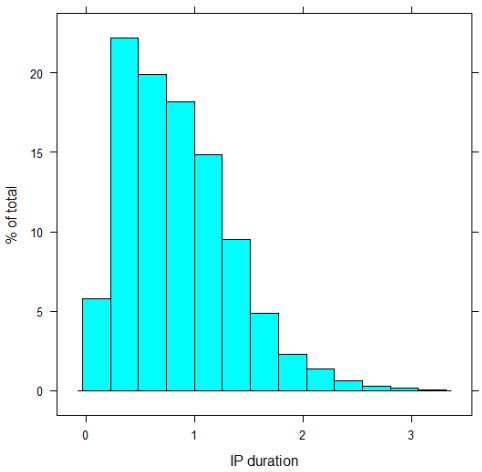


Figure 3: Distribution of Intonational Phrase durations

On the next step, we compared IP length distributions for two speakers separately (Table 1). In our study, one of the speakers tends to produce shorter units than the other, but the mean values are still under the level observed in earlier studies of reading. The differences between the speakers could be related both to speech rate differences (cf. [17] on longer units in fast speech) and to different strategies for prosodic structuring of speech (this hypothesis could be tested by comparing unit lengths in number of syllables).

Table 1. *Description of IP durations distributions for two speakers*

Speaker	AB	CM
Min	0.08	0.084
1st quartile	0.528	0.358
Median	0.845	0.625
Mean	0.905	0.709
3rd quartile	1.208	0.96
Max	3.194	2.844

3.2.2. *AP durations*

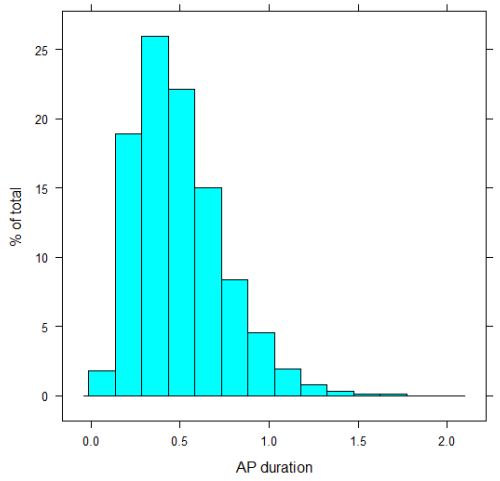


Figure 4: Distribution of Accentual Phrase durations.

The AP lengths in our corpus present the similar distribution as IP lengths with mean AP duration being of 0.496s. (75% of units being shorter than 0.636s. with substantial internal variation, C_v = 0.501).

3.2.3. *Speech-in-interaction effect*

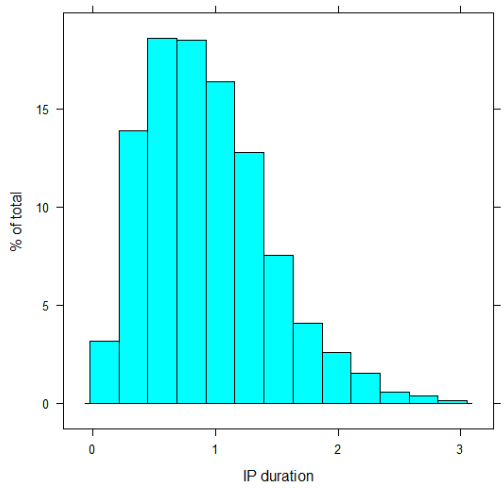


Figure 5: Distribution of Intonational Phrase durations in spontaneous monologues

As described earlier, our corpus contains both task- and strategy negotiation sequences and monologues from one and the other speakers. During these monologues, the participation of the second interlocutor is limited to backchannels and some rare questions and reactions. Thus, to sort out speech-in-interaction effect we limited ourselves to monologue sequences only and looked for IP duration distributions in this sub-corpus. IP number was thus limited by 30%. The mean IP duration for sub-corpus is slightly higher than for the totality of data: mean IP duration is of 0.937 and 75% of IPs are shorter than 1.234s, variation coefficient value (C_v=0.541) indicating great dispersion of values around the mean. This finding goes in the sense of our hypothesis

4. Discussion and Conclusions

Based on a large corpus, our study is about prosody production. Starting with the idea that prosodic constituents previously identified in laboratory speech are relevant for the analysis of conversational prosody, we investigated hierarchical relation between major (Intonational phrases) and minor (Accentual phrases) prosodic units in French. In the analysis of the results we took into account both phonological constraints on prosodic phrasing and details of phonetic (temporal) organisation of these units.

We found that Intonational phrases containing only one Accentual Phrase are dominating in our corpus, quite in agreement with its style. In this aspect, our corpus contrasts with the read speech and news broadcasts [15], though we do not possess directly comparable data in French. Moreover, Intonational units containing no more than two Accentual Units cover 80% of the data. This finding has its implications for the issue of number of levels in prosodic structuring in French: in fact, it means that there are the appropriate conditions for the emergence of an intermediate level of phrasing in only 20% of contexts. We need to further test the hypothesis if there really are any phonetic indices (greater pre-boundary lengthening and/or tonal cues) indicating the presence of an intermediate phrase boundary in these potential locations. We propose that further modeling of prosodic phrasing in spontaneous speech integrates this probabilistic data of the number of phrasing levels.

As to the temporal dimension of prosodic units, we obtained that spontaneous speech is structured in shorter units than read speech (mean IP duration of 0.841s in our study and that of 1.2 -1.6 s. communicated in reading). For the restricted corpus of monologue sequences, mean IP length slightly augments up to 0.937s. Restricted corpus contained less short units which function as backchannels, though there is still a number of context setting phrases promulgated to the level of IPs in conversation. Both the data on IP duration and that on AP duration exhibit a great dispersion around the mean value. At the same time, our data on conversational speech show that mean IP duration could potentially be one of the parameters allowing for tagging apart conversations from other oral documents.

We observed as well speaker's effect on phrase duration, our data attesting the tendency of one of the speaker to produce shorter units than the other. This difference could be imputed to speech rate differences and subsequent tendency to produce larger units (cf. [17] uncovering at least two mechanisms available to the speaker in fast speech rate which have an impact on the number and the strength of prosodic boundaries) or individual strategies in speech structuring. Both hypotheses need further investigation.

5. References

[1] Nespor, M. and Vogel, I., Prosodic Phonology, Foris Publication, Dordrecht, 1986.

[2] Beckman, M.E. and Pierrehumbert, J.B., "Intonational Structure in Japanese and English", Phonology Yearbook, 3:255-309, 1986.

[3] Jun, S.A. and C. Fougeron.,"A Phonological Model of French Intonation", Probus, 14, 147-172, 2000.

[4] Di Cristo, A., "Intonation in French". In D. Hirst, & A. Di Cristo (Ed), Intonations systems: A survey of twenty languages, Cambridge: Cambridge University Press, 195-218, 1998.

[5] Vaissière, J., "Rhythm, Accentuation and Final Lengthening in French". In J. Sundberg, L. Nord and R. Carlson [Ed], Music, language, speech and brain, 59,108-120, Stockholm : Wenner-Gren, International Symposium Series, 108-120, 1992.

[6] Post, B. "Tonal and Phrasal Structures in French Intonation". The Hague: Holland Academic Graphics, 2000.

[7] Pasdeloup, V., "Modèle de règles rythmiques du français appliqué à la synthèse de parole". Doctoral thesis, Université de Provence, 1990.

[8] Di Cristo, A., "De la microprosodie à l'intonosyntaxe". Aix-Marseille : Université de Provence, 1985.

[9] Selkirk, E., Phonology and Syntax: The Relation Between Sound and Structure, Coll. Current Studies in Linguistics Series, 10. Cambridge, MA, USA : The MIT Press.

[10] Frazier, L., Clifton, Ch. Jr., and Carlson, K., "Don't break, or do: prosodic boundary preferences", Lingua, Vol. 114(1), 2004, 3-27.

[11] Portes, C., Betrand, R., " Permanence et variation des unités prosodiques dans le discours et l'interaction", Journal of French Language Studies, 2011.

[12] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and S. Rauzy, "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle", Traitement automatique des langues (TAL), 49(3):105-134, 2008.

[13] Cohen, J., "A coefficient of agreement for nominal scales", Educational and Psychological Measurement, Vol. 20, 1960, 37-46.

[14] Nesterenko, I., Bertrand, R., Rauzy, S., "Prosody in a corpus of French spontaneous speech: perception, annotation and prosody ~ syntax interaction", Speech Prosody, Chicago, USA, 2010.

[15] Dainora, A., "An Empirically based Probabilistic model of Intonation in Americain English", PhD Dissertation, University of Chicago, IL, 2001.

[16] Nesterenko, I., Skreline, P., "Some evidence on the phonetics and phonology of prosodic phrasing in Russian". Interspeech 2007 : 438-441

[17] Jun, S.A. and Fougeron, C., "The Realizations of the Accentual Phrase in French Intonation". Probus, 14, 147-172, 2002.

[18] D'Imperio, M., Michelas, A., « Embedded register levels and prosodic phrasing in French ». Speech Prosody, Chicago, USA, 2010, 100879:1-4

Intonational Patterns of Telephone Numbers in Brazilian Portuguese

Oyedeji Musiliyu [1], *Miguel Oliveira Jr.* [2]

[1, 2] Universidade Federal de Alagoas, Alagoas, Brasil

bodeses@yahoo.fr, miguel@fale.ufal.br

Abstract

The main purpose of this paper was to identify intonational patterns of a quite common type of numeric grouping in Brazilian Portuguese: the one associated with telephone numbers. To this aim, 30 samples of spoken telephone numbers, read aloud by 85 native speakers of Brazilian Portuguese were analysed. The description of the intonation contours was observed by using *Momel/Intsint* [1] and *ProsodyPro* [2] scripts for *Praat* (version 5.3.53) [3], through a semi-automatic analysis of pitch variations in numeric groupings that form the telephone numbers. The results show a pattern of intonation and numeric grouping strategy that are sufficient enough to characterize prosodically different types of spoken telephone numbers in Brazilian Portuguese.

Keywords: telephone numbers, intonation, Brazilian Portuguese.

1. Introduction

In the past few years, speech technology advancements have made frequent the use of automatic speech recognition and synthesis systems in multiple applications. Some services based on these automated systems make use of concatenated numeric digits for various purposes, such as: activation of credit cards, banking information, telephone directories, booking inquiries and assistance services for blind and visually impaired individuals, for example.

Services that make use of numerical groupings as input or output data depend on a good voice or textual data processing system of alphanumeric digits information. An efficient system will recognize adequately spontaneous speech and produce a voice or textual output corresponding to adequate enunciation of a given numeric grouping.

In many cases, however, the performance of these systems are considered deficient, either for not processing adequately natural speech (in the case of speech recognition systems), or for not presenting in its production the expected characteristics of naturally produced speech (in the case of speech synthesis systems). This is due in part to the fact that such systems are mostly based on outdated, non-spontaneous speech data.

Progress has been achieved in that respect as a result of the description of enunciation of natural numbers in various languages, such as German [4], Japanese [5] and French ([6] and [7]). To this date, however, no study has yet described, in a systematic and comprehensive manner, the various acoustic features of the organization of natural numbers in predetermined structures in Brazilian Portuguese.

1.1. Aims of the study

The aim of this study is two-fold: firstly, to segment the number sequence into prosodic sub-groupings in order to investigate the grouping and wording strategy applied to telephone numbers of different length in Brazilian Portuguese. Secondly, to conduct a prosodic analysis in order to determine the typical intonation contours of telephone numbers as spoken by native speakers. The results of such study may serve as valuable information in improving automatic speech recognition and synthesis systems in application connected to telephone numbers in Brazilian Portuguese.

2. Experiment

2.1. Speech materials

The corpus of the present study consists of read-aloud samples of a total of 30 telephone numbers, listed in Table 1. The numbers were all selected from a local telephone book. The numbers were chosen randomly to cover (a) the landline and mobile telephone numbers with eight digits, (b) the special service numbers with three digits and (c) the toll free number with eleven digits.

In order to test a possible relationship between the graphic displays of numbers and the spoken strategy applied in their utterance, the landline numbers were presented in three different types of graphic displays: (i) into two groups of four digits (NNNN-NNNN), (ii) into a group of three digits followed by two groups of two digits each (NNN-NN-NN) and (iii) a group of eight digits (NNNNNNNN). Telephone books contain numbers with these three types of grouping, although the type (i) is the most common.

The 30 telephone numbers were presented in a slideshow at regular interval of seven seconds. The numbers were recorded from 85 adult, native speakers of Brazilian Portuguese (48 women and 37 men). Speech materials were recorded using a minidisc (Sony, MZ-R700) through a digital microphone (Sony, ECM-MS907) placed at 15 centimetres of participants' mouth. Before the recording, the participants did a quick rehearsal with six numbers. They were instructed to utter spontaneously the numbers that were displayed in a computer screen, and were informed that there is no correct way of speaking telephone numbers and that they were not being evaluated in that task.

Table 1: *List of telephone numbers use in the analyses of this study.*

Telephone numbers			
3 digits	8 digits		11 digits
120	32224034	2226 31 96	08002812112
104	32514251	3221 47 54	08007010114
147	33274686	3271 00 84	08007011566
190	34238577	3428 09 24	08007070044
193	34412276	3465 30 46	08007704418
	3228 6924	8803 91 48	
	3251 7343	9605 36 81	
	3424 2767	9619 94 53	
	3452 1425	9909 62 94	
	3465 2746	9948 09 93	

2.2. Segmental analysis

The spoken telephone numbers were extracted from the recordings and archived in *.wav* format using *Praat* as

previously shown [3]. Each spoken telephone number was segmented into intonation units and transcribed orthographically into grouping represented by the digit 1 (unary), 2 (binary), 3 (ternary), or 4 (quaternary), and into wording represented by the letter U (units), D (tens), C (hundreds) or M (thousands) as exemplified in Figure 1.

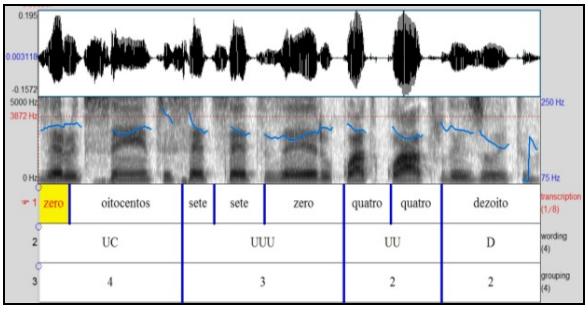


Figure 1: *Praat's textGrid showing the waveform, spectrogram, pitch contour, segmentation and transcription of 08007704418 as spoken by participant woman_2.*

2.3. Intonation contours analysis

Intonation contours of the spoken telephone numbers were extracted through the following method: Firstly, melodic variations of each spoken telephone number were analysed in a semi automatic way with the *Momel/Intsint* script as previously shown [1] in *Praat* (version 5.3.53), using the script' default values (This plugin allows a *Praat* user to convert the pitch contour into a stylized curve, which can be afterwards modified manually according to an auditive evaluation, and finally annotated following the *INTSINT* notation. *INTSINT* is "a transcription system by means of which pitch patterns can be coded using a limited set of abstract tonal symbols, {T,M,B,H,S,L,U,D} (standing for: Top, Mid; Bottom, Higher, Same, Lower, Upstepped, Downstepped respectively)" [8]). Figure 2 shows an example of a telephone number prosodically annotated with the *Momel/Intsint* script.

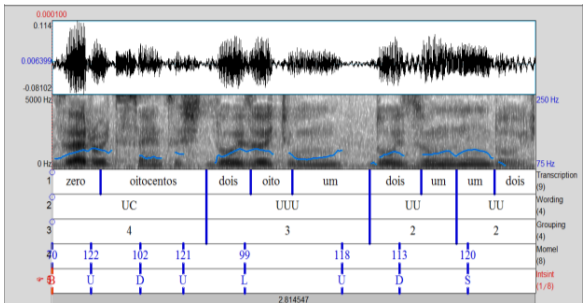


Figure 2: *Praat's textGrid merged showing the waveform, spectrogram, pitch contour, fundamental frequency values, contour variations' coding, transcription and segmentation of 08007704418 as spoken by participant man_69.*

After that, each wording corresponding to a grouping in the spoken telephone numbers was extracted and analyzed with the *Prosodypro* script for *Praat*. In equidistant time intervals, the script was set up to select ten values of fundamental frequency (F0) in the wordings, thereby making it possible to observe adequately intonation contours of wordings of different length through graphic representations. An example is shown in Figure 3.

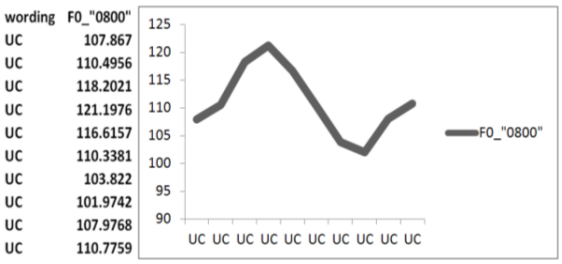


Figure 3: *Graphic representation of intonation contour of "0800" in "08002812112" as spoken by participant man_69.*

3. Results

3.1. Grouping strategy

For eight-digit telephone numbers, the most common grouping strategy is of type "2-2-2-2" in 85% of all cases, which is, four binary groupings.

The grouping strategy with three-digit telephone numbers, in its totality, is of type "3" in 100% of all cases, which is, a ternary grouping.

And with eleven-digit telephone numbers, the grouping strategy is mostly of type "4-3-2-2" in 98% of all cases, which is, a quaternary grouping followed by a ternary and two binary groupings.

The predetermined grouping of the telephone numbers did not seem to influence the grouping strategy adopted by the participants. The grouping strategy of type "2-2-2-2" was preferred in 85% of predetermined grouping of type "NNNN-NNNN", in 80% of predetermined grouping of type "NNNNNNNN" and in 81% of predetermined grouping of type "NNNN-NN-NN".

3.2. Wording strategy

Grouping strategy of type "2-2-2-2" with eight-digit telephone numbers, mostly presented wording of type "$UU_{(1)}$-$UU_{(2)}$-$UU_{(3)}$-$UU_{(4)}$" in 48% of all cases, which is, four binary groupings rendered as units. Other relatively significant wording types are "UU-UU-D-D" in 9%, "D-D-D-D" in 9% and "UU-UU-D-UU" in 9% of all cases.

Grouping strategy of type "4-3-2-2" with eleven-digit numbers, preferentially presented wording of type "UC-UUU-$UU_{(5)}$-$UU_{(6)}$" in 19% of all cases, that is, a quaternary grouping rendered as units and hundreds, followed by a ternary and two binary groupings rendered as units. Other relatively significant wording types are "UC-C-D-D" in 17%, "UC-UUU-D-D" in 16% and "UC-C-D-UU" in 13% of all cases.

Grouping strategy of type "3" with three-digit numbers, mostly presented wording type "C" in 66% of all cases, that is, a ternary grouping rendered as hundreds. Other relatively significant wording type is "UUU" in 34% of all cases.

3.3. Coding of intonation

In spoken telephone numbers with three digits of wording type "C", the most recurrent intonation coding generated by *Intsint/Momel* script in *Praat* is of type "MUD", that is: a Mid

tone, followed by an Upstepped tone and ending with a Downstepped tone.

In telephone numbers with eight digits of wording type "$UU_{(1)}$-$UU_{(2)}$-$UU_{(3)}$-$UU_{(4)}$", the most recurrent intonation coding generated in the first grouping "$UU_{(1)}$" is of type "MUD", that is: a Mid tone, followed by an Upstepped tone and ending with a Downstepped tone. In the second grouping "$UU_{(2)}$", the most generated intonation coding is of type "DU", that is: a Downstepped tone ending with an Upstepped tone. The third grouping "$UU_{(3)}$", mostly, presented the intonation coding of type "UD", that is: an Upstepped tone ending with a Downstepped tone. In the last grouping "$UU_{(4)}$", the most generated intonation coding is of type "UB", that is: an Upstepped tone ending with a Bottom tone..

Lastly, in telephone numbers with eleven digits of wording type "UC-UUU-$UU_{(5)}$-$UU_{(6)}$", the most generated intonation coding in the first grouping "UC" is of type "MUDU", that is: a Mid tone, followed by an Upstepped tone, a Downstepped tone and ending with an Upstepped tone. In the second grouping "UUU" the most generated coding is of type "MD", that is: a Mid tone ending with a Downstepped tone. In the third grouping "$UU_{(5)}$" the most generated coding is of type "UD", that is: an Upstepped tone ending with an Upstepped tone. And in the last grouping "$UU_{(6)}$", the most generated intonation coding is of type "UB", that is: an Upstepped tone ending with a Bottom tone.

3.4. F0 contour

We analyzed variation of the ten selected F0 values at equidistant time intervals in each grouping of spoken telephone number. F0 contours and their patterns were represented.

With eight-digit telephone numbers, Figure 4, 5, 6 e 7 show, respectively, intonation contours of exemplified groupings of type $UU_{(1)}$, $UU_{(2)}$, $UU_{(3)}$ and type $UU_{(4)}$ and their global patterns as spoken by participants.

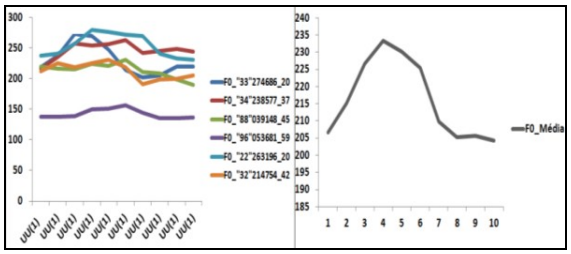


Figure 4: *Intonation contours of exemplified groupings of type $UU_{(1)}$ and their global patterns.*

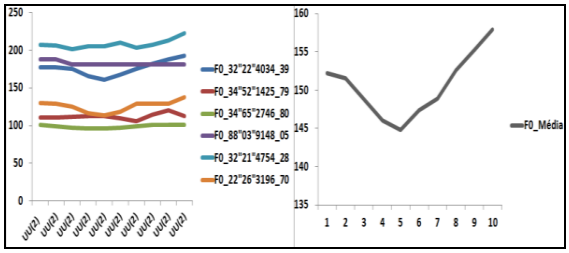


Figure 5: *Intonation contours of exemplified groupings of type $UU_{(2)}$ and their global patterns.*

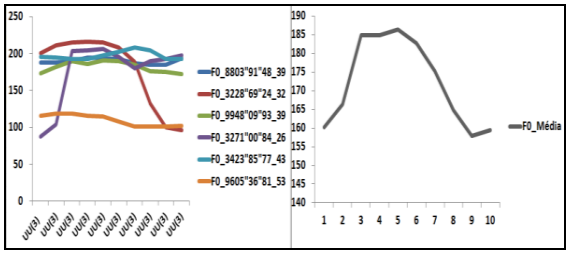


Figure 6: *Intonation contours of exemplified groupings of type $UU_{(3)}$ and their global patterns.*

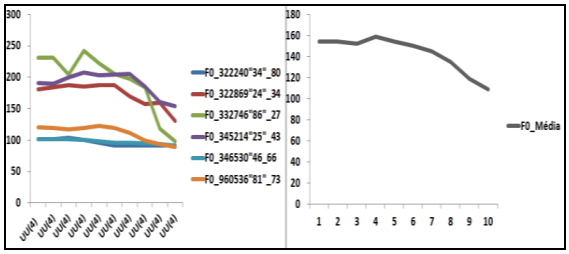


Figure 7: *Intonation contours of exemplified groupings of type $UU_{(4)}$ and their global patterns.*

With eleven-digit telephone numbers, Figure 8, 9, 10 and 11 show, respectively, intonation contours of exemplified groupings of type UC, UUU, $UU_{(5)}$ and type $UU_{(6)}$ and their global patterns as spoken by participants.

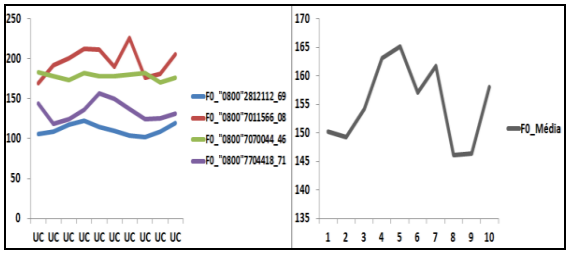


Figure 8: *Intonation contours of exemplified groupings of type UC and their global patterns.*

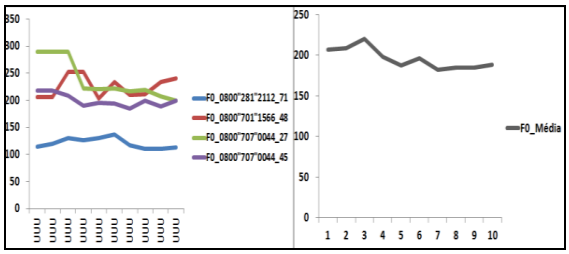


Figure 9: *Intonation contours of exemplified groupings of type UUU and their global patterns.*

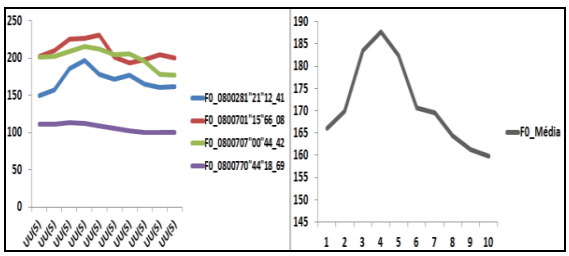


Figure 10: *Intonation contours of exemplified groupings of type $UU_{(5)}$ and their global patterns.*

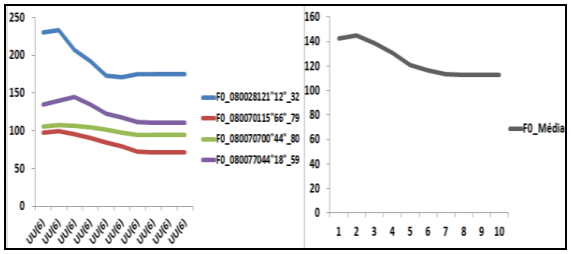


Figure 11: *Intonation contours of exemplified groupings of type $UU_{(6)}$ and their global patterns.*

Figure 12 shows intonation contours of exemplified groupings of type "C" and their global pattern with three-digit telephone numbers as spoken by the participants.

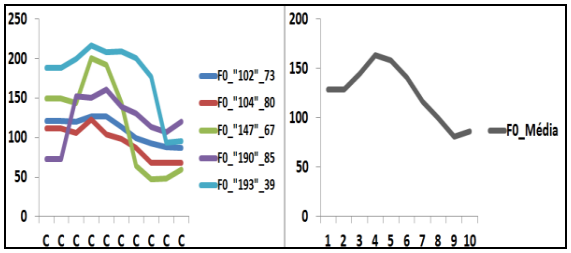


Figure 12: *Intonation contours of exemplified groupings of type "C" and their global pattern.*

3.5. Intonation contour

Results of analyses have shown grouping, wording strategies and intonation contour patterns in spoken telephone numbers in Brazilian Portuguese.

In three-digit telephone numbers, a ternary grouping rendered as hundreds is preferred. As example, the telephone number "190" is spoken as "*cento e noventa*" with the intonation contour shown in Figure 13.

Figure 13: *Intonation contour and corresponding intonation coding of three-digit telephone numbers as spoken by native speakers of Brazilian Portuguese.*

In eight-digit telephone numbers, a four binary groupings strategy rendered as units is preferred. As example, the telephone number "3445 2348" is spoken as "*três quatro*"-"*quatro cinco*"-"*dois três*"-"*quatro oito*" with the intonation contour shown in Figure 14.

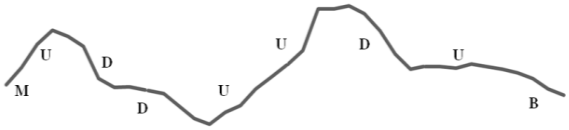


Figure 14: *Intonation contour and corresponding intonation coding of eight-digit telephone numbers as spoken by native speakers of Brazilian Portuguese.*

In eleven-digit telephone numbers, the preferred strategy is a quaternary grouping rendered as units and hundreds, followed by a ternary and two binary groupings rendered as units. As example, the telephone number "08002812112" is spoken as "*zero oitocentos*"- "*dois oito um*"- "*dois um*"- "*um dois*" with the intonation contour shown in Figure 15.

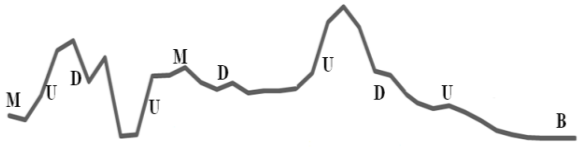


Figure 15: *Intonation contour and corresponding intonation coding of eleven-digit telephone numbers as spoken by native speakers of Brazilian Portuguese.*

4. Conclusion

The main purpose of the present study was to identify intonational patterns of spoken telephone numbers in Brazilian Portuguese. It was demonstrated that intonation alone is enough to characterize prosodically three different types of telephone numbers in Brazilian Portuguese.

A comparison of this study's results with those of [9] shows that the final fall trend reproduce by *Intsint* in Brazilian declarative sentences is also evident in telephone numbers and transcribed by *Intsint*, in this study, with the symbols D (Downstepped tone) and B (Bottom tone) as presented in Figure 13, 14 and 15 respectively. We assume that the intonational patterns of Brazilian Portuguese sentences are used to express the telephone numbers, but the phrasing is adjusted to deliver the sense of grouping.

Other finding of this study is that a similarity in melodic variation exists between telephone numbers in Brazilian Portuguese and those of other languages. The final fall tone of telephone numbers in Brazilian Portuguese is also observed in studies realized on telephone numbers in German [4], Japanese [5] and French [7].

It would of course be interesting to investigate other prosodic features in the characterization of telephone numbers in Brazilian Portuguese, such as intensity and duration. A future investigation involving these features is planned for future analyses. It is also planned to run perceptual tests, based on the intonational patterns identified here in order to find out whether the patterns, applied to synthetic speech, are considered acceptable by Brazilian Portuguese speakers. This research may be regarded as a contribution to the study of spoken numbers in general and to the improvement of automated dialogue systems connected to numbers in particular.

5. Acknowledgements

The completion of this study was made possible with the assistance of *FonUFAL* (Fonetics and Fonology study group of the *Universidade Federal de Alagoas, Brasil*) and the financial support of *CNPq* (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*).

6. References

[1] Hirst, D.J., “A Praat plugin for Momel and Intsint with improved algorithms for modelling and coding intonation”, In Proceedings International Conference on Phonetic Sciences, Saarbrücken, 2007

[2] Xu, Y. (2005-2013), “ProsodyPro.praat”: http://www.phon.ucl.ac.uk/home/yi/ProsodyPro/, accessed on 19 October 2013.

[3] Boersma, P. and Weenink, D. “Praat: doing phonetics by computer. Versão 5.3.53”: http://www.praat.org/, accessed on 20 September 2013.

[4] Baumann, S. and Trouvain, J., “On the prosody of German telephone numbers”, In Proceedings of The 7th Conference on Speech Communication and Technology. Aalborg, Denmark, 2001. P. 557-560.

[5] Amino, K and Osanai, T., “Realisation of the prosodic structure of spoken telephone numbers by native and non-native speakers of Japanese”, In proceeding of The 17th International Congress of Phonetic Sciences (ICPhS XVII). Honk kong, China, August 17-21, 2011.

[6] Bartkova, K. and Jouvet, D., “Selective prosodic post-processing for improving recognition of French telephone numbers”, In Proceedings of The 6th Eurospeech. Budapest, Hungary, 1999.

[7] Martin, P. La prosodie. 2007. Available in: < http://www.linguistes.com/phonetique/prosodie.html>. Accessed on 23 February 2014

[8] Hirst, D.J., Di Cristo, A. and Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. Prosody: Theory and Experiment. Kluwer Academic Press.

[9] Reis, c. & Celeste, l.c. Análise entonativa formal:INTSINT aplicado ao Português. Journal of of Speech Sciences, Vol 2, Iss 2, Pp 3-21 (2012). Available in :< http://www.journalofspeechsciences.org/index.php/journalofspeechsciences/article/view/49/47>. Accessed on 23 Febrary 2014.

Song and speech prosody influences VOT in stuttering and non-stuttering adolescents

Simone Falk[1, 2], *Elena Maslow*[3]

[1] Laboratoire Parole et Langage, CNRS UMR 7309, Aix-en-Provence, France
[2] Institut für dt. Philologie, LMU München, Germany
[3] Institut für Phonetik und Sprachverarbeitung, LMU München, Germany
simone.falk@lpl-aix.fr, elena.maslow@campus.lmu.de

Abstract

Since a long time, it is known that singing helps persons who stutter to produce their utterances more fluently. The prosodic characteristics of spoken and sung utterances differ considerably in their rhythmic and tonal structure. Therefore, it has been proposed that song prosody helps stutterers to improve their rhythmic planning of verbal material [1]. In order to investigate this idea, we examined temporal aspects, namely Voice Onset Time (henceforth, VOT) of voiceless plosives, in sung and spoken utterances of young German stutterers and non-stuttering controls. VOT tends to be reduced in song compared to speech. We expected a more important reduction in the stuttering group as voice onset timing should be facilitated in song compared to speech. Eight stuttering adolescents and eight normal fluent peers read and sang an altered version of "Happy Birthday" with test words containing the three voiceless stops /p/, /t/, /k/. Results showed that stuttering as well as non-stuttering adolescents reduced VOT during singing compared to speech. In contrast, only adolescents who stutter were less variable in their VOT production in song compared to speech. Additional analyses indicated further group differences in vowel duration following the stop consonant. These findings suggest that young stutterers benefit from sung prosody in their timing abilities.

Index Terms: Stuttering, Song and Speech Prosody, Voice Onset Time

1. Introduction

When stutterers sing, their disfluencies can reduce in a substantial way [2, 3, 4]. The reasons for this intriguing phenomenon are still unclear. In the 70ies, Wingate (1969) proposed that stuttering occurs in part because of a rhythmic deficit that can be attenuated by altering prosodic variables of the vocalization [1]. Both parts of this idea, the rhythmic deficit hypothesis [5] and the altered vocalization hypothesis [6] have been pursued in subsequent research. With respect to altered vocalization, it has been found that prolonged phases of phonation are good predictors of fluency in stutterers [7, 8]. In singing, phonation is fostered by prolongation of vocalic and sonorant portions of the speech signal [9]. However, recent research suggests that increased phonation time is not sufficient to explain all fluency-evoking conditions (e.g., paced speech [10]). This raises the question if rhythmic processes could play a more important role. In song, temporal intervals between syllables become more regular and hence, more predictable due to musical beat structure [11]. This could help stutterers by timing segmental material at the right moment during their production [12] and thereby enhance their fluency [1].

In this study, we aim to investigate temporal processes during singing in stuttering children and adolescents. Voice Onset Time is an indicator of fine-grained motor and timing control of laryngeal processes [13]. Stuttering (young) adults tend to show longer and more variable VOTs than non-stuttering adults, even in fluent speech [14, 15]. For children, the situation is less clear. In children of 4 and 9 years on average, no significant differences in VOT were found compared to age-matched peers, but their productions were overall more variable [16, 17].

In this paper, we examine how VOT of voiceless plosives changes in stuttering adolescents during singing compared to speaking, a question that was not addressed so far. An age-group was chosen that was intermediate between the groups tested in previous research (young stuttering children, adults). In general, VOTs are compressed in singing compared to spoken speech [18]. If singing enhances temporal planning and control, stuttering adolescents should benefit from this in their VOT productions. VOTs in spoken speech should be longer and more variable than in age-matched controls. On the other hand, VOTs in singing shouldn't differ from controls as far as timing processes are facilitated by song. Therefore, we hypothesize that VOTs will reduce more in the stuttering group than in controls when comparing spoken vs. sung productions. In addition, several other acoustic parameters such as stop closure, vowel duration and pitch were examined in order to better understand the influence of prosodic differences between sung and spoken utterances in both groups.

2. Method

2.1. Participants

Eight stuttering children aged from 11 to 15 years (M = 12.4, SD = 1.9, 6 males) participated in the study. Their stuttering symptoms ranged from mild to severe on the SSI-3 scale [19]. The participants were tested while attending the therapy course "SAS - Stärker als Stottern" in Starnberg near Munich. The control group consisted of 8 age-matched children (M = 13.0, SD = 2.3, 6 males). The control group was recorded at home or at the Institute of Phonetics and Speech processing, LMU Munich. All participants were native speakers of German and were untrained singers (no choral activity, no singing lessons).

2.2. Procedure

The experiment had two parts. First, participants were asked to read repeatedly a simple text consisting of three sentences which was based on the lyrics of the song "Happy Birthday". Participants were instructed to read the sentences at a

comfortable moderate reading pace. At this time, the participants did not know that the lyrics were derived from the "Happy Birthday" song. In the second part of the experiment, participants sang the text to the melody of "Happy Birthday". Again, they were asked to sing at a moderate tempo that was comfortable to them. The material was presented on handouts and participants had the opportunity to practice their reading and singing with one repetition or stanza before recording. All participants declared to be familiar with the song.

Each participant was recorded separately in a quiet room with the experimenter present who controlled the recording. The participant was seated at a table and had the visual handouts before him / her on the desk. The sung and spoken texts were repeated nine times with different test words per repetition (see Material). Data collection was done with a Beyerdynamic headset (TG H54c) and a H-4N Zoom recorder. The data was recorded at 44100Hz/24bit.

2.3. Material

The three German voiceless stops /p, t, k/ were inserted in bisyllabic nonsense words. Each test word had a trochaic strong-weak accent pattern. The critical stop was in the onset of the first, stressed syllable followed by a long vowel. The test words were /'pi:ta/, /'pa:ta/, /'pu:ta/, /'ti:ta/, /'ta:ta/, /'tu:ta/, /'ki:ta/, /'ka:ta/, /'ku:ta/. These words were inserted in the German version of "Happy Birthday". In the German version, the name of the birthday child is used in every sentence ("Dear X, all the best to you"). The test words were inserted five times in each repetition at the place where the name occurs as demonstrated for /'ki:ta/ in (1).

(1) *Liebe* /'ki:ta/, *viel Glück, liebe* /'ki:ta/, *viel Glück, liebe* /'ki:ta/, *liebe* /'ki:ta/, *liebe* /'ki:ta/, *viel Glück*.

"Dear /'ki:ta/, all the best to you, etc..."

Overall, 15 read / spoken samples were collected for each stop per participant. This resulted in overall 45 × 2 items per participant for analysis.

2.4. Analyses

VOT was measured manually for each plosive by inspecting the oscillogram and spectrogram of the audio signal in Praat [20]. VOT was annotated for each plosive [13]. The time interval was marked between the onset of the stop release and the beginning of vocal fold vibration of the following vowel onset. The second zero crossing of the glottal pulse was consistently used to mark the beginning of voicing. Furthermore, standard deviations of VOT were calculated for each participant in order to assess variability in sung and spoken VOT productions. Some additional measures were taken. First, closure duration was measured from the point where vocal fold vibration of the preceding Schwa-vowel stopped until the onset of the stop release. Second, vowel duration following the stops was extracted from the recordings. The left vowel boundary was determined by marking the beginning of periodic glottal pulses in the oscillogram (second zero crossing). The right boundary was assessed by inspecting the formant patterns and the subsequent spectrum of the consonantal context. Pitch targets were used for the evaluation of vocalic pitch. Maximum pitch per vowel was analyzed using the pitch analysis algorithm in Praat (range: 75-600 Hz) [20]. Furthermore, overall pitch characteristics (median pitch and range) as well as the total duration of each utterance (i.e., one stanza or repetition of the text / song) were assessed. Pauses (e.g., respirations, hesitations, stuttering) were subtracted from the total utterance duration.

3. Results

Test words containing disfluencies or errors of any kind (such as misreading, stuttering) were discarded from analysis. Overall, stuttering symptoms were not very frequent in the test group, and occurred equally often in speech and song (i.e., in 1.6 % of all words in the material). Mean duration and variability for VOT values per stop consonant as well as closure and vowel duration (all vowels confounded) were calculated. Before averaging, outliers were excluded (20 values of closure duration in both groups). These values as well as overall duration and pitch of the utterance and vocalic pitch are displayed for for both groups in Table 1 and Fig. 1.

	Adolescents who stutter		Control group	
Measures	Reading	Singing	Reading	Singing
Utterance				
Mean overall duration (s)	8.25 (1.56)	8.88 (1.63)	7.05 (0.88)	8.94 (0.88)
Median pitch (Hz)	198 (51)	213 (48)	200 (49)	232 (61)
Mean pitch range (st)	11.2 (2.8)	12.2 (3.4)	12.6 (5.9)	13.0 (2.9)
Stop and subsequent vowel				
Mean closure duration (s)	0.093 (0.016)	0.076 (0.011)	0.080 (0.013)	0.067 (0.013)
Mean vowel duration (s)	0.099 (0.033)	0.174 (0.042)	0.115 (0.022)	0.224 (0.039)
Maximum vocalic pitch (Hz)	222 (55)	241 (61)	216 (52)	255 (85)

Table 1. *Overall mean values of prosodic utterance, stop and vowel measures (in seconds (s), Hertz (Hz) and semitones (st)) of the sung and read performance in both groups of participants. Standard deviations are displayed in brackets.*

As can be seen in Fig. 1, sung stops displayed smaller VOT values than read stops in both groups. The VOT data were entered in a three-way mixed-design Analysis of Variance (ANOVA) with the within-subject factors Stop (/p/, /t/, /k/) and Vocalization (reading vs. singing) and the between-subject factor Group (test vs. control). Results showed main effects for Vocalization (F (1, 14) = 62.13, $p < .001$) and Stop ($F(2, 28) = 34.55$, $p < .001$). No interactions or group differences were found. Additional pairwise comparisons confirmed that VOT differed significantly in all three stop classes (with /k/ > /t/ > /p/). This is in line with previous research reporting differences between stop categories [13].

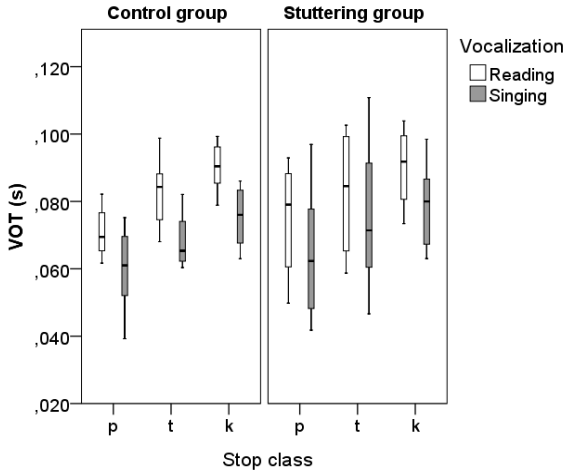


Figure 1. *Boxplots of VOT values of three German voiceless stops in sung and read speech displayed for adolescents who stutter and for the control group.*

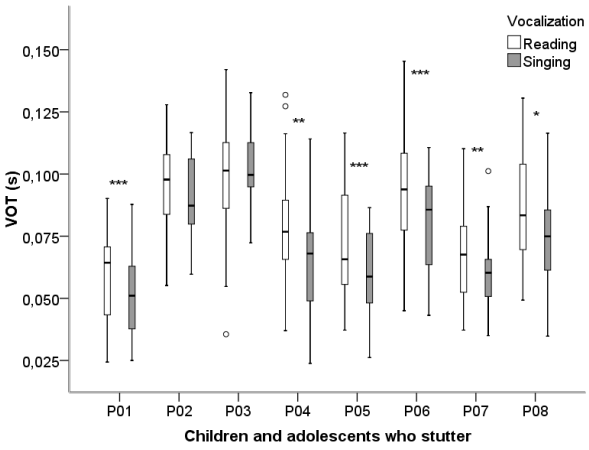


Figure 3. *Individual VOT reduction patterns in read and sung speech in participants who stutter. Stars indicate significance levels (*** $p < .001$, ** $p < .01$, * $p < .05$).*

Another ANOVA with the same factors reported above was performed for VOT variability (the standard deviation of the mean per participant, see 2.4). Main effects were found for Vocalization (F (1, 14) = 19.46, $p < .005$) and Group (F (1, 14) = 6.68, $p < .05$) as well as a significant interaction between both factors (F (1, 14) = 10.77, $p < .01$). The interaction is displayed in Fig. 2. Two-sided t-tests (Bonferroni-corrected) confirmed that adolescents who stutter significantly reduced VOT variability from speech to song (t (7) = 6.12, $p < .001$). This was not the case for the control group.

In order to examine individual VOT patterns, especially in the test group (see Fig. 3), statistical analyses were performed on VOTs for each participant separately (two-sided paired-samples t-tests on VOTs in sung vs. read speech averaged over all three stop classes).

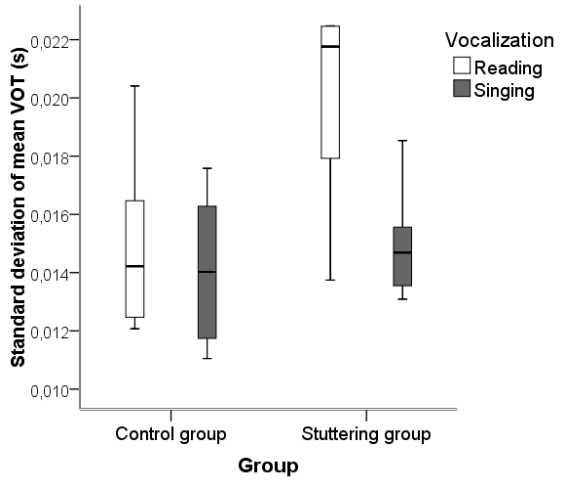


Figure 2. *Variability of VOT in sung and read speech displayed for adolescents who stutter and for the control group.*

Results revealed that VOT reduction was significant only in six out of eight stuttering children. In the control group, all children showed a significant reduction of VOT from speech to song (with at least $p < .01$). In order to examine further temporal differences between speech and song and their realizations by the test and control group, additional statistical analyses were performed on stop closure, vowel and utterance duration and their variability (see Table 1). The same ANOVA design as for the VOT analyses was used except that for vowel and utterance duration, no third factor (Stop) was needed.

For closure duration, a similar reduction pattern from speech to song was found as for VOT (F (1, 14) = 32.58, $p < .001$). A main effect was also found for Stop class (F (2, 28) = 47.72, $p < .001$). Groups did not differ, there was only a marginal trend for the test group showing longer closure than the control group ($p = .098$). Closure variability was also reduced from speech to song (F (1, 14) = 7.63, $p < .05$) with again a marginal trend for the test group to be more variable than controls ($p = .088$). In contrast to VOT and closure duration, vowels were lengthened in song vs. speech (F (1, 14) = 87.15, $p < .001$). Adolescents who stutter displayed overall shorter vowel duration than the control group (F (1, 14) = 4.89, $p < .05$). Moreover, vowel variability was greater in song than in speech in both groups (F (1, 14) = 6.31, $p < .05$). Finally, total utterance duration was longer in song than in speech (F (1, 14) = 14.86, $p < .005$) and less variable in song than in speech (F (1, 14) = 7.66, $p < .05$). Adolescents who stutter merely showed a trend to have longer ($p = .075$) and more variable utterance duration ($p = .063$).

As we did not control for tempo during the recordings between speech and song, we evaluated to what extent VOT or closure duration and reduction and vowel duration and lengthening was related to utterance duration differences. First, we correlated utterance duration with mean VOT and closure duration as well as mean vowel duration per utterance for sung and spoken speech separately. VOT as well as vowel duration, but not closure duration, were positively correlated with utterance duration in both speech (VOT: r (142) = 0.23, $p < .01$; vowel: r (142) = 0.29, $p < .001$) and song (VOT: r (142) = 0.25, $p < .01$; vowel: r (142) = 0.38, $p <$

.001). In other words, longer utterance duration was related to longer VOT and longer vowel duration in both reading and singing.

Second, we assessed if differences in utterance duration were likely to influence VOT reduction. As children and adolescents who stutter tended to show longer utterance duration, but still similar VOT values as the control group, it might be the case that we underestimated group differences without a temporal normalization. In order to answer this question, we calculated difference scores between speech and song by subtracting utterance duration in speech from utterance duration in song for each trial. The same calculation was performed for VOT reduction which was estimated by subtracting sung mean VOTs per trial from read VOTs. A linear regression with the predictor variable utterance difference and the dependent variable VOT reduction was performed. No significant relation was found between utterance duration differences and VOT reduction. Hence, we conclude that the results from the VOT reduction data are independent of utterance duration differences.

4. Discussion

In the present study, we investigated the impact of sung prosody on temporal aspects of young stutterers' verbal productions. In a control group, the VOT of voiceless German plosives was found to reduce significantly in sung speech compared to read speech. This result extends prior research as we show the same result for children and adolescents as done for adult populations [18]. VOT reduction was also found in a group of eight stuttering children and adolescents. Importantly, VOT variability (defined as the standard deviation of the mean) was a good indicator of group differences. VOT variability was significantly reduced during singing compared to reading in adolescents who stutter whereas the control group did not show differences in both tasks.

These findings suggest that singing influenced timing aspects in the present group of young stutterers. First, we demonstrated that VOT reduction occurs as well in stuttering adolescents as in age-matched peers, but against our initial hypothesis, this reduction was not more important in the test group. VOT values were comparable to age-matched peers in sung and, in particular, in spoken productions. This is in line with previous research on VOT in the speech of stuttering children (mean age < 10) [16, 17], but differs from results on adults that showed longer VOT in spoken speech [14, 15]. It is likely that VOT differences only arise consistently at an adult age. Children and adolescents follow individual patterns of development and are in general more variable in their timing capacities [21]. For instance, individual differences were found in our group of adolescents who stutter. Two participants did not show a VOT reduction. Future studies should investigate if the degree of VOT reduction is a predictor of other variables as for example stuttering severity or the efficiency of altered vocalization in reducing disfluencies.

Although VOT duration per se did not allow to distinguish between test group and controls, VOT variability did. Adolescents who stutter were significantly less variable in their VOTs during singing than during reading. It has been shown in previous studies [16], that persons who stutter have more variable VOT in speech production compared to controls. Similar results have been found for perception as well [22]. Our finding shows for the first time that singing has the potential to reduce VOT variability in the productions of adolescents who stutter. This lends support to the hypothesis that singing helps stutterers to improve their timing abilities. Consequently, as suggested by Wingate [1], disfluencies would reduce because rhythmical planning is facilitated. In our sample, disfluencies were rarely occurring in either spoken or sung speech, probably due to the simple and repetitive nature of the German text of "Happy Birthday". A more difficult text/song could be used in future studies in order to observe fluency enhancement during singing in relation to VOT variability reduction.

Further variables were examined in order to better understand the temporal differences between speech and song in our participants. It is well-known that vowel lengthening is very prominent in singing [9] and this tendency was also confirmed by our data. In contrast, previous research on VOT [18] suggests that consonantal parts of the segmental stream undergo a durational compression in song unless consonantal lengthening is part of the phonological system of a language [23]. Our data confirmed compression for plosives by showing the same temporal reduction of closure duration in song vs. speech as found in VOT. While no group differences appeared for closure duration, vowel duration was shorter in participants who stutter. This was unexpected as stutterers tended to show longer utterances which should have led to longer vowel duration. Previous studies have found shorter vowel duration in stutterers' speech [24]. The shorter vowels in our spoken and sung material are probably reflections of this segmental characteristic of stuttering.

In sum, our findings show that singing positively impacts on timing abilities in adolescents who stutter. Future research should address the question to what extent rhythmic characteristics interact with tonal aspects in singing to reduce disfluencies. In a previous study, Glover et al. (1996) discussed the possibility that fluency-enhancement in singing could also arise due to a better representation of global melodic structures, normally not present in speech [6]. For instance, in the domain of neurological rehabilitation, Melodic Intonation Therapy [25] combines both melodic and rhythmic alterations of verbal productions in order to enhance fluency in aphasic patients. In fact, both, rhythm and melody have been shown to impact on production in this task [26, 27]. Future research using neuro-imaging and electrophysiological methods could further unravel the role of melody and rhythm for fluent sung and spoken productions in persons who stutter [e.g., 10, 28].

5. Acknowledgements

The research leading to these results has received funding from a SSHRC-MCRI AIRS (www.airsplace.ca) research grant as well as from the European Union Seventh Framework Program [FP7/2007-2013; FP7-PEOPLE-2012-IEF] under grant agreement n° 327586 to Simone Falk. We thank Thilo Müller and the team of the SAS course for their help with recording the stuttering participants as well as Jonathan Harrington, Phil Hoole, Ulrich Reubold and Florian Schiel of the IPS Munich, Simone Dalla Bella, Euromov, Montpellier, and two anonymous reviewers for their advice and helpful comments.

6. References

[1] Wingate, M.E. (1969). Sound and pattern in "artificial" fluency. *Journal of Speech and Hearing Research*, *12*, 677-686.

[2] Andrews, G., Howie, P.M., Dozsa, M., & Guitar, B.E. (1982). Stuttering: speech pattern characteristics under fluency-inducing conditions. *Journal of Speech, Language, and Hearing Research, 25*, 208-216.

[3] Healey, E.C., Mallard, A.R., & Adams, M.R. (1976). Factors contributing to the reduction of stuttering during singing. *Journal of Speech and Hearing Research, 19*(3), 475-480.

[4] Johnson, W., & Rosen, L. (1937). Studies in the psychology of stuttering: VII. Effect of certain changes in speech pattern upon frequency of stuttering. *Journal of Speech disorders*, *2*, 105-109.

[5] Olander, L., Smith, N., & Zelaznik, H.N. (2010). Evidence that a motor timing deficit is a factor in the development of stuttering. *Journal of Speech, Language, and Hearing Research*, *53*, 867-886.

[6] Glover, H., Kalinowski, J., Rastatter, M., & Stuart, A. (1996). Effect of instruction to sing on stuttering frequency at normal and fast rates. *Perception and Motor Skills*, *83*(2), 511-522.

[7] Colcord R.D., & Adams, M.R. (1979). Voicing duration and vocal SPL changes associated with stuttering reduction during singing. *Journal of Speech and Hearing Research, 22*(3), 468-479.

[8] Davidow, J.H., Bothe, A.K., Andreatta, R.D., & Ye, J. (2009). Measurement of phonated intervals during four fluency-inducing conditions. *Journal of Speech and Hearing Research, 52*(1), 188-205.

[9] Eckardt, F. (1999). *Sprechen und Singen im Vergleich artikulatorischer Bewegungen* [Comparing the articulatory movements in speaking and singing]. Darmstadt: Thiasos Musikverlag.

[10] Stager, S.V., Jeffries, K.J., & Braun, A. R. (2003). Common features of fluency-evoking conditions studied in stuttering subjects and controls: an H215O PET study. *Journal of Fluency Disorders, 28*, 319-336.

[11] London, J. (2004). *Hearing in time*. New York, Oxford: Oxford University Press.

[12] Harrington, J. (1988). Stuttering, delayed auditory feedback, and linguistic rhythm. *Journal of Speech and Hearing Research, 31,* 36-47.

[13] Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical Measurements. *Word, 20*(3), 384-422.

[14] Hillman, R. E., & Gilbert, H. R. (1977). Voice onset time for voiceless stop consonants in the fluent reading of stutterers and nonstutterers. *Journal of the Acoustical Society of America, 61*(2), 610-611.

[15] Metz, D. E., Conture, E. G., & Caruso, A. (1979). Voice onset time, frication, and aspiration during stutterers' fluent speech. *Journal of Speech and Hearing Research, 22*(3), 649-656.

[16] De Nil, L. F., & Brutten, G. J. (1991). Voice onset times of stuttering and nonstuttering children: The influence of externally and linguistically imposed time pressure. *Journal of Fluency Disorders, 16*, 143-158.

[17] Zebrowski, P. M., Conture, E. G., & Cudahy, E. A. (1985). Acoustic analysis of young stutterers' fluency: Preliminary observations. *Journal of Fluency Disorders, 10*, 173-192.

[18] McCrea, C. R. & Morris, R. J. (2007). Effects of vocal training and phonatory task on voice onset time. *Journal of Voice, 21*(1), 54-63.

[19] Riley, G. D. (1994). A stuttering severity instrument for children and adults (SSI-3). Austin: Pro Ed.

[20] Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341-345.

[21] Falk, S., Müller, T., & Dalla Bella, S. (in preparation). Poor sensorimotor timing in children and adolescents who stutter.

[22] Neef, N.E., Sommer, M., Neef, A., Paulus, W., von Gudenberg, A.W., Jung, K., & Wüstenberg, T. (2012). Reduced speech perceptual acuity for stop consonants in individuals who stutter. *Journal of Speech, Language and Hearing Research, 55*(1), 276-289.

[23] Falk, S. (2011). Temporal variability and stability in infant-directed sung speech: evidence for language-specific patterns. *Language and Speech, 54*(2), 167-180.

[24] Howell, P., & Vause, L. (1986). Acoustic analysis and perception of vowels in stuttered speech. *Journal of the Acoustical Society of America, 79*(5), 1571-1579.

[25] Albert, M.L., Sparks, R.W., & Helm, N.A. (1973). Melodic intonation therapy for aphasia. *Archives in Neurology*, *29*, 130-131.

[26] Schlaug, G., Norton, A., Marchina, S., Zipse, L., & Wan C.Y. (2010). From singing to speaking: facilitating recovery from nonfluent aphasia. *Future Neurology, 5*(5), 657-665.

[27] Stahl, B., Kotz, S. A., Henseler, I., Turner, R., & Geyer, S. (2011). Rhythm in disguise: Why singing may not hold the key to recovery from aphasia. *Brain*, *134*(10), 3083-3093.

[28] Toyomura, A., Fujii, T., & Kuriki, S. (2011). Effect of external auditory pacing on the neural activity of stuttering speakers. *Neuroimage*, *57*, 1507-1516.

Some aspects on individual speaking style features in Hood German

Stefanie Jannedy [1], *Melanie Weirich*[2]

[1] Center for General Linguistics Berlin (ZAS), Germany
[2] Friedrich-Schiller-Universität Jena, Germany
jannedy@zas.gwz-berlin.de, melanie.weirich@uni-jena.de

Abstract

Multiethnic urban German (Hood German) as spoken by adolescents in Berlin differs in several significant ways from more standard varieties of Berlin German. It is characterized by a variety of morpho-syntactic alternations and phonetic variants uncommon to the regional standard spoken in Berlin. Previous quantitative corpus analyses have shown that overall speakers of the multiethnic youth style German have a strong tendency to centralize /ɔɪ/ compared to speakers rendering the local regional standard.

This paper now summarizes this centralization tendency and investigates auditory salient realizations of variation by individuals which show tendencies towards a hiatus in the diphthong /ɔɪ/, breaking the nucleus and the off glide. Moreover, there are other prosodic and segmental co-occurring features in the speech of some adolescents which are displayed since it is suspected that some of these may be(come) markers of Hood German.

Index Terms: speaker specificity, diphthong, Hood German

1. Introduction

Hood German [1], a youth style multiethnolect as spoken in larger urban areas of Germany is characterized by several morpho-syntactic alternations such as the 'overuse' of the discourse particle *so* 'like' marking unspecificity ([2]), the use of bare NPs (Hast Du problem? 'Do you have (a) problem?'), lack of prepositions (*Ich gehe Schule*. 'I go (to) school'), the lack of copula verbs (*München weit weg*. 'Munich (is) far away') or the lack of congruency (*mein Schwester hat …* 'my-NOM sister has…') ([3]). There are several qualitative accounts of such alternations from a range of different places in Germany, such as Berlin ([4], [5]; [6]; [7]), Böblingen, Munich, Nürnberg, and Urbach near Stuttgart, ([8], cited in [3]), Frankfurt ([9]), Hamburg ([3]; [10]); and Mannheim ([11]).

Just recently, we have conducted quantitative corpus studies on this youth variety. Data was collected through interviews with adolescents from the Berlin districts Wedding, Kreuzberg and Neukölln. All interviews were orthographically transcribed and then added to the ZAS spontaneous speech corpus on Hood German which is based on the Labb-CAT system [12] . Moreover, data was also collected from Berliners speaking the local standard variety of Berlin German. With the help of this corpus the phonetic realization of /ç/ as [ʃ] or [ɕ] was documented to be a feature of the youth style multiethnic urban variety spoken in Berlin, i.e. Hood German ([13]). This feature was also found to be a stigmatized and highly salient marker of *Hood German* for monoethnic Berliners ([1]). In our perception study ([1]) listeners were asked to categorize identical stimuli (comprising resynthesized fricatives ranging from more /ç/-like tokens to more /ʃ/-like tokens) as either *Fichte* /fɪçtə/ ('spruce') or *fischte* /fɪʃtə/ (1st person sg. 'to fish') depending on the presence of a prime (names of two different neighborhoods of Berlin or no additional information). Results revealed that older listeners categorized more stimuli as *fischte* when they were primed with the name of a multiethnic neighborhood (Kreuzberg) than when primed with the name of a monoethnic neighborhood (Zehlendorf) or no prime at all. Interestingly, younger listeners rated most stimuli as *fischte* in the control condition with no added information. From that, we suggested a potential sound change in progress in terms of the loss of the phoneme contrast between /ç/ and /ʃ/ in Hood German.

Labov's Martha's Vineyard study ([14]; [15]) illustrated that the centralization of the diphthongs /aɪ/ as in 'right' and 'light' and /aʊ/ as in 'cow' and 'loud' serves as a linguistic marker of social and geographical identity. He describes this alternation as regional in character and as a feature of the speech of the people of Martha's Vineyard. He found that this linguistic marker spread from the community of fishermen to the general population of the island and was used by the islanders as a social marker to set themselves off from the economically more powerful tourists and visitors coming to the island for summer vacations.

In addition to the variable realizations of /ç/ as [ʃ] or [ɕ], we also noticed auditory striking diphthong realizations in the spontaneous speech of the Hood German speakers in our ZAS Corpus. Standard German has three diphthongs /aɪ/, /aʊ/ and /ɔɪ/ of which two seem to be realized differently in Hood German. Thus, we have now started to investigate the diphthong realizations of Hood German speakers. In a recent study we have analyzed the F2 patterns in the diphthongs /ɔɪ/ and /aɪ/ to detect potential measurable differences in the spectral patterns ([7]).

1.1. Diphthong Characteristics

In our ongoing analysis, so far, we have investigated F2 patterns from 18 female mono- and multiethnic speakers. From self-identification and the use of other Hood German features speakers in our corpus were categorized in Berlin German speakers and Hood German speakers (note that also monoethnic monolingual German speakers can use Hood German). Approximately 2750 tokens of /ɔɪ/ and /aɪ/ realizations extracted from our spontaneous speech data base have been labeled and analyzed. Formant measurements were taken at 5 equi-distant points throughout the diphthong. The onset and offset of each diphthong was marked off in Praat ([16]). A third value was logged in the middle of the vowel and two more points were logged in between the onset and the middle of the diphthong, as well as between the middle and the offset of the glide. Thus, as the green lines in Figure 1 show, measurements were taken at 0% (start), 25% (early), 50% (mid), 75% (late) and 100% (end) throughout the diphthong.

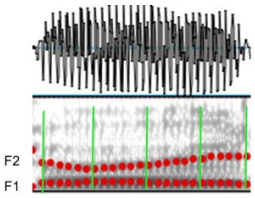


Figure 1: *Formant measurement points during the diphthong /ɔɪ/.*

Results revealed significantly higher F2 values for the Hood German speakers than for the Berlin German speakers at the positions start, early and mid of the diphthong. No such differences were found for /aɪ/ (cf. Figure 2). The centralization of /ɔɪ/ (in terms of a higher F2 in the nucleus of the diphthong) was independent of the speakers' other strong language (Turkish, Arabic, none) and thus, we suggest that it is a phonetic feature of Hood German which serves to express group membership in the multiethnic youth community. Descriptions of the regional Berlin standard ([17]) do not mention such a difference.

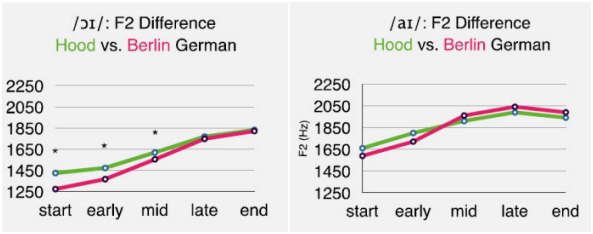


Figure 2: *Average formant patterns during the diphthongs of Hood German (green) and Berlin German (red) speakers.*

In addition, we have been trying to track some linguistic influences on the centralization of /ɔɪ/ in Hood German. Currently, we are investigating the impact of the preceeding and following segmental environment. Preliminary results are in line with the findings obtained by Labov for Martha's Vineyard: obstruents and voicelessness favor centralization ([18]).

For our analysis of the interaction of sentence accent and stress on centralization we were not able to conduct a reliable statistical analysis because only very few words exist that contain the /ɔɪ/ diphthong in a lexically unstressed syllable.

However, in our investigation of the diphthong characteristics of /ɔɪ/, we have been noticing that some speakers of Hood German produce a very different variant of this /ɔɪ/ diphthong. While it sounds centralized to trained ears, there is something else - possibly a hiatic tendency - setting some tokens produced by some speakers off from their other tokens, and from tokens produced by speakers that do not show this variant. A hiatus changes the prosodic structure of the word, adding another syllable nucleus by breaking up a diphthong into its two individual components.

To investigate this more deeply and in a more controlled way, additional recordings were made at a school in Kreuzberg with students, most of them having a multiethnic background and showing features of Hood German. While we also find it desirable to quantify our data and correlate the occurrence of hiatus with social variables, the aim of this first study is not to quantify and parameterize the general spectral and temporal patterns of the diphthong realization in Hood German, but rather to spot auditory salient variants of these diphthongs (and possibly other sounds contained in the speech material) shown by some individual speakers. Of course, individual differences might be due to speaker specific characteristics independent of the speech style within this community. However, we know that individual speakers can act as leaders in their speech community and the once individual features may spread (especially when these features are found in more than one speaker and in several tokens).

2. Method

A set of 33 adolescents from a school in Kreuzberg (Grade 8 and 9; age 13 through 16; 14 male, 19 female) participated in this experiment. Kreuzberg is one of the multiethnic neighborhoods of Berlin where Hood German is spoken by many adolescents. In contrast to our corpus study we used read speech. In this way, the recorded speech data is more controlled. Furthermore, it is well established that there is a register difference between spontaneous and read speech with the latter being more formal and pronounced ([19]). For this reason, we elicited read items containing the diphthong /ɔɪ/, hoping to get more exaggerated versions of the auditory salient, potentially hiatic alternations.

2.1. Speech material

Speakers were asked to read a list of words in the order given on a piece of paper and then repeat the list with the order of words reversed (from bottom to top). Included within this list were five words containing the diphthong /ɔɪ/. In this way 330 tokens of /ɔɪ/ realizations (33 speakers * 5 words * 2 repetitions) were recorded. The word list contained the following tokens:

1. Deutschland (Germany) /dɔɪt͡ʃland/
2. Euro (Euro) /ɔɪʁo/
3. heute (today) /hɔɪtə/
4. Kräuter (herbs) /kʁɔɪtɐ/
5. Kreuzberg (district of Berlin) /kʁɔɪt͡sbɛːɐ̯g/

The selection of words was done so that there is high familiarity for the readers in their daily language use. The words were also selected to allow for comparison with renditions of these lemmas in our spontaneous speech corpus. We deliberately elicited word lists rather than carrier phrases which may have seemed like a larger reading task to the students.

2.2. Analyses

First, both authors conducted an auditory analysis independently of each other. This was done to spot salient realizations of individual tokens. Second, representative items of these perceptually salient tokens were then investigated in more detail in spectral and temporal terms to look for quantifiable acoustic features that characterize these tokens.

3. Results

At this point, we are in the process of describing this youth style variety. We are taking inventory of the differences between Hood German (HG) and the regional standard German (RSG) as spoken in Berlin and we will not provide quantitative analyses of the items containing diphthongs that sound

different to trained ears. At a later point we plan to establish which features are due to individual variation and speaking styles and which may have a wider distribution within the speech community. This will be done by quantifying our observations through corpus studies on data drawn from our spontaneous speech database of Hood German and standard regional Berlin German.

Here, we will concentrate on several speakers which were rated as showing diphthong realizations particularly interesting to the authors. In addition, these speakers also revealed other audibly very salient prosodic features such as tensing of the final unstressed schwa vowel in the orthographic sequence <-er> but also fronting of coronal stops, final stop deletion or dark /ɬ/ coloring as expressions of their individual speaking styles. It is noteworthy, that not every speaker reveals all features but that speakers use this repertoire of features to draw from.

3.1. Diphthong realizations

3.1.1. Speakers HGm1 and HGm2

The four renditions of the /ɔɪ/ diphthong in the word *heute* 'today' shown in Figure 3 were produced by two 16 year old male Hood German speakers HGm1 (upper panel) and HGm2 (lower panel).

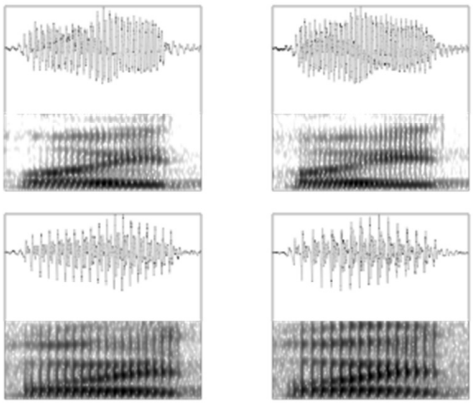


Figure 3: *Spectrograms of the diphthong /ɔɪ/ in the word heute 'today' by two 16 year old male speakers HGm1 (upper panel) and HGm2 (lower panel).*

The boys have very similar multiethnic backgrounds, yet, their /ɔɪ/ diphthong sounds rather different. The speaker whose speech is depicted on the top panel sounds to the authors, as if there are differences in timing compared to the regional standard variety spoken in Berlin.

The bottom panel shows a spectrogram by a speaker where we did not get these auditory impressions and who we believe adheres to the regional standard with regard to this feature.

There seem to be two spectrally observable differences between these productions of /ɔɪ/ for these two speakers: 1. Speaker HGm1 (top panel) produces a much greater discontinuity between the nucleus and the offglide compared to HGm2. There is spectral evidence for this in the waveform and in the spectrogram accompanied by an amplitude change visible in the waveform. Moreover, for HGm2, the F2 rises more steadily and gradual whereas for the first speaker, the F2 remains somewhat low for about one third of the diphthong and then rises.

Secondly, speaker HGm1 reaches the maximum F2 earlier within the diphthong than speaker HGm2. For most tokens of these speakers that we have in our data, these qualitative observations seem to hold.

3.1.2. Speakers HGf1 and HGf2

The diphthongs in Figure 4 were also excised from the word *heute* as produced by two 14 year old female Hood German speakers (HGf). Both renditions produced by speaker HGf1 (upper panel) resemble each other acoustically and leave the auditory impression of almost two syllables being produced rather than a monosyllabic diphthong. This is very similar to HGm1 (upper panel Fig. 3) who also showed hiatic tendencies. Just as the diphthongs by speaker HGm1 , the renditions on the top left of Fig. 4 also reveal some discontinuities. In addition, both tokens by HGf1 reach a high F2 at an early stage throughout the diphthong.

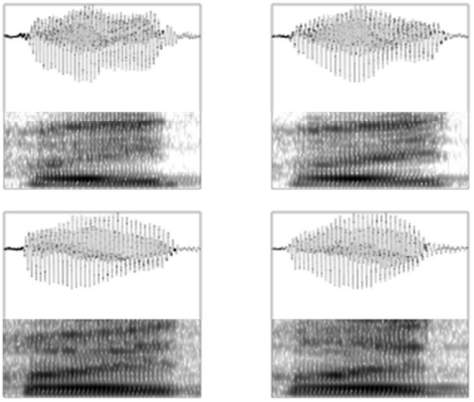


Figure 4: *Spectrograms of the diphthong /ɔɪ/ in the word heute 'today' by two 14 year old female speakers HGf1 (upper panel) and HGf2 (lower panel).*

Note that for HGf1, the energy distribution in the diphthong is different compared to that of the boys (Fig. 3). There is more energy in the initial nucleus part compared to the offglide. It is also evident through the initially raised F2 that the diphthong is strongly centralized compared to the regional standard variety ([7], [18]). Here, too, the discontinuity in the amplitude of the waveform can be observed; however, F2 seems to rise more continuously in the right panel compared to the left.

The spectrograms on the bottom of Fig. 4 show renditions of the diphthong /ɔɪ/ by HGf2 who seems to have a more equal distribution of energy throughout the diphthong. While her first iteration sounds rather hiatic, the second one sounds like regional standard German. A possible explanation is that the F2 maximum is reached somewhat earlier in the diphthong in rendition on the left which then leads to a noticeable difference in the realization of the diphthongs, best described as semi-hiatic.

3.2. Other features

There are several other features in the speech of these adolescents that are worth exploring further to establish an inven-

tory of differences between regional standard German and Hood German.

3.2.1. Tensing of final <-er>

While some features are rather segmental in nature, pertaining to the quality of the segment, one prosodic difference that we found to be rather salient was the tensing and change of quality and duration of the word final unstressed syllable as orthographically represented by <-ter> in words such as *Kräuter* (herbs) /kʁɔɪtɐ/ [kʁɔɪta]. The displays in Fig. 6 show examples of this observation by the two 14 year old female Hood German speakers HGf1 and HGf2.

HGf1 on the left emphasizes the final syllable which is untypical for regional standard German. The exact nature of this emphasis is neither well described nor well understood yet. However, perceptually, it is rather salient. It resembles what is expressed in US-American English as ´gansta´ for *ganster* and may as well originate from musical rap culture or possibly from Turkish language influence. HGf2 on the right uses the regional standard pronunciation.

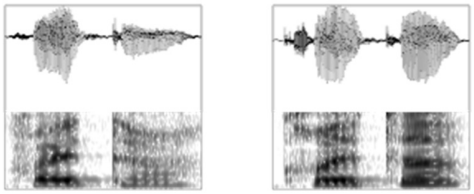


Figure 5: *Spectrograms of the word Kräuter (herbs) by the female speakers HGf1 (left) and HGf2 (right).*

Also notable here is that HGf2 on the right has a much stronger onset of the /kʁ/ sequence which we also perceive as rather typical for Hood German and which we observed with other speakers, too. The perceived quality of the uvular/velar German /ʁ/ strongly resembles that of a velar fricative /x/.

3.2.2. Dental release of /t/

The frontal release of word medial /t/ in *heute* 'today', we also observe to be perceptually very salient. This may be spectrally evidenced, too, by a higher mean burst frequency ([20]) that is indicative of a more frontal release.

Figure 6 shows two renditions of the final syllable /tə/ of *heute* by the 14 year old female speaker HGf1. The one on the left is realized more alveolar while the one on the right is produced more dental.

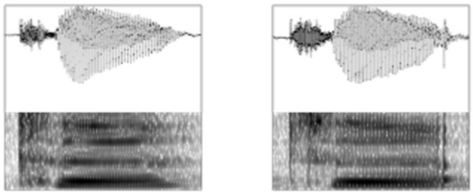


Figure 6: *Spectrograms of the final syllable /tə/ in the word heute 'today' by the 14 year old female speaker of Hood German HGf1.*

This is not only audible but also reflected in the different energy distribution of the /t/ burst. Notable also is the glottal stop at the end of the final unstressed schwa in the second panel. This, however, seems to be an idiosyncrasy of the speaker more so than a general pattern observable across different speakers.

3.2.3. Coloring of /l/

The regional standard variety of German spoken in Berlin does not usually have a dark /ɬ/ quality. However, several adolescent speakers realize the /l/ as [ɬ] in our word lists, in the onset of a syllable before a back vowel. The auditory dark lateral is produced with a more retracted tongue position and therefore spectrally reflected by a lowered F2 value and a higher F1 formant configuration.

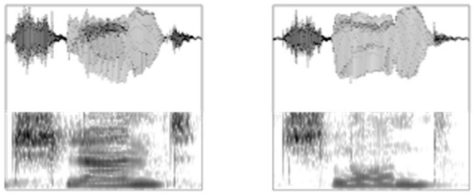


Figure 7: *Spectrograms of the sequence /t͡ʃland/ excised from the word Deutschland (country) by the female speakers RSGf1 (left) and HGf2 (right).*

The velarization of /ɬ/ is visible in the right spectrogram, compared to that seen in the left display of Fig. 7 with a light /l/ which has a relatively high F2 and low F1 ([21]). The spectrogram on the left is based on the speech of a regional standard speaker (RSGf1) who does not have this /l/ coloring. While this feature may originate from speakers with a Turkish cultural and ethnic heritage born in Germany, it is also being used by monolingual monoethnic Hood German speakers.

4. Discussion

In Hood German, we noticed prosodic and segmental deviations from the regional standard of Berlin German. They include the timing patterns of the diphthong /ɔɪ/ and the tensing and lengthening of phonologically unstressed final orthographic <-er> which in German goes along with changes in duration. We also observed the dental release of /t/ and the realization of a dark /ɬ/ in Hood German in syllable onset position before low back vowels. While none of our speakers displayed all features, some speakers show several idiosyncrasies at the same time. Thus, it seems that adolescent speakers of Hood German chose from an inventory of markers and features to create their own individual speaking style within this urban variety of German.

If however, these are reliable features of Hood German or markers of individual speaking styles remains to be seen as we are now at a stage taking inventory of the phonetic/phonological alternations evident in this youth style multiethnolect. We will take our results from the reading task (word lists as a more formal register) and compare it to iterations found in our spontaneous speech data base.

Nevertheless, the differences in realizations of the diphthong ranging from a hiatus like /ɔ.ɪ/ to the diphthong /ɔɪ/ (which is canonical in the regional standard) are variable within speakers and also across speakers. At this point, we are still investigating the acoustic properties that can best account for this observation and which may as well be individual expressions of speaking style.

We suspect that this is a new feature of Hood German as spoken in Berlin which will eventually spread through the community as already now, it is one of the most prevalent markers of the as of yet under described phonetic/phonological features.

5. Acknowledgements

This work was supported by the Federal German Ministry for Education and Research under Grant No. 01UG1411. Much gratitude is due to our research assistants Iona Gessinger, Luisa Helmeke und Sophie Arndt.

6. References

[1] Jannedy, S. & Weirich, M., (2014) Sound Change in an Urban Setting: Category Instability of the Palatal Fricative in Berlin *Journal of Laboratory Phonology* 5(1):91-122.

[2] Jannedy, Stefanie (2010) The Usages and Meanings of 'so' in Spontaneous Berlin Kiezdeutsch. In Melanie Weirich & Stefanie Jannedy (eds.) *ZAS Papers in Linguistics (ZASPiL) 52*, 43-61.

[3] Auer, P. (2003) Türkenslang: Ein jugendsprachlicher Ethnolekt des Deutschen und seine Transformationen. In A. Häcki-Buhofer (Hrsg.), *Spracherwerb und Lebensalter,* Tübingen: Francke, 255 – 264.

[4] Wiese, Heike (2009) Grammatical innovation in multiethnic urban Europe: new linguistic practices among adolescents. *Lingua* 119: 782–806.

[5] Wiese, Heike (2012) *Kiezdeutsch. Ein neuer Dialekt entsteht*. München: C. H. Beck.

[6] Jannedy, Stefanie (2012) Urbanes Deutsch und seine Rezeption. In *Jahrbuch der Geisteswissenschaftlichen Zentren Berlins (GWZ). Bericht über das Forschungsjahr 2011*, 74–95.

[7] Jannedy Stefanie, & Weirich, Melanie. (2013) /ɔɪ/ as an identity marker of Hood German in Berlin. *Proceedings on Meetings on Acoustics* (POMA) 19(060096).

[8] Füglein, Rosemarie. (2000) *Kanak Sprak. Eine ethnolinguistische Untersuchung eines Sprachphänomens im Deutschen*. Diplomarbeit, Fakultät für Sprach- und Literaturwissenschaften der Otto-Friedrich-Universität Bamberg (unpublished).

[9] Tertilt, Hermann. (1996) *Turkish power boys*. Ethnographie einer Jugendbande. Frankfurt: Suhrkamp.

[10] Dirim, İnci, & Peter Auer. 2004. *Türkisch sprechen nicht nur die Türken. Über die Unschärfebeziehung zwischen Sprache und Ethnie in Deutschland*. Berlin: De Gruyter.

[11] Keim, Inken (2008). *Die "türkischen Powergirls"—Lebenswelt und kommunikativer Stil einer Migrantinnengruppe in Mannheim*. Studien zur deutschen Sprache (2nd ed.), 39. Tübingen: Narr.

[12] Fromont, R. & Hay, J. *LABB-CAT* formerly known as Onze-Miner (http://onzeminer.sourceforge.net/)

[13] Jannedy, Stefanie, Melanie Weirich, & Jana Brunner (2011). The effect of inferences on the categorization of Berlin German fricatives. In *Proceedings of the International Congress of Phonetic Sciences* (ICPhS) 14, Hong Kong, China, 962–965.

[14] Labov, W. (1972a). The social motivation of a sound change. In Labov, W. *Sociolinguistic patterns*, 1–42. Philadelphia: University of Pennsylvania Press.

[15] Labov, W. (1972b). The social setting of linguistic change. In Labov, W. *Sociolinguistic patterns*, 260 – 325. Philadelphia: University of Pennsylvania Press.

[16] Boersma, P. and Weenink, D. (2012) *Praat: doing phonetics by computer* [Computer program]. Version 5.3.23, retrieved August, 7th, 2012 from http://www.praat.org/

[17] Schönfeld, H. & Schlobinski, P. (1995) After the Wall: Social Change and Linguistic Variation in Berlin. In Stevenson, P. (ed.) *The German Language and the Real World* Ch. 6, 117-134.

[18] Jannedy, S. and Weirich, M. (accepted) *Linguistic Influences on Diphthong Realization of /ɔɪ/ in Hood German*. ISSP, Köln.

[19] Labov, W. (1986) *The Social Stratification of English in New York City.* Center for Applied Linguistics, Washington D.C.

[20] Sundara, M. (2005) Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *J. Acoust. Soc. Am*. 118:2, 1026 –1037

[21] Yuan, Jiahong & Liberman, Mark (2009) Investing /l/ variation in English through Forced Alignment. *Proceedings of Interspeech, Brighton*, UK, ISCA, 2215 –2218.

Automatic extraction of the general tendency of F0 patterns Cross linguistic study on laboratory data

Katarina Bartkova, Mathilde Dargnat

University of Lorraine, ATILF, France
katarina.bartkova@atilf.fr, mathilde.dargnat@atilf.fr

Abstract

The goal of our study is to use an automatic approach to extract the general prosodic tendencies of the speech signal conveyed by the F0 pattern. The speech signal is prosodically annotated by an automatic prosodic transcriber and then prosodic patterns are extracted from this annotation. The pertinence of the pattern extraction is tested here on laboratory data containing isolated sentences in French and English uttered by native and non-native speakers. An analysis of the extracted parameters shows how the prosody of the sentences is defined by their shared syntactic structures and also indicates to what extent the prosodic features used by the two languages are similar or different. It appears from the analyzed data that the extraction of parameters via automatic processing can yield relevant information for a cross-linguistic study of prosody.

Index Terms: prosodic annotation, automatic pattern extraction, native & non-native prosody

1. Introduction

The use of an automatic approach for prosodic annotation of speech is useful, especially as agreement on manually annotated prosodic events (boundary levels, disfluences and hesitation, perceptual prominences) between expert annotators is quite low [15]. When manual coding of pitch level is carried out, there is the risk that human annotators can be influenced by the meaning of the speech. Moreover a human transcriber may be also influenced by what he considers to be the norm, thereby standardizing the transcription of prosodic phenomena and ignoring the reality of the speech signal.

A further advantage of automatic processing is that, once the values of the parameters are normalized, they are then compared to the same threshold value. This is difficult to achieve with manual annotation because of the inherent subjectivity of this approach.

The goal of the present study is to extract relevant prosodic tendencies of the F0 pattern. This approach is then tested in a cross-linguistic study of speech prosody in French and English.

2. French & English prosody

Many studies have described the specificities of French and English prosody. According to these studies, French uses a combination of segmental and tonal cues to signal prosodic phrases, and differs in this respect from a language like English, which relies almost exclusively on tonal boundaries [7] [14]. In French, lexical stress is mostly quantitative [8], and the final syllable is the one which undergoes a potential lengthening. However, lengthening of the last syllable of the word corresponds also in French to final (pre-boundary) lengthening, which concerns rhythm, and is not an accentual lengthening as in English [6

French is generally considered as a language with mostly 'rising' F0 patterns [12] accompanied by a lengthening of final syllables [20]. French prosodic phrasing was described by Delattre's functionalist approach [9] Though extended by more recent studies [11], [13], [10]. Delattre's work still remains seminal for studies on French prosody. In French spontaneous speech data, a melodic rise is generally produced at the end of the clause. It indicates that the clause is an unfinished constituent at the discourse level, and that it can be associated with the term of "**major**" or "**minor**" **continuation contour**, according to Delattre's approach.

French and English intonation are sometimes described by a set of contours. Delattre [9] considers that 10 basic contours can describe the most frequent intonation patterns in French; [18] also distinguish 10 contours though their contours differ from those proposed by Delattre. As far as English is concerned, 22 pertinent intonation contours are proposed by [17] to describe English intonation.

It is common to use the term of *assertion intonation* or *question intonation* to refer to falling or rising contours: falling contours are associated with assertion or assertiveness (Bartels 1999), whereas rising contours are associated with questions or aspects of questioning (uncertainty, ignorance, call for a response or feedback from the addressee, etc.). Although prototypical assertions are uttered with a falling contour and prototypical confirmation or verifying questions are uttered with a rising contour, occurrences of assertions with a rising contour and occurrences of confirmation or verifying questions with a falling contour are far from rare in everyday conversations [4].

In the following paragraphs F0 contours in French and English sentences spoken by native and non-native speakers are measured and compared and their differences are statistically evaluated.

3. Prosodic annotation

Prosodic parameters are subject to parameter values governing prosodic coherence along the prosodic group. It was observed in automatic speech processing (in diphone and data driven speech synthesis) that a sudden unjustified change in F0 or sound duration (beyond stressed syllables or prosodic junctures), is perceived either as a corruption of the speech signal or as an occurrence of a misplaced contrastive stress [5]. Most of the time researchers focus on the transcription of parameter values of syllables considered as linguistically prominent, carrying pertinent linguistic information. The other linguistically non prominent syllables, remain generally uncoded although their prosody contributes to an overall perception of a correct pattern. Therefore we believe that in order to keep a faithful prosodic transcription of the speech signal, all the parameters of the syllables should be annotated.

3.1. Prosodic labelling

Speech data processing was carried out in several stages. First, prosodic parameters were extracted from the speech signal. In order to segment the speech data, a text-to-speech forced alignment was carried out using the CMU sphinx speech recognition toolkit [16] yielding an automatic segmentation of the speech signal at the phoneme level. This automatic segmentation of the speech signal was then manually checked by an expert phonetician.

For the F0 pattern analysis, F0 values in semitones were estimated every 10 ms by the software Aurora [19]. A simple F0 parameter smoothing was carried out by our annotation software to eliminate corrupted F0 values.

Prosodic annotations were yielded by the language independent automatic annotation tool PROSOTRAN [2]. This tool requires no specific linguistic knowledge, therefore it is well-adapted for cross-linguistic studies. PROSOTRAN yields various numeric and symbolic prosodic annotations for each syllable of the speech signal; however, from this data, only F0 range values and sound durations are used in this study. Sound duration is normalized and transformed to a symbolic duration annotation. For the representation of F0 patterns, a melodic range is calculated between the maximum and the minimum values of the F0 in semi-tones. All speech material for each speaker is used to build a histogram of the distribution of the F0 values. To avoid extreme, often wrongly detected F0 values, 6% of the extreme F0 values (3% of the highest and 3% of the lowest ones) are discarded. The resulting range is then divided into several zones (9 in our case) and is coded into levels (from 0 to 9). By calculating FO in this way, value normalization is enabled and also inter-speaker comparison of FO patterns.

3.2. Corpus

The corpus used in this study was recorded as part of the Intonal project, focusing on the study of intonation in French and English. The recorded corpus contains 40 short sentences belonging to 8 syntactic categories using 20 French and 20 English native speakers. The French speakers uttered French and English sentences, and constitute our non-native English speaker group.

The corpus sentences contain sentences with two kinds of non-conclusive F0 slope configurations as well as interrogative and declarative sentence final configurations. Our study analyses mainly the F0 contours on discourse level (the F0 value of the final segment of declarative clauses connected by a discourse relation, marked or not by a conjunction) and on syntactic level (F0 pattern on the final segment of declarative and interrogative sentences).

3.2.1.1 F0 tendency extraction

The goal of the F0 pattern extraction is to get the most representative F0 pattern(s) for a given sentence for a group of speakers keeping one or, if necessary, several F0 values per syllable (Figure 1).

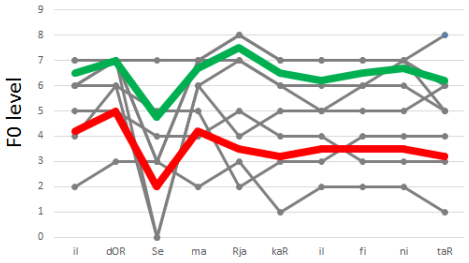


Figure 1: *Representative patterns (in green and red) of F0 levels for a sentence uttered by 8 speakers (in grey)*

The symbolic annotation yielded by PROSOTRAN is used for each syllable and for the whole speaker group of each language to identify the values that represent the general tendency of the F0 pattern. An empirical approach was adopted allowing the emergence of maximally two F0 values per syllable (as the number of speakers is relatively small in each group – maximally 18 for English native speakers). The F0 values coded by their range level are split into two groups using an adjusted median value keeping F0 values belonging to the same symbolic code in the same group. This way the division of the following symbolic F0 values [10 9 9 8 8 8 7 7 5 6] occurs after the last "8" (the 6th F0 value and not after the 5th value as expected). Each grouping obtained is represented by a mean value (V1 & V2, cf. Figure 2) and the two F0 values per syllable are maintained only if their difference is higher than 3 semi tones and when the number of F0 values in a group is higher than 2. If not, the groups are merged and a general mean value (V(1,2)) is calculated using the values of all the speakers for a given syllable.

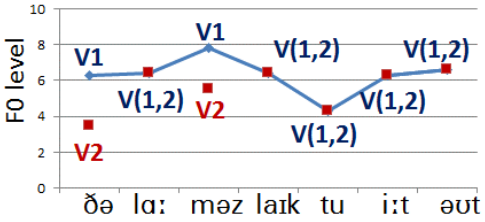


Figure 2: *F0 pattern general tendency: V1 first F0 value, V2 second F0 value, V(1,2) first and second F0 values merged*

The number of merged values was 70% for the French native speakers, 74% for English natives but only 50% for non-native speakers. The F0 patterns of non-native speakers were less consistent and have more variability in their pronunciation.

For the different sentences, the succession of the F0 values is recovered and the preferred F0 pattern tendency observed. For example, for the sentence in figure 2, the preferred tendency as to the succession of the F0 values is represented on Figure 3.

6.3 (V1) 6.4 (V1) 7.8 (V1) 6.4 (V1) 4.3 (V1) 6.3 (V1) 6.6 (V1)	**[4]**
3.5 (V2) 6.4 (V1) 5.5 (V2) 6.4 (V1) 4.3 (V1) 6.3 (V1) 6.6 (V1)	**[2]**
6.3 (V1) 6.4 (V1) 5.5 (V2) 6.4 (V1) 4.3 (V1) 6.3 (V1) 6.6 (V1)	**[2]**

Figure 3: Succession of the F0 range values and their codes (between parentheses) and the number of the pattern observed (between brackets)

3.3. Analysis of results

Our approach of automatic extraction of F0 tendency is tested on 4 sentence types from our laboratory data; that is on continuative, paratactic, interrogative and declarative sentences. Each sentence group contains 5 different sentences

of the same syntactic structure uttered by a group of at least 8 speakers - that is a corpus of about 160 sentences. In the following paragraphs only results obtained for one sentence (containing about 8 utterances) per sentence type will be discussed, however the results obtained for the remaining sentences of the same sentence type obtained very similar results.

Continuative sentences: *(two clause sentence, with coordinating conjunction, ("He has seen **Maria** because he has come" "Il va voir **Maria** car il en a envie").* (cf. Figure 4).

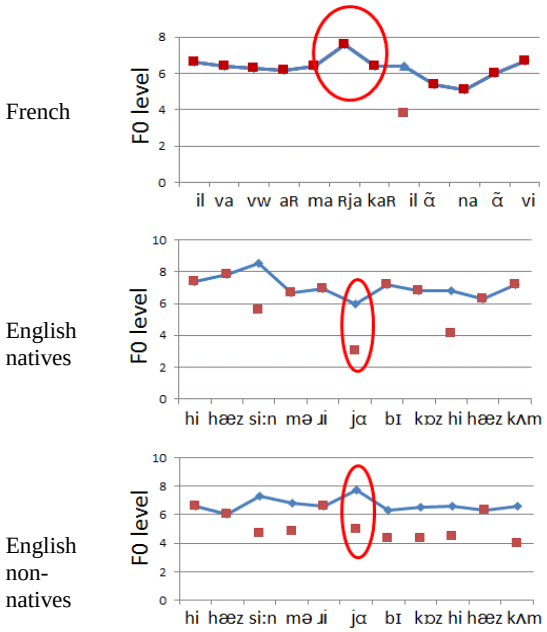


Figure 4: *F0 pattern in a continuative sentence ("He has seen **Maria** because he has come" "Il va voir **Maria** car il en a envie"); red circle: major prosodic boundary*

French speakers marked the continuation (red circle on the figure) with a rising F0 while English speakers prosodically coded the same syntactic boundary with a lowering F0. Non-native English speakers use more rising patterns than falling ones. In French, the general rising tendency of the F0 is not very high but the prosodic boundary is also indicated with lengthened vowel duration. On the other hand, the downwards movement of the F0 in English was more important but there is no vowel lengthening in the final syllable. A high prosodic agreement in these sentences is in the realization of the major prosodic boundaries: the rising tendency on the NP boundaries is respected by the majority of the French speakers and the falling F0 pattern by the majority of English speakers. In the non-native group there is little agreement as to the F0 pattern on major prosodic boundaries; in fact most of the time two F0 values are extracted: a high value indicating a rising F0 movement and a low value indicating a falling F0 movement.

Paratactic sentences: *(two clause sentence, without coordinating conjunction, "Il dort chez **Maria**, il va finir tard. / He'll sleep at **Maria's**, he'll finish late.")* (cf. Figure 5)

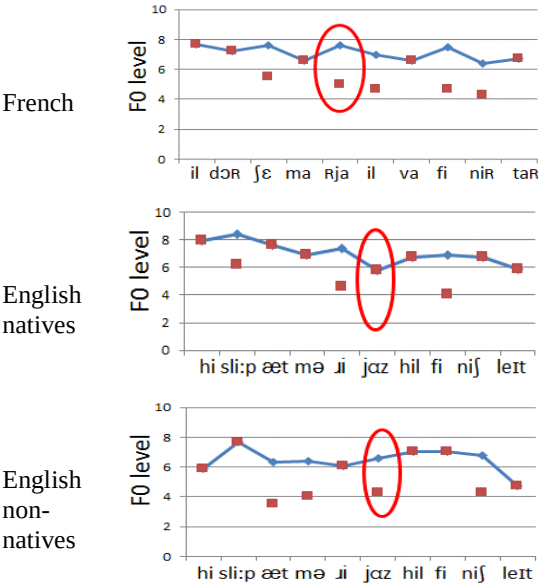
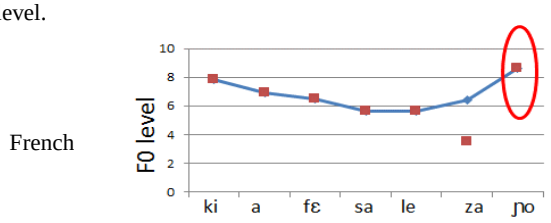


Figure 5: *F0 pattern in a paratactic sentences ("He slips at **Maria's** he'll finish late" "Il dort chez **Maria**, il va finir tard"); red circle: major prosodic boundary*

In French little prosodic agreement is found on the major non-final prosodic boundary (red circle on the figure): the F0 level fluctuated and most of the time two F0 level values are extracted. In these sentences the speaker's prosodic production of the first clause is similar to a final prosody signaling the end of the first clause in a sentence and its syntactic independence from the second clause. However, a strong agreement in the F0 values is observed in English native data where all the clause final F0 values are falling. As far as the non-native English group is concerned, their F0 realizations are again closer to the French group than to the English group.

Interrogative configuration *(simple subject NP: "Qui a fait ça? **Les agneaux**? / Who did this? **The lambs**?")* (Figure 6)

In French interrogative sentences, generally, a huge level rise is preceded by a rather flat F0 level. The F0 pattern in English interrogative sentences contains a more moderate F0 upward movement or (Figure 6) a lowering F0 movement (despite the interrogative character of the sentence). As the interrogative character of the sentence is also expressed by question words ("who" in this example), there is probably no real need for prosodic marking. Again, the French speakers of English are closer with their F0 pattern realizations to French prosody, (rising F0 values) than to English prosody. Other noteworthy findings for this sentence type: the interrogative character is prepared from the beginning (onset) of the sentence: for the 3 speaker groups the interrogative sentences start at a high F0 level.



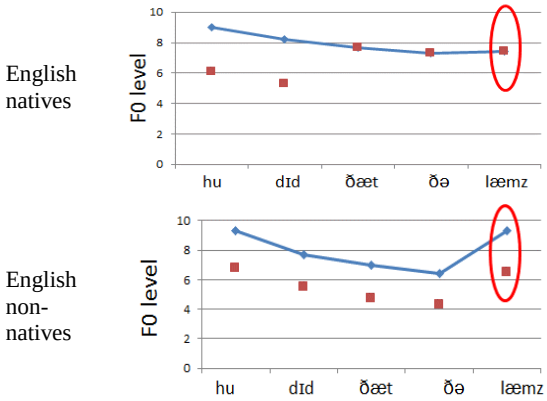


Figure 6: *Interrogative sentence types ("**The lamas** like to eat oats" "**Les lamas** aiment bien l'avoine"; red circle: interrogative sentence final boundary*

Declarative sentences *(Longer declarative sentence: "Il dort chez **Maria**. /He'll sleep at **Maria's**")* (cf. Figure 7)

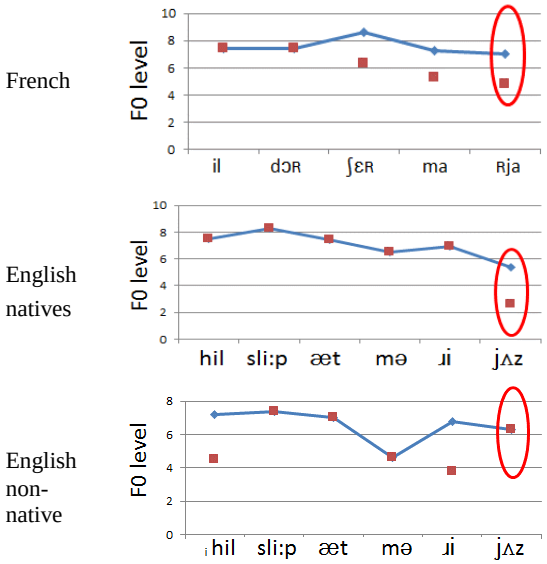


Figure 7: *F0 pattern in a sentence with continuative configuration on a subject NP ("**The lamas** like to eat oats" "**Les lamas** aiment bien l'avoine"); red circle: final declarative boundary*

In French the preferred F0 pattern at the end of the declarative sentence types is only slightly falling or it remains flat. This finding can be partly explained by the fact that, in French, the final pattern of a declarative sentence is also marked by the last syllable duration lengthening which is more moderate than the syllable duration lengthening on major continuation prosodic boundaries [1]. In English, on the other hand, the end of the declarative sentences is marked by a systematically falling F0 pattern. As for the non-native speaker group, their F0 pattern is also either slightly falling or remains flat i.e. they are closer to the native French group F0 patterns than to the native English group patterns.

3.4. General discussion

From the previous analysis some general tendencies can be identified as to the differences between French and English F0 patterns. In the phrases studied here French speakers gave preference to more rising F0 patterns on prosodic boundaries while English native speakers uttered the English version of the sentences with a preferential falling F0 pattern. The final F0 movement in the sentence is falling in both languages (cf. Table I), however the slope of the F0 level change is steeper in English than in French.

The English non-native speaker's prosody remains somehow influenced by both French and English prosody: for example in interrogative sentences their F0 pattern is clearly similar to French, while in continuative sentences the non-native's F0 pattern is more similar to English (cf. Table I).

Table I: *Amount of F0 levels in rising (+) or falling (-) patterns on major prosodic boundaries for the 4 sentence types*

Sentence type	French	English native	English non-native
Continuative	+2.6	-1.5	-1.5
Paratactic	+1.1	-1.5	-0.5
Interrogative	+3	+0.5	+2.5
Declarative	-1.8	-3	-1.5

The intra-speaker variability of the F0 levels (2 F0 level values per syllable) of the native (French & English) groups occurs more often in linguistically less important syllables. So in French only 1.7 syllables/sentence type (or 35% of linguistically pertinent syllables), and in English 1.5 syllables/sentence type (or 30% of linguistically pertinent syllables) are coded by two F0 levels on the linguistically pertinent prosodic boundaries. On the other hand, non-native speakers are less consistent in their prosodic production with more variability in F0 values: 2.8 syllables/sentence type (or 55% of linguistically pertinent syllables) are captured by 2 F0 level values.

Finally, variability is observed also with respect to the number of F0 levels used by the speakers of the 3 groups: the French and English native speakers used up to 6 F0 levels for their F0 patterns while the non-native speakers used only 4 F0 levels. That means that the F0 patterns of non-native speakers are more monotonous than the F0 patterns of the native speakers.

4. Conclusions

The goal of our study was to use an appropriate coding schema for prosody representation in a cross linguistic study of French and English prosody. The data used for testing the proposed method were laboratory data produced by a group of French and English native speakers and they contained sentences sharing the same syntactic structures in both languages. This syntactic specificity of the data base was well adapted to a cross-linguistic study as it allowed for comparison of prosodic phenomena relatively easily.

The methodological problem addressed here was how to represent prosodic parameters in such a way that comparison of the occurrences of these parameters in different sentences and languages would be pertinent. The aim of the study was to represent the general tendency of the F0 pattern by extracting one, or maximum two, F0 values per syllable, coded in terms of 9 F0 levels calculated from the voice range of each speaker. In the present study for each syllable one or maximum two F0 values were kept to capture the prosodic tendency of the sentence. However, in future, a more general automatic decision algorithm should be used to make the decision of the number of representative values of the F0 more data driven.

5. References

[1] Bartkova, K., Sorin, C.: A model of segmental duration for speech synthesis in French. Speech Communication 6(3): 245-260, 1986.

[2] Bartkova, K., Delais-Roussarie, E., Santiago-Vargas, F.: "PROSOTRAN : a tool to annotate prosodically non-standard data", Proceedings of Speech Prosody, Shangaï, China, 22-25 mai, 2012.

[3] Bartels, C.: *The Intonation of English Statements and Questions*, New-York: Garland Publishing, 1999.

[4] Beyssade, C., Marandin, J.-M., Rialland, A.: "Ground / Focus: a perspective from French". In R. Nunez-Cedeno *et al. (*eds), *A Romance perspective on language knowledge and use: selected papers of LSRL 2001*. Amsterdam/Philadelphia: Benjamins. pp. 83-98, 2003.

[5] Boidin, C. : *Modélisation statistique de l'intonation de la parole expressive*, Thesis, Université Rennes 1, 2009.

[6] Campbell, W.N. : "Syllable-based segmental duration". In *Talking machines: theories, models and design*, Bailly & Benoît (eds). Amsterdam: Elsevier, 211-224, 1992.

[7] Crystal, D.: *Prosodic systems and intonation in English*, Cambridge University Press, 1969.

[8] Delattre, P.: "A comparative study of declarative intonation in American English and Spanish", *Hispania* XLV/2, pp. 233-241, 1938.

[9] Delattre, P. : "Les dix intonations de base du français", *The French Review*, 40/1, 1-14, 1966.

[10] Delais-Roussarie, E. : "Vers une nouvelle approche de la structure prosodique", *Langue Française*, 126 : 92-112. Paris: Larousse, 2000.

[11] Di Cristo, A. : A propos des intonations de base du français. Unpublished ms., 2010.

[12] Fónagy, I., Bérard, E. : "Questions totales simples et implicatives en français parisien, Interrogation et Intonation: Studia Phonetica no 8, Ed. by Grundstrom A., Léon P. Paris: Didier. pp. 53-98, 1973.

[13] Fónagy, I. : "L'accent français, accent probabilitaire: dynamique d'un changement prosodique", in *L'accent en français contemporain*, Fónagy & Léon (eds), *Studia Phonetica* 15, 123-233, 1980.

[14] Gussenhoven, C.: On the grammar and semantcis of sentence accents. Dordrecht: Foris, 1984.

[15] Lacheret-Dujour, A., Obin, N., Avanzi, M.: "Design and evaluation of shared prosodic annotation for French spontaneous speech: from expert's lnowledge to non-experts annotations", in Proceedings of the 4th Linguistic Annotation Workshop, Uppsala, Sweden, 2010.

[16] Mesbahi, L., Jouvet, D., Bonneau, A., Fohr, D., Illina, I., Laprie, Y.: "Reliability of non-native speech automatic segmentation for prosodic feedback", Proceedings of SLaTE, 2011.

[17] Pierrehumbert, J.: *The phonology and phonetics of English intonation*, PhD thesis, published 1988 by IULC, 1980.

[18] Post, B.: *Tonal and phrasal structures in French intonation*, The Hague: Holland Academic Graphics, 2000.

[19] Speech Processing, Transmission and Quality Aspects (STQ): "Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression Algorithms, ETSI ES pp. 202 212, 2005.

[20] Vaissière, J.: "Cross-linguistic prosodic transcription: French vs. English. In *Problems and methods of experimental phonetics. In honour of the 70th anniversary of Pr. L.V. Bondarko*, Volskaya, Svetozarova & Skrelin (eds). pp. 147-164, 2002.

14 Thursday 3

Implicit Prosodic Priming and Autistic Traits in Relative Clause Attachment

Sun-Ah Jun[1], *Jason Bishop*[2,3]

[1]Department of Linguistics, University of California, Los Angeles
[2]Department of English, College of Staten Island, City University of New York
[3]Linguistics Program, The Graduate Center, City University of New York
jun@humnet.ucla.edu, jbishop@gc.cuny.edu

Abstract

Using the structural priming paradigm, the present study explores predictions made by the Implicit Prosody Hypothesis by testing whether an implicit prosodic boundary generated from a silently-read sentence influences attachment preference for a novel, subsequently read sentence. Results indicate that such priming does occur, although the patterns are highly dependent on individual differences in listeners' "autistic" traits.

Index Terms: Implicit Prosody, non-restrictive RC, autistic traits, prosodic disambiguation, AQ

1. Introduction

1.1 Implicit Prosody in Relative Clause Attachment

In a sentence such as (1), it is ambiguous whether the relative clause (RC) modifies NP1 *the servant* (high attachment) or NP2 *the actress* (low attachment):

(1) *Someone shot the servant of the actress who was on the balcony*. ([1])

Although the details of attachment preference are language-specific ([2],[3]), it is known that, cross-linguistically, attachment decisions are sensitive to the sentence's prosodic characteristics, including the location of a prosodic boundary ([4], [5], [6], [7], [8]). This fact has been used to support the Implicit Prosody Hypothesis (IPH; [2], [9]), which states that, in silent reading, a default prosodic contour is projected onto the sentence, influencing syntactic ambiguity resolution. Other things being equal, the parser favors the syntactic analysis associated with the most natural (default) prosodic contour for the construction. Fodor and colleagues claimed that speakers interpret a prosodic break before an RC as a marker of a stronger syntactic boundary, which prompts high attachment. This suggests that the human sentence parser would favor low attachment when the RC forms a single prosodic phrase with NP2, but favors high attachment when a prosodic break directly precedes the RC.

1.2 Accessing Implicit Prosody

An important question is how implicitly-generated prosody can be investigated. Fodor and colleagues (e.g., [10],[6],[11],[7]) assumed that implicit prosody is equal to explicit, or overt, prosody such as that associated with an out-of-the-blue reading. However, recent production studies ([12],[13],[14]) suggest that this is not necessarily the case. These studies found that a majority of English speakers produced a large prosodic boundary directly before the RC, despite English's status as a language with a low-attachment bias. Thus it appears that implicit prosody is not easily accessible via the investigation of overt prosody.

If implicit prosody is not reliably studied by way of overt prosody, it may be possible to study it more directly, by influencing implicit prosody itself and observing the effects on processing. One possible way to accomplish this is to manipulate visual cues that impose implicit boundaries in reading materials. For example, recent studies examining individual differences in attachment preference show working memory capacity to be positively (e.g., [15], [16] for on-line data) or negatively (e.g., [17] for off-line data) correlated with high attachment responses when the target sentence is presented on a single line. However, when the sentence is displayed on two lines, i.e., visually chunked so that the RC is separated from the two head nouns (i.e., …NP1 NP2 // RC...), working memory plays a weaker role and subjects prefer high attachment, regardless of their working memory capacity. This suggests that visual cues, such as visual discontinuity, may influence attachment by imposing implicit prosodic juncture in reading.

Another approach to the manipulation of implicit prosody would be to try to "prime" it using overt prosody. In recent work ([18]), we carried out an experiment intended to do just this, using a novel prosodic adaptation of the structural priming paradigm ([20], [21], [22]). In particular, listeners were auditorily presented with ambiguous RC-sentences that contained a prosodic boundary either before or after NP2, and then had to read another, also ambiguous RC target sentence. Subjects then made attachment decisions about the silently read target. If the overt prosody in the prime sentence influenced the implicit prosody generated during the reading of the target sentence—and implicit prosody influences attachment—this would be observable in attachment decisions.

In fact, a certain subset of listeners were influenced by the primes as predicted by the IPH; after hearing sentences with a boundary after NP2, these subjects were more likely to interpret the silently read target as having a high-attaching RC. Interestingly, this group of subjects consisted of those with prominent "autistic" traits along the "communication" dimension (indicating poorer, more autistic-like communication skills), as measured by the Autism Spectrum Quotient ([23]). At present it is unclear why these individuals should be especially sensitive to prosodic boundaries in this way, although high scores on this dimension (indicating poorer, more

autistic-like communication skills) have been shown inversely related to the use of pragmatic information in sentence processing (e.g., [23], [24]), and, similarly, to the accentual patterns that typically encode such information ([25]). (For further discussion of the individual sensitivity to prosodic structure, and its consequences for sentence parsing, see [19]).

The results of the experiment just described indicate a correlation between the use of a prosodic boundary and attachment of an RC in a way that supports the basic prediction of the IPH. A question remains, however: what was the mechanism? That is, although a boundary separating the RC and the two head nouns was (for subjects with more prominent autistic traits) associated with high attachment, it is unclear whether this happened as a result of syntactic structure priming or prosodic structure priming. In the first case, the overt prosody of the primes would have influenced the syntax assigned to the primes themselves, at which point that syntactic structure could be re-used to parse the target sentence. This is syntactic priming in the typical sense ([20]). In the second case, however, the overt prosody would have influenced the targets more directly, by influencing the implicit prosody generated for those targets. However, both of these scenarios are possible in the experiment just described, it is unclear which is the correct one.

This is the matter that we attempt to better understand in the present study. In the experiment presented below, native English speakers took part in a more traditional (i.e., reading only) structural priming task, involving targets containing an RC with ambiguous attachment, such as in (1) above. However, prime sentences were designed not to have RCs with one or the other attachment possibilities, but to have RCs that modified a head noun restrictively or non-restrictively. The reason for this was that this contrast—orthogonal to the attachment ambiguity—is distinguished primarily by the presence versus absence of a prosodic boundary before the RC in speech; it is represented visually by the presence versus absence of a comma in orthography. An example of the contrast is shown in (2a) and (2b):

(2) a. Restrictive RC in the prime sentence:
The newspaper reporter phoned the secretary who was annoyed.

b. Non-restrictive RC in the prime sentence:
The newspaper reporter phoned the secretary, who was annoyed.

If the implicit prosody (of targets) can be influenced by the implicit prosody (of primes), we expect targets like (1) to be parsed differently depending on whether they are read following a sentence with a restrictive versus non-restrictive RC. In particular, because the non-restrictive RC in the primes contains a boundary before the RC (cued by the comma), it should increase the probability that a boundary will be inserted before the RC in the targets. Based on the results of previous studies (e.g., [27], [17]), we predict that this will induce a greater likelihood of high attachment responses. Based on [18, 19], we also predict this effect to depend somewhat on autistic traits. As described above, autistic traits are measured in the neurotypical population using the AQ, a self-report questionnaire; this measure is composed of five subscales measuring *social skills*, *attention to detail*, *attention switching*, *communication skills*, and *imagination*. Rather than the entire score combined, in our own work, we have found scores on the communication subscale to be inversely related to the use of prosodic prominence in speech [26], and directly related to a sensitivity to prosodic boundaries [18, 19].

This was tested in the experiment presented below in Section 2; a discussion and conclusion based on the results follows in Section 3.

2. Experiment

2.1. Method

2.1.1. Stimuli

Sixteen sentences containing RCs of medium length (4-6 syllables) were created to serve as target sentences to be read by participants. These sentences, based on sentences used in previous studies (e.g., [28], [29], [30], [31], [32], [33]), were designed to lack any grammatical or semantic bias towards high or low attachment. An example of such a sentence is shown in (1), with some additional examples listed in Appendix I.

Prime sentences, to be presented and read immediately before the targets, were based on 30 sentences of similar length. Like the targets, primes contained RCs that were 4-6 syllables in length, but they differed from targets in two ways. First, the RCs in primes followed a single head noun, and so they had unambiguous attachment to that head noun. Second, there were two versions of each prime, one which was to be interpreted as having a restrictive RC, and the other which had a nonrestrictive RC. The non-restrictive RC was marked by the standard orthographic convention, i.e., a comma preceding the RC; the restrictive RC versions lacked any such comma. An example of each version is shown in (2); five additional examples are listed in Appendices II and III.

2.1.2. Participants

Participants were 120 native speakers of American English, mostly undergraduate students at the University of California, Los Angeles. None of the participants reported any speech or communication disorders, and all received either course credit or monetary compensation.

2.1.3. Procedures

Participants read the prime and target sentences, and answered attachment questions, at their own pace. A MATLAB script was used to present participants with the sentence materials on a computer screen. On each experimental trial, the script selected one of the 16 target sentences and three prime sentences from one of the two prime conditions; the order of presentation of primes was randomized on each trial. The subject then proceeded through these four sentences, first the three primes, then finally the target, at their own pace, pressing a computer key to remove one sentence and display the next. Following a key press after the target sentence, however, a question appeared, asking the participant the standard RC-

attachment question (e.g., Who was on the balcony? In the case of (1), above), presenting the two possible head nouns as the options "A" and "B". Whether the high attachment response (i.e., NP1) appeared on the left as "A" or on the right as "B" was counterbalanced for each participant.

Filler trials (consisting of 28 filler targets and 30 filler primes) proceeded in the same manner as experimental trials, with the following exception. On filler trials, a question appeared for one of the three prime sentences, selected at random. This was to prevent participants from knowing exactly which sentence in a trial (i.e., every fourth sentence presented to them) would be the one requiring the answering of a question.

Participants carried on through all experimental and filler trials (randomized for each participant), and the assignment of a particular target sentence to a prime condition (restrictive vs. non-restrictive RC) was counterbalanced across subjects. The task took participants approximately 15-20 minutes. Following the reading task, participants completed the AQ ([23]), a 50-item self-report questionnaire, requiring an additional 10 minutes.

2.2. Results

Two rounds of mixed-effects logistic regression modeling took place; the first was aimed at a simple, overall test of the effect of primes, without considering possible individual differences related to autistic traits. The second round of modeling included AQ scores.

Results of the first model, shown in Table 1, indicated that primes did in fact influence how participants interpreted the ambiguous targets; as predicted, participants chose high attachment significantly more often after reading sentences in the non-restrictive RC condition that contained a comma than in the restrictive RC condition that did not contain a comma (see Figure 1). That is, the implicit prosody of primes influenced the implicit prosody of the targets, which then influenced the attachment resolution, thus supporting the IPH. Additionally, the significant effect of trial indicated that, as the experiment went on, high attachment responses became more likely overall.

A second model of mixed effects logistic regression included the participants' scores on each of the subscales of the AQ (Communication, Social Skills, Attention to Detail, Attention Switching, and Imagination), in addition to the factors in the first model. The results of the model (see Table 2) showed significant interaction between Prime Type and AQ-Communication scores. As shown in Figure 2 (next page), the influence of primes was stronger for individuals with higher scores on this subscale (indicating more autistic-like communication skills).

There was also a significant main effect for scores on the AQ-Attention Switching subscale; as shown in Figure 3, higher scores in AQ-Attention scale were associated with a greater likelihood of a "high attachment" response. Finally, the significant effect of trial found in the simpler model also held here, indicating high attachment was more likely on later trials.

Table 1. *Estimates, standard errors, z- and p-values for the first model testing the effect of primes on high attachment responses. Positive estimates indicate the amount of increase in log-odds relative to the intercept.*

Fixed effects:	*β*	*SE*	*z*	*p*
(Intercept)	-.389	.213	-1.83	.068
Trial	.009	.003	2.55	**.011**
PrimeType(*nonrestrictive*)	.220	.104	2.11	**.035**

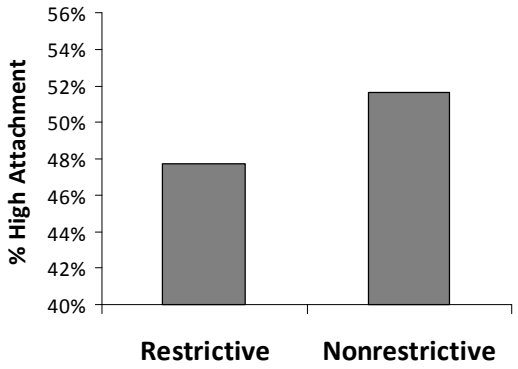


Figure 1: High attachment responses to target sentences in the two prime conditions, based on RC-type.

Table 2. *Estimates, standard errors, z- and p-values for the second model of high attachment responses, adding AQ subscales.*

Fixed effects:	*β*	*SE*	*z*	*p*
(Intercept)	-1.797	.933	-1.93	.054
Trial	.009	.004	2.45	**.014**
AQ-Attention Switching	.096	.036	2.70	**.007**
AQ-Communication	-.052	.036	-1.45	.147
PrimeType(*nonrestr.*)	-.932	.511	-1.82	.068
PrimeType(*nonrestr.*) × AQ-Communication	.060	.026	2.30	**.021**

3. Discussion and Conclusion

The structural priming experiment presented above was intended to test a basic prediction of the Implicit Prosody Hypothesis ([2], [9]), which says that a strong prosodic boundary generated implicitly during reading influences the attachment of a relative clause. In particular, a strong prosodic boundary directly before the RC is predicted to encourage a high attachment parsing of the RC. Our prime sentences were designed to manipulate the presence versus absence of such a boundary. In fact, results showed that, after reading primes with a boundary, participants were more likely to attach the RC high

in a novel, structurally ambiguous sentence. First, and significantly, this suggests that implicit prosodic structure, like syntactic structure, can be primed; second, in relation to our main goal, it demonstrates that silently-generated prosodic boundaries influence attachment—supporting the IPH.

The results also confirm previous findings that autistic traits in the neurotypical population are relevant to predicting sensitivity to prosody in sentence processing ([26]). In particular, the communication subscale of the AQ seems to be correlated with sensitivity to prosodic boundaries ([18, 19)]. Participants with higher AQ-communication scores (e.g., those who are not good at social chit-chat, or are slower in understanding the point of a joke), chose more high attachment responses for targets following non-restrictive RC primes than those following restrictive RC primes. Though the mechanisms underlying this tendency require further study, we hypothesize that individuals with high AQ-communication scores, rather than incorporating the prosodic boundary to attachment, might have been disrupted by the juncture, prompting closure at the location of boundary.

On the other hand, a second finding, not previously reported, is also related to autistic traits. Namely, high scores on the attention-switching subscale of the AQ (i.e., worse attention-switching abilities) were associated with higher overall rates of high attachment. The relation between attachment preference and attention-switching therefore resembles the one between attachment preference and verbal working memory capacity ([16], [17]), possibly because both of these reflect similar general processing resources. That is, those who lack attention shifting abilities and those who have lower working memory capacity might both have difficulty integrating multiple sources of information. This may be supported by the main effect of trial, which indicated that, as the experiment went on (and participants possibly became more fatigued), high attachment decisions became more likely. Again, further research is needed to understand the processing mechanism underlying the performance of individuals with more prominent autistic traits.

In sum, the present study provides crucial evidence for the Implicit Prosody Hypothesis with respect to the relation between implicit boundaries in reading and attachment preferences. Further, we have shown that implicit prosodic structure, like syntactic structure, can be primed, and that the details of this priming depend on autistic traits in the neurotypical population.

4. Acknowledgements

The authors thank Henry Tehrani for his help in writing the MATLAB script and undergraduate research assistants Katie Brown, Sewon Na, and Hannah Kim for help running the experiment. Work on this project was facilitated by a UCLA Faculty Research Grant to Sun-Ah Jun.

5. Appendices: Example Stimuli (part)

I **Example Target Sentences**

1 Jennifer blackmailed the boss of the clerk that was dishonest.
2 Susanna was dating the cousin of the artist that was a veteran.
3 The lady mended the sleeve of the shirt that had been stained.

II **Example Prime Sentences (Restrictive, without comma)**

1a The inspector photographed the boat's cover that was yellow.
2a The coach looked at the varsity players who were very happy.
3a The picky journalist hated the soldiers who were sitting down.

III **Example Prime Sentences (Non-restrictive, with comma)**

1b The inspector photographed the boat's cover, which was yellow.
2b The coach looked at the varsity players, who were very happy.
3b The picky journalist hated the soldiers, who were sitting down.

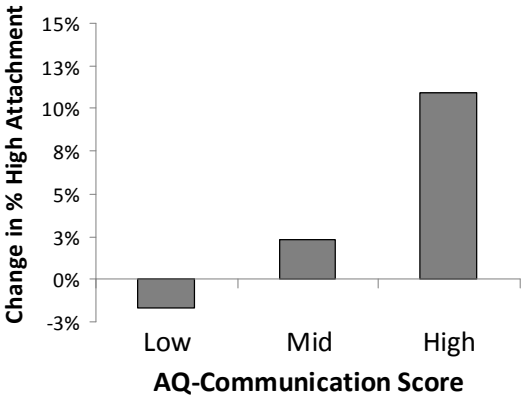


Figure 2: *Increase in high attachment responses to targets in the non-restrictive RC prime condition (relative to the restrictive RC prime condition) as a function of AQ-Communication scores. The three levels of AQ refer to the group distribution: "Mid" are subjects scoring within 1 SD of the mean, the "Low" and "High" levels below or above 1 SD.*

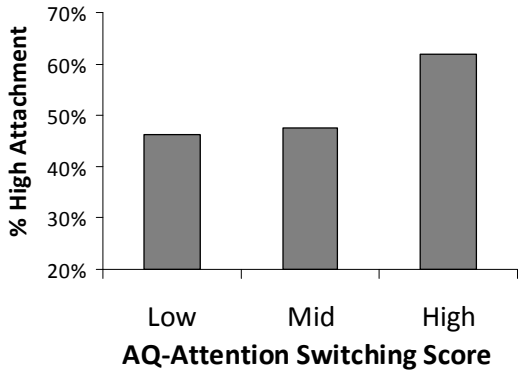


Figure 3: *High attachment responses as a function of scores on the Attention-Switching subscale of the AQ. The three levels refer to the group distribution: "Mid" are subjects scoring within 1 SD of the mean, the "Low" and "High" levels below or above 1 SD.*

6. References

[1] Cuetos, F., & Mitchell, D. C. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition*, *30*, 73-105.

[2] Fodor, J. D. 1998. Learning to parse. *Journal of Psycholinguistic Research, 27(2),* 285-319.

[3] Fernández, E. M. 2003. *Bilingual sentence processing: Relative clause attachment in English and Spanish*. Amsterdam: John Benjamins.

[4] Schafer, A. J. 1997. *Prosodic parsing: The role of prosody in sentence comprehension.* Ph.D. Dissertation, University of Massachusetts.

[5] Kjelgaard, M. M., & Speer, S. R. 1999. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *J. of Memory and Lang.*, *40*, 153-194.

[6] Quinn, D., Abdelghany, H., & Fodor, J. D. 2000. More evidence of implicit prosody in reading: French and Arabic relative clauses. Poster presented at the 13th Annual CUNY Conference on Human Sentence Processing, La Jolla, CA, March 30-April 1.

[7] Lovrić, N., Bradley, D., & Fodor, J. D. 2001. Silent prosody resolves syntactic ambiguities: Evidence from Croatian. Presented at the SUNY/CUNY/NYU Conference, Stonybrook, NY.

[8] Jun, S.-A. 2003. Prosodic Phrasing and Attachment Preferences. *Journal of Psycholinguistic Research*, *32(2)*, pp. 219-249.

[9] Fodor, J. D. 2002. Prosodic Disambiguation in Silent Reading. *NELS* 32, 113-32.

[10] Maynell, L. A. 1999. Effect of pitch accent placement on resolving relative clause ambiguity in English. Poster presented at the 12th Annual CUNY Conference on Human Sentence Processing, New York.

[11] Lovrić, N., Bradley, D., & Fodor, J. D. 2000. RC attachment in Croatian with and without preposition. Poster presented at the AMLaP Conference, Leiden.

[12] Bergmann, A., Armstrong, M., & Maday, K. 2008. Relative clause attachment in English and Spanish: A production study. *Proceedings of Speech Prosody* 2008, Campinas, Brazil.

[13] Bergmann, A., & Ito, K. 2007. Attachment of ambiguous RCs: A production study. Talk given at the 13th annual conference on architectures and mechanisms for language processing (AMLaP), Turku, Finland, 24-27 August 2007.

[14] Jun, S.-A. 2010. The Implicit Prosody Hypothesis and Overt Prosody in English. *Language and Cognitive Processes*, *25*, 1201-1233.

[15] Traxler, M. J. 2007. Working memory contributions to relative clause attachment processing: A hierarchical linear modeling analysis. *Memory and Cognition*, *35*, 1107-1121.

[16] Traxler, M. J. 2009. A hierarchical linear modeling analysis of working memory and implicit prosody in the resolution of adjunct attachment ambiguity. *J. of Psycholinguistic Research*, *38*, 491-509.

[17] Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. 2007. The role of working memory in syntactic ambiguity resolution: A psychometric approach. *J. of Experimental Psychology: General*, *136*, 64-81.

[18] Jun, S.-A. & Bishop, J. 2013. Implicit prosody priming in relative clause attachment. Paper presented at the 87th Meeting of the Linguistic Society of America, Boston, 3-6 January 2013.

[19] Jun, S.-A. & Bishop, J. in progress. Prosodic Priming and Individual Differences. ms. UCLA.

[20] Bock, J. K. 1986. Syntactic persistence in language production. *Cognitive Psychology, 18*, 355-387.

[21] Scheepers, C. 2003. Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. *Cognition, 89,* 179-205.

[22] Loncke, M., Van Laere, S. M.J., & Desmet, T. 2011. Cross-structural priming. *Experimental Psychology, 58(3)*, 227-234.

[23] Baron-Cohn, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. 2001. The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians. *J. of Autism and Developmental Disorders, 31(1)*, 5-17.

[24] Nieuwland, M., Ditman, T., & Kuperberg, G. 2010. On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *J. of Memory and Lang.*, *63*, 324-346.

[25] Xiang, M., Grove, J., & Giannakidou, A. 2013. Dependency dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Front. Psychol.* 4:708.doi: 10.3389/fpsyg.2013.00708.

[26] Bishop, J. 2013. *Prenuclear accentuation: Phonetics, phonology, and information structure*. Ph.D dissertation, UCLA.

[27] Carlson, K., Clifton, C. Jr., & Frazier, L. 2001. Prosodic boundaries in adjunct attachment. *J. of Memory and Lang.*, *45,* 58-81.

[28] Dussias, P. E. 2003. Syntactic ambiguity resolution in second language learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition*, 25, 529-557.

[29] Felser, C., Marinis, T. & Clahsen, H. 2003. Children's processing of ambiguous sentences: a study of relative clause attachment. *Language Acquisition* 11, 127-163.

[30] Fernández, E. M., & Bradley, D. 1999. Length effects in the attachment of relative clauses in English. Poster presented at the 12th annual CUNY conference on human sentence processing, New York.

[31] Carreiras, M., & Clifton, C. Jr. 1993. Relative clause interpretation preferences in Spanish and English. *Language and Speech*, *36*, 353-372.

[32] Frazier, L. 1990. Parsing modifiers: Special purpose routines in the human sentence processing mechanism? In D. A. Balota, G. G. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 303-330). Hillsdale, NJ: Lawrence Erlbaum.

[33] Frazier, L., & Clifton, C. Jr. 1996. *Construal*. Cambridge, MA: MIT Press.

Listening for sound, listening for meaning: Task effects on prosodic transcription

Jennifer Cole [1], *Timothy Mahrt* [1], *José I. Hualde* [1,2]

[1] Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[2] Department of Spanish, Italian and Portuguese , University of Illinois at Urbana-Champaign, Urbana, IL, USA

jscole@illinois.edu, tmahrt2@illinois.edu, jihualde@illinois.edu

Abstract

The perception of prosodic structure (phrasal prominences and boundaries) may depend in part on acoustic cues in the speech signal and in part on utterance meaning as related to syntactic structure and discourse context. In this study we ask if listeners are able to differentially weigh acoustic and meaning-based cues to prosody. We test naïve subjects' transcription of prominences and boundaries in spontaneous American English under three different conditions, all of which involve listening to audio recordings and marking prominences and boundaries on a transcript. The three conditions differ in the instructions given to transcribers. In one condition, subjects were instructed to transcribe prominence and boundaries based on meaning criteria, in a second condition they were told to transcribe based on criteria of acoustic salience, and a third condition had less specific instructions, without explicit reference to either meaning-based or acoustic cues. Our results show that subjects perform differently when focusing on meaning than when focusing on acoustics, especially for prominence marking, where partially different sets of words are selected as prominent under the two tasks. Boundary marking is more similar under the two instructions, with acoustic criteria resulting in more listeners marking a given word as pre-boundary, but with boundaries marked largely on the same words in both tasks. With non-specific instructions, performance was similar to that obtained under acoustic-based instructions. We report on agreement rates within and across conditions. This study has implications for models of prosody perception and the methodology of prosodic transcription.

Index Terms: prosody, prominence, boundaries, prosodic transcription

1. Introduction

Prosodic prominences and boundaries are assigned to utterances based on many factors related to syntactic structure and discourse context. Because of these dependencies, the acoustic cues that signal prosody also serve as cues to the linguistic context of the prosodically marked word and the utterance to which it belongs. For instance, listeners interpret syntactic structure based in part on acoustic cues that signal prosodic boundaries [1,2,3,4], and the interpretation of the focus and information status of a word is influenced by acoustic cues to prominence [5,6,7,8,9]. The influence of prosodic cues on discourse processing is such that a mismatch between the discourse context and a word's prosodic form can disrupt processing, as shown by evidence from eye-tracking [6,8,10] and ERP studies [11].

The studies cited above, and many others, demonstrate the role of prosody in communicating meaning related to the syntactic, semantic and discourse context of an utterance. While the evidence shows that listeners attend to the prosodic cues present in the acoustic speech signal, it's possible that listeners' perception of prosody is also driven by expectations about the prosodic form of an utterance given its syntactic properties and its semantic and discourse context, in the same way that expectations play a crucial role in word recognition (and in the visual domain). In this paper we examine prosody perception due to acoustic cues and due to expectations from factors related to syntactic, semantic and discourse context (hereafter *meaning-based cues*), and ask whether listeners can focus differentially on acoustic and meaning-based cues in identifying prosodic prominence and phrase boundaries in spontaneous speech. We examine listeners' perception of prosody using the method of Rapid Prosody Transcription (RPT) developed by one of the authors (JC) for investigating prosody through the analysis of judgments made by naïve native speakers of English [12]. This method and prior findings are introduced in the next section.

1.1. Rapid Prosody Transcription

Under the RPT methodology multiple listeners (between 10-20 in prior experiments) make auditory judgments about the location of prosodic phrase boundaries and prominences in an audio speech recording, based only on (the individual listener's) auditory impression and with no visual inspection of the graphical speech display. For each word of transcribed speech two continuous-valued prosody features are calculated, representing the proportion of transcribers who perceived the word as prominent (the p-score), and the proportion who perceived the word as final in a prosodic phrase (the b-score). A p-score of 0 shows agreement among all transcribers that the word is not prominent while a p-score of 1 shows agreement that the word is prominent. Values in between 0 and 1 reflect disagreement among transcribers. The prosody scores can be viewed as a measure of the probability that a random listener (from the same speech community) will perceive a given word as prominent, or as preceding a prosodic phrase boundary. Fig. 4 below shows an example of the p- and b-scores for each word in a fragment of a speech sample from this study, based on the aggregated transcriptions of 16 listeners.

Cole and her colleagues conducted two studies of prosody perception with American English spontaneous speech using RPT, testing the relative contribution of signal-based processing (from acoustic cues) and expectation-based processing (from syntactic and information-based cues) in non-expert listeners' judgments of prosodic prominence and phrase boundaries. Cole et al. [13] investigated p-scores and their relationship to various acoustic cues previously found to correlate with prosodic prominence such as increased duration, increased intensity, and the presence of a pitch accent. They found a positive correlation between these acoustic measures

and p-scores: higher values of the acoustic measures (e.g., longer duration, higher intensity) predict higher p-scores (indicating higher agreement among listeners that a word is prominent) for a given word. However, p-scores were also correlated with measures of word surprisal—non-acoustic cues such as word frequency and number previous mentions of a word in the discourse. Similarly, Cole et al. [14] found that syntactic context predicts the perception of boundaries in spontaneous speech, in addition to and partly independent of acoustic cues.

These two prior studies using RPT provide evidence that in perceiving prosodic prominences and phrase boundaries, listeners are influenced both by acoustic cues and by cues related to the syntactic role of a word and its meaning in relation to discourse context, and that these cues function at least partly independently of one another. However, these studies do not fully indicate to what extent acoustic and meaning-based cues are different and whether listeners can attune their attention to either acoustic or meaning-based cues, diminishing the other. These issues are the point of departure for the present study.

2. Methodology

2.1. Subjects and Materials

This experiment uses the Rapid Prosody Transcription method [12] to obtain prominence and boundary judgments from 15 naïve native speakers of English, all students at the University of Illinois. Sixteen short excerpts (~18 s each) from sixteen different speakers in the Buckeye corpus of spontaneous English speech [15] were used in this study. The total number of words summed over all excerpts was 925. This dataset is a subset of the dataset used in the study by [12] and comparisons with the findings of that study are included in Section 3. The transcription experiment was conducted in a quiet, computer-equipped room. Subjects proceeded through the experiment at their own pace using LMEDS, a customized software application developed by the authors.

2.2. LMEDS

A customized web interface, LMEDS, the Language Markup and Experimental Design Software, was developed to administer the experiment materials electronically. LMEDS is a generic toolset that simplifies the creation of custom experimental setups, such as those needed for an RPT experiment, as well as the aggregation of data collected during the experiment.

In this LMEDS experiment each excerpted speech sample was presented on its own page. Each of these pages presents a button for playing the audio file, a transcript where each word is clickable, and a button to progress to the next phase (Fig. 1). Each participant (hereafter, transcriber) first listened to the audio twice while clicking on individual words to mark a perceived boundary after the word. The transcribers had no training in phonetics and were not shown any visual display of the speech waveform, spectrogram or pitch track. After two passes through the file, listening and marking boundaries, transcribers then listened to the audio passage two more times, clicking on words perceived as prominent. The interface displayed the location of a selected boundary with a thick vertical line and indicated a word marked as prominent by changing the font color of the word to red. While marking prominences, transcribers were able to see, but not modify, the boundaries they had just placed. After annotating a transcript for both boundaries and prominences, subjects would progress to the next page in the experiment.

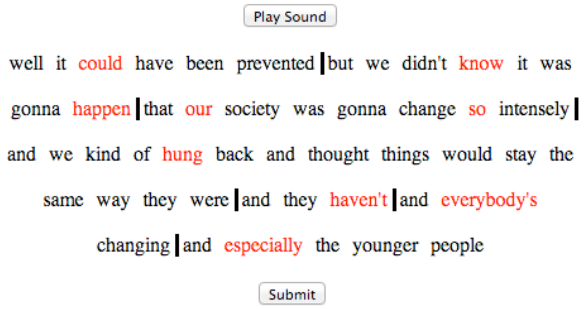


Figure 1. Example transcript and action buttons for a speech excerpt displayed in LMEDS, with prominences and boundaries as marked by an individual transcriber.

2.3. Task instructions

The experiment was divided into two blocks. Each block specified the criteria transcribers were to use in making judgments about the location of prominences and boundaries. In the *acoustics* block, subjects were asked to mark a boundary where they heard a 'break, discontinuity or disconnection in the speech stream, strong or subtle' and were asked to mark a prominence where they heard a word stand out by 'being louder, longer, more extreme in pitch, or more crisply articulated.' In the *meaning* block, subjects were asked to mark boundaries where the audio could be 'segmented with minimal disruption of the meaning of the speech' and were asked to mark the words that 'convey the main points of information as you think the speaker intended.'

3. Results

In our analysis we consider the RPT task under three different instruction sets: the explicit acoustic-based and meaning-based instruction sets that were run for the present study and the less explicit instructions used in [12,13,14].

Our first question is whether or not subjects performed differently across these three tasks. One way to assess differences due to task instructions is through inter-transcriber agreement, as reflected in p-scores and b-scores. Greater agreement would indicate stronger and more consistent cues to prosody under the stated criteria (acoustic or meaning-based). Fig. 2 shows that the distribution of b-scores and p-scores across all of the recordings are largely similar across the three transcription tasks. Most words receive a b-score of zero, indicating that no transcribers marked a boundary following the word, with a nearly flat distribution of b-scores greater than zero. The distribution of p-scores is similar, with the majority of words receiving a p-score of zero (again, indicating that no transcribers marked the word as prominent). However, there are also a significant number of words with low, non-zero p-scores (i.e, words that very few transcribers marked as prominent). Although there is some variation across tasks, the overall trend in transcriber agreement for p-scores or b-scores is the same.

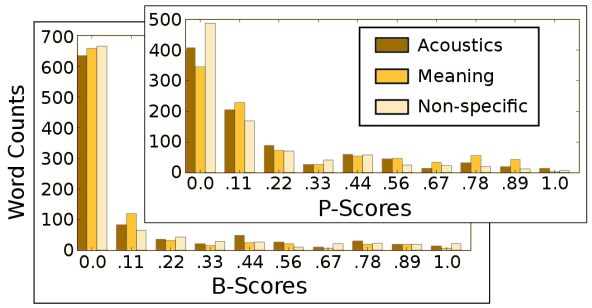


Figure 2. Distribution of b- and p-scores across the three tasks differing in transcription instruction.

The distributions of b-scores and p-scores also show that across tasks transcribers are marking a similar number of words as prominent or as preceding a boundary. For instance, looking at the histogram of b-scores we see that the number of words with a b-score of zero (indicating that no transcriber marked the word for a boundary) is very similar across the three tasks. We also observe that boundary labeling is more conservative than prominence labeling: in all tasks there are more words with a b-score of zero than there are words with a p-score of zero, which means that there are relatively fewer words being marked by transcribers as pre-boundary compared to the number of words marked as prominent. One exception to the overall similarity in prosody scores across tasks, as shown in Fig. 2, is the lower number of words with a p-score of zero under the meaning-based and acoustic criteria compared to the non-specific criteria. This finding indicates that transcribers are more likely to judge a word as prominent when attending to specific cues than when judging prominence in a non-specific way .

The distributions of p-score and b-score values for all tasks show a high agreement among transcribers on words that are not prominent (p-score=0), and on words that are not preceding a boundary (b-score=0), but it reveals little about the patterns of agreement on individual words as marked under different transcription instructions. We are interested to know if the transcribers weigh acoustic and meaning-based criteria differently on the basis of the task instructions, marking different words as prominent/important or as pre-boundary, across tasks.

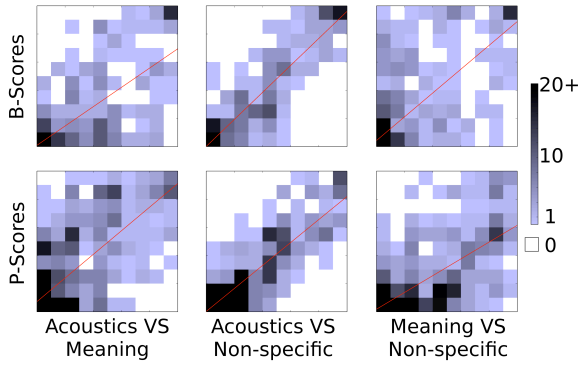


Figure 3. Linear correlations on top of density plots comparing b-scores (top) or p-scores (bottom) in one task against the corresponding score for the same word in a different task. The plots represent data density (# of datapoints in the same region of the plot) with values on the color scale as shown in the legend.

We examine task effects on transcription by comparing b-scores and p-scores for each word from one task with the scores for the same word from a different task, using correlation and linear regression analysis (Fig. 3). All correlations are significant, and r^2 values are above 0.5 for all comparisons (Table 1), indicating that the selection of words that are prominent and in pre-boundary position is similar across tasks, though not identical. The transcriptions that are most correlated (with the highest r^2) are those based on acoustic and non-specific criteria. The meaning-based transcription is less correlated to transcription under either acoustic or non-specific criteria. These findings show that transcribers are able to weigh acoustic and meaning-based cues differently when specifically instructed, but that in the absence of instructions calling for attention to meaning-based criteria, transcribers rely more on acoustic cues in marking prominences and boundaries. Thus, providing different instructions to subjects can indeed cause them to attune their attention to different types of information in speech.

	Acoustics VS Meaning	Acoustics VS Non-specific	Meaning VS Non-specific
B-scores	0.616	0.892	0.528
P-Scores	0.575	0.828	0.540

Table 1. Linear regression coefficients (r^2) for data plotted in Figure 3. ($p < 0.01$ for all reported values)

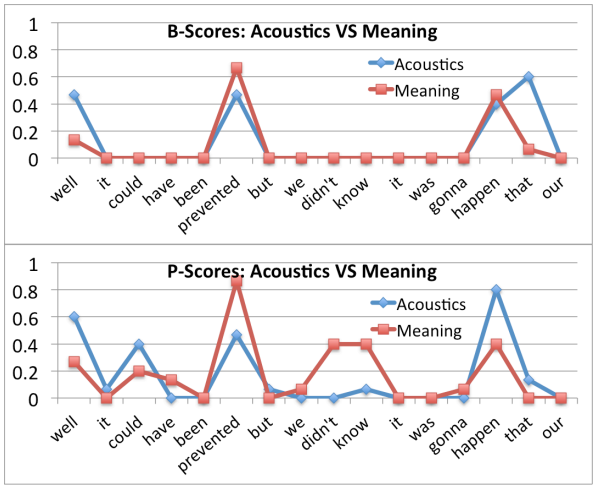


Figure 4. B-scores and p-scores for words from a fragment of a single excerpt, comparing the acoustics-based task with the meaning-based task.

The relatively lower correlation between acoustic and meaning-based prosody scores tells us that there are frequent mismatches in transcribers' marking of words in these two tasks. There are two scenarios that can explain such differences. One possibility is that transcribers in one task are marking a subset of the words that are marked in the other task. Taking p-scores as an example, it could be words that are selected as prominent under meaning-based criteria are also selected under acoustic criteria, but not vice versa. Under this scenario, (nearly) all words with p-scores greater than zero in the meaning-based transcription would also have non-zero p-scores in the acoustic transcription, but in addition, some words with non-zero p-scores in the acoustic transcription

would have a p-score of zero in the meaning-based task. The second scenario is where *different* words are marked as prominent under the two transcription criteria. This pattern of mismatch would result in words that have p-scores of zero in one task and non-zero p-scores in the other task, and vice-versa.

Fig. 4 plots b-scores and p-scores for individual words for a fragment of one speech excerpt, comparing acoustic-based with meaning-based transcription. Recall that the regression analysis (Table 1) shows substantial disparity between tasks in both b-scores and p-scores. The example in Fig. 4 suggests that task-related differences pattern differently for p-scores compared to b-scores. The b-scores under the two tasks (top panel) are similar in the selection of words that are marked as pre-boundary by any transcriber—the lines graphing b-scores nearly lie on top of one another—with b-scores in the two tasks differing mostly in the *number* of transcribers who mark a word as pre-boundary. On the other hand, the p-scores under the two tasks (bottom panel) differ substantially in which words are marked as prominent—the lines graphing the p-scores do not appear so nearly as lying on top of one another.

To further investigate the nature of the disparity between acoustic and meaning-based transcription, we compare prosody scores between the tasks after binning all scores into two values: 0 and 1. All b-score and p-score values of zero remain at zero, and values greater than zero are set to 1. This amounts to labeling a word as prominent if one or more transcriber marks it so, and otherwise labeling it as not-prominent. B-scores are similarly transformed to the labels boundary and not-boundary. Using these binned results, we can now ask how often two tasks align in their scores, which reveals the extent to which the same words are selected as prominent or as pre-boundary in both tasks, disregarding differences in the number of transcribers (>0) who agree in marking the word. In this analysis we ask whether a word that is labeled as prominent by acoustic criteria is also labeled as prominent under meaning-based criteria, and similarly for boundary labeling.

Fig. 5 shows the patterns of agreement between acoustic and meaning-based transcription for prominence and boundary labels (P/B) from the binned p-score and b-score data. Words counted in the Agree groups have the same P (or B) label in both tasks (acoustic, meaning), where 1 marks P (or B) and 0 marks not-P (or not-B). Words counted as Disagree are labeled differently in the two tasks. We observe several interesting findings. First, the vast majority of the words in the speech samples (89%) have the same boundary label across tasks (words in the 'Agree' groups), with most words assigned the not-boundary label. This meets our expectation, given that most prosodic phrases contain more than one word. However, we also note that there is high agreement across tasks for words with the boundary label, and only 11% of words are assigned different boundary labels in the two tasks. Turning to the prominence labels, we again note that overall more words are marked as prominent than are marked as (pre-)boundary, but there is also a lower overall level of agreement between tasks in prominence labeling, with only 76% of words assigned the same prominence label (the Agree groups). In other words, 24% of the words in this sample are marked as prominent in one task but not in the other. Words with disagreeing labels include those marked as prominent under acoustic criteria but not under meaning-based criteria, and vice-versa. These findings confirm the patterns shown in Fig. 4, that differences in p-scores between the tasks reflect the selection of different words marked as prominent, where differences in b-scores tended to reflect differences in the number of transcribers marking a word as pre-boundary more than differences in which words are marked.

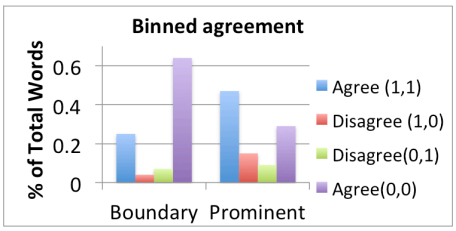


Figure 5. Number of words (% of total) that agree or disagree in prominence and boundary labels across acoustic and meaning-based tasks.

In both the histograms (Fig. 2) and the agreement ratings (Fig. 5) we see that there are more words have a b-score of 0 than there are words having a p-score of zero. For the p-scores, that mass is mostly redistributed to the low, non-zero p-scores (between 0.1-0.3). One reason for this discrepancy between prominence and boundary marking might be the number and variety of factors that condition the placement of prosodic prominence vs. boundaries. Thus, a boundary may be placed at a major syntactic juncture and also preceding disfluency. Prominence, on the other hand, seems to be conditioned by a greater variety of factors. Importantly, words that are acoustically salient do not necessarily convey new or pragmatically important information [16]. There may be differences among transcribers and/or across tasks in the weighting of these factors in the perception of prominence, resulting in greater variability in prominence marking. Another consideration is that the acoustic correlates of prominence seem to be more variable across speakers and utterances compared to the acoustic correlates of boundaries [17], and it's possible that transcribers vary in their sensitivity to individual cues.

4. Conclusions

This study compared prosody transcription under task conditions that focus listeners' attention on acoustic vs. meaning-based criteria. Transcription of prominence and boundaries in spontaneous American English was conducted by non-expert listeners. The findings show similar frequency of boundary and prominence marking across tasks, a lower frequency of boundaries than prominences, and higher agreement among transcribers in the location of boundaries. Task-related differences were also observed: more frequent prominence marking under meaning-based criteria, and a greater disparity between tasks in the individual words that are marked as prominent than there is for words marked as preceding a boundary. Overall there is more uniformity across transcribers and across tasks in boundary marking, parallel to results on inter-transcriber reliability for the ToBI prosodic transcription system [18,19]. This finding calls for future work on the status of prominence in speech production and perception, and on the criteria for prominence transcription. Our ongoing work investigates acoustic cues and also compares p-scores and b-scores gathered in a text-only condition with those obtained under the conditions described in this paper.

.

5. Acknowledgements

This study is supported by NSF BCS 12-51343.

6. References

[1] Schafer, A., Speer, S., Warren, P., & White, S. D. Intonational disambiguation in sentence production and comprehension. Journal of Psycholinguistic Research, 29(2):169-182, 2000.

[2] Carlson, K., Clifton, C., Jr., & Frazier, L. Prosodic boundaries in adjunct attachment. Journal of Memory and Language, 45:58–81, 2001.

[3] Clifton, C., Carlson, K., & Frazier, L. Informative prosodic boundaries. Language and Speech, 45:87–114, 2002.

[4] Snedeker, J. and Truesell, J. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. Journal of Memory and Language, 48:103-130, 2003.

[5] Birch, S., & Clifton, C. Jr. Focus, accent, and argument structure: Effects on language comprehension. Language and Speech, 38(4):365-392, 1995.

[6] Dahan, D., Tanenhaus, M. K., & Chambers, C. G. Accent and reference resolution in spoken-language comprehension. Journal of Memory and Language, 47(2):292-314, 2002.

[7] Chen, A., den Os, E. & de Ruiter, J.P. Pitch accent type matters for online processing of information status: Evidence from natural and synthetic speech. The Linguistic Review, 24, 317–344, 2007.

[8] Ito, K., & Speer, S. R. Anticipatory effects of intonation: Eye movements during instructed visual search. Journal of Memory and Language, 85(2):541-573, 2008.

[9] Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. Interpreting pitch accents in on-line comprehension: H* vs. L_H*. Cognitive Science, 32, 1232-1244, 2008.

[10] Arnold, J. E. (2008). *THE BACON* not *the bacon*: How children and adults understand accented and unaccented noun phrases. Cognition, 108(1), 69-99.

[11] Magne, C., Astésano, C., Lacheret-Dujour, A., Morel, M., Alter, K., & Besson, M. On-line processing of "pop-out" words in spoken French dialogues. Journal of cognitive neuroscience, 17(5):740- 756, 2005.

[12] Mo, Y., Cole, J., Lee, E. Naïve listeners' prominence and boundary perception. In Proceedings of the 4th Speech Prosody, Campinas, Brazil, 2008.

[13] Cole, J., Mo, Y., & Hasegawa-Johnson, M. Signal-based and expectation-based factors in the perception of prosodic prominence. Laboratory Phonology, 1:425–452, 2010.

[14] Cole, J., Mo, Y., & Baek, S. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. Language and Cognitive Processes, 25(7):1141–1177, 2010.

[15] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. Buckeye corpus of conversational speech (2nd release). Columbus, OH: Department of Psychology, Ohio State University, 2007. Retrieved March 15, 2006, from www.buckeyecorpus.osu.edu.

[16] Calhoun, S. The centrality of metrical structure in signaling information structure: A probabilistic perspective. Language, 86:1-42, 2010.

[17] Mo, Y. Prosody production and perception with conversational speech. PhD Thesis, University of Illinois, 2011.

[18] Pitrelli, J.F., Beckman, M.E., & Hirschberg, J. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, 123-126, 1994.

[19] Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In Proceedings of the International Conference on Speech and Language Processing, Jeju, Korea, 2729-2732, 2004.

Acoustic-Prosodic Characteristics of Sleepy Speech – Between Performance and Interpretation*

Florian Hönig[1], *Anton Batliner*[1,2], *Elmar Nöth*[1,3], *Sebastian Schnieder*[4], *Jarek Krajewski*[4]

[1]Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
[2]Institute for Human-Machine Communication, Technische Univ. München, Munich, Germany
[3]Electrical & Computer Engineering Dept., King Abdulaziz University, Jeddah, Saudi Arabia
[4]Experimental Industrial Psychology, University of Wuppertal, Germany
{hoenig,batliner}@cs.fau.de

Abstract

When we address speaker states like sleepiness, two partly competing interests can be observed: both within applications and engineering approaches, we aim at utmost performance in terms of classification or regression accuracy – which normally means using a very large feature vector and a brute force approach. The other interest is interpretation: we want to know what tells apart atypical (here: sleepy) speech from typical (here: non-sleepy) speech, i.e., their respective feature characteristics. Both interests cannot be served at the same time. In this paper, we preselect a small number of easily interpretable acoustic-prosodic features modelling spectrum and prosody, based on the literature and on the general idea of sleepiness being characterised by relaxation. Performance obtained with these single features and this small feature vector is compared with the performance obtained with a very large feature vector; moreover, we discuss to which extent the features chosen model relaxation as sleepiness characteristic.

Index Terms: paralinguistics, sleepiness, prosody, brute forcing, interpretation

1. Introduction

Sleepiness is definitely an interesting research topic, both for practical reasons – the detection of sleepiness is highly relevant in scenarios where sleepiness can cause accidents (driving, flying, operating of machines), and for general reasons – it is ubiquitous, we face it several times a day. As a multi-modal phenomenon, it can be perceived/measured within all modalities, be this speech, facial gestures, eye movements, gait, body posture, or biosignals. Each of these modalities has its pros and cons, as far as processing is concerned: for video processing, light conditions should be favourable; biosensors are intrusive; audio recordings are non-intrusive and possible even under less favourable noise conditions. In this paper, we will concentrate on audio. Moreover, we concentrate on one specific research problem which might not be formulated that often explicitly but sort of gets into our way very often: do we want to get better, or do we want to get any wiser? For getting better, i.e., for obtaining the highest accuracy in classification or regression, we normally employ a very large feature vector (with or without subsequent feature selection): the baseline result for the Interspeech 2011 sleepiness challenge [1] was obtained using the openSMILE tool and 4368 features. Finding features that are most relevant for performance and at the same time easily interpretable is not an easy task [2]. For getting any wiser, so far, we are confined to the rather 'traditional' way of doing research: we employ a small set of promising features and, if possible, formulate a working (alternate) hypothesis on what we expect to find. These promising features are at the same time easily interpretable such as *F0 mean*, in contrast to complex and at the same time opaque features such as the *75% quantile of the 10th MFCC coefficient on consonantal frames* which will turn out as the 2nd most important feature obtained in our data-driven feature selection, cf. Section 4. However, we will definitely not get the highest possible classification/regression performance when using the 'traditional' approach. Here, we try to combine these two different approaches. After presenting the database in Section 2, we sketch in Section 3 the feature sets employed. The experiments reported in Section 4 are discussed in Section 5.

* The authors have received funding from the German Research Council (DFG) under grant agreements KR 3698/4-1 and NO 444/6-1. The responsibility lies with the authors.

2. Data and Annotation

We employ the Sleepy Language Corpus (SLC) from the Interspeech 2011 Speaker State Challenge [1, 3]. Ninety-nine German speakers took part in six partial sleep deprivation studies (mean age 24.9 years, standard deviation 4.2 and a range of 20–52 years; recordings in a realistic car environment or in lecture-rooms; microphone-to-mouth distance 0.3 m, sampling rate 16 kHz, quantisation 16 bit). We disregard the isolated vowels and use the remaining five subsets (7745 speech files ("turns", units of analysis in our regression approach), about 20 hours of speech): read speech: the story of "Die Sonne und der Nordwind" ('the North Wind and the Sun'); commands/requests: simulated driver assistance system commands/requests, e. g. "Ich suche die Friesenstraße" ('I am looking for the Friesen street'); simulated pilot-air traffic controller communication statements (non-native English); descriptions of pictures; a PowerPoint guided, but non-scripted 20 minutes presentation in front of 50 listeners. A well established, standardised subjective sleepiness questionnaire, the Karolinska Sleepiness Scale (KSS, [4]), was used by the subjects (self-assessment) and by the three assistants who had supervised the experiments, using all available information (audio/video/context); they had been formally trained to apply a standardised set of judging criteria. Scores range from 1 to 10: extremely alert (1), very alert (2), alert (3), rather alert (4), neither alert nor sleepy (5), some signs of sleepiness (6), sleepy, but no effort to stay awake (7), sleepy, some effort to stay awake (8), very sleepy, great effort to stay awake, strug-

gling against sleep (9), extremely sleepy, cannot stay awake (10). The labels were given not to single turns but to 'recording units' consisting of up to 20 turns (9.4 on average) such as stories or sequences of commands. This constitutes an optimal and smooth reference; accordingly, mean pairwise Pearson correlation between self-assessment and observers is very high: 0.89 (0.88 between two observers). The scores from self-assessment and observers are averaged to form the reference sleepiness values (mean/standard dev.: 6.1 ± 2.3 for females, 5.9 ± 2.5 for males). A more detailed description of the data is given in [5, 6]. For the 2011 Challenge, a subdivision of the data into three speaker-disjunct sets for training, development and test was defined. Here, we always report the results on the test set (*TEST*, 19 females, 14 males, 2466 turns, 6.6 hours), estimating parameters on the union of the original training and development set (henceforth *TRAIN*, 37 females, 29 males, 5279 turns, 13.2 hours). Gender is a bit imbalanced: 73% and 64% of the utterances of TRAIN and TEST, respectively, are from female speakers.

3. Features

We employ 3705 acoustic-prosodic features described in [7]; here, we only can give the general idea. For segmenting pauses, vowels, consonants, and speaker noise, we use the phoneme recognizer of the Brno Univ. of Technology [8]. Then, pseudo-syllables are derived in four different ways, taking: (1) the nucleus (i. e. consecutive vowels), (2) nucleus + coda (consecutive vowels plus trailing consecutive consonants), (3) onset + nucleus (leading consonants plus consecutive vowels), and (4) onset + nucleus + coda (leading consonants plus consecutive vowels plus trailing consonants – these syllables overlap). We compute four low-level descriptors on a frame-by-frame basis: F0, formants, formant bandwidths, and Mel frequency cepstral coefficients (MFCC) as a more fine-grained and robust, yet less explicit representation of articulators. For each syllable, we compute micro-structural prosodic descriptors such as loudness [9]. F0 is suitably interpolated, normalized per utterance, and perceptually transformed. Normalized versions of energy and duration remove phoneme-intrinsic influences. To obtain a fixed number of features per utterance, we compute twelve functionals that characterise the statistical and temporal properties of these local descriptors: mean, standard deviation, minimum, maximum, median, quantiles 5%, 25%, 75%, 95%, average absolute local change (similar to Grabe's raw pairwise variability index rPVI [10]), root average squared local change, and slope of the regression line. Depending on the type of descriptor, these functionals are computed across syllables and vocalic/consonantal frames. Additionally, we compute features developed for describing speech rhythm [10, 11, 12]. From this brute-force set, we now manually select and combine the following features suitable to capture the acoustic correlates of sleepiness that can be expected according to the pertinent literature (see e. g. [3] for an overview). The expected sign given in parentheses indicates falling/lower ('$-$') or rising/higher ('$+$') values for sleepy speech; an appended question mark indicates ambiguous tendencies.

spectral features: formants

(1) g-mean(F1–4)_V_mean ($-$): the ***geometric mean of formants F1–F4*** per frame, averaged across vocalic frames. The circadian rhythm includes body core temperature variation [13]; we can assume some decrease with sleepiness. With that, the temperature of the exhaled air drops, too. Therefore, formant frequencies should be shifted slightly downwards [14, 15].

(2) mean(FBW1–4)_V_mean ($+$): the ***arithmetic mean of the formant bandwidths FBW1–FBW4 per frame***, averaged across vocalic frames. Reduced body temperature and muscular relaxation might lead to vocal tract softening and stronger dampening of the signal due to yielding walls [16]. We expect glottal loss and cavity-wall loss for the lower formants, and radiation, viscous and heat-conduction loss for the higher formants [17]. Consistent with that is the increased time of high values for Formant 1 bandwidth in [6].

(3) F1_V_std * F2_V_std ($-$): the ***product of the standard deviations of F1 and F2*** across vocalic frames. The reduced cognitive processing speed going along with sleepiness might lead to impaired neuromuscular motor coordination processes, slowing down the transduction of neuromuscular commands into articulator movement and affecting the feedback of articulator positions [18, 19], possibly leading to aversion of spending compensatory effort [20]. Thus, sleepy speech could exhibit slurred, less crisp pronunciation, mispronunciations, abrupt articulatory changes, speech errors, or hesitations. A less crisp pronunciation might result in vowel centralization and a reduced area covered by the first two formant frequencies, which account for most of the discriminability across vowels.

(4) F1_V_mean ($-?$): the ***average of F1 across vocalic frames***. Sleepy speech is also expected to exhibit changes in speech quality such as tensed, nasal, or breathy speech due to, e. g. impaired coordination of velum closure [21]. The effects of increased nasality are complex: the first formant (F1) gets weaker, and its position moves higher, because nasals are usually pronounced more open. Yet, F1 is likely to be masked by the appearance of the lower and louder first nasal formant, resulting in an opposite tendency. Thus, nasality seems to be difficult to quantify with a simple acoustic parameter. The decrease in F1 for sleepy speech reported in [6] thus cannot be readily assessed: most likely, the decrease is due to the first nasal formant showing up more, and possibly also to reduced body temperature as shown above, these two effects outweighing the opposite tendency due to the expected more open pronunciation.

spectral features: MFCC

(5) MFCC2_V_mean ($+?$): the ***average of the second MFCC coefficient*** across vocalic frames as an estimate of the negative spectral tilt; we expect the spectral tilt to fall with sleepiness, and thus a rise of this feature. Increased breathiness in sleepy speech – see feature (4) – should lead to a negative spectral tilt for high frequencies [22], which seems to be confirmed by [23] where a decrease of the slope of the long term average spectrum is reported. An opposite effect could be caused, however, by a stronger high-pass effected by a more closed mouth position (centralisation) compatible with reduced muscular tension.

(6) MFCC1_V_mean / MFCC1_C_mean ($-$): the ***ratio of the first MFCC coefficient averaged across vocalic frames to its average over consonantal frames.*** The abovementioned losses in resonance (muscular relaxation and reduced body temperature) could lead to a reduced energy in vocalic segments compared to consonantal segments; the first MFCC coefficient is a measure of energy.

(7-10) MFCC2_V_std, MFCC3_V_std, MFCC2_C_std, MFCC3_C_std ($-$): the ***standard deviations of the second and third MFCC computed separately across vocalic and consonantal segments.*** These features describe coarsely the spectrum of the vowels (F1, F2) and of the consonants and can capture less diligent pronunciation (centralisation) [7].

prosody: F0

(11) F0_V_mean ($-?$): the ***average of pitch estimates across vocalic frames.*** The muscular relaxation going along with

sleepiness might lead to a reduced fundamental frequency (F0) as reported in [6, 24, 25], although [26] report the opposite. Since increased breathiness should also go along with reduced F0, there is one more reason to assume a decrease.

(12) F0_V_std / F0_V_mean ($-?$): the ***standard deviation of F0, normalized to the mean F0***, across vocalic frames. For sleepy speech, we anticipate monotonic and flattened intonation [3]: [27, 25] report a decreased standard deviation of F0 although an opposite result has been published in [28]. The standard deviation of pitch is correlated with the absolute pitch level of a speaker; this effect is removed by the normalization.

(13) syl-F0-mean_std ($-?$): the ***standard deviation of the syllables' average F0***; here and in the following, we use the 'nucleus + coda' pseudo-syllables. Now we apply our microstructural prosodic features, where F0 undergoes a different normalization, and perceptual scaling.

(14) syl-F0-max_mean ($-?$): the ***syllables' F0 maxima*** averaged across syllables. Flattened intonation should lead to less pronounced F0 maxima.

(15) syl-F0-min_mean ($+?$): the ***syllables' F0 minima*** averaged across syllables which should rise.

(16) syl-F0-slope_mean ($-$): the ***F0 slope within syllables***, averaged across syllables, expected to fall with sleepiness because of flattened intonation.

prosody: energy

(17) syl-energy-mean-norm_std ($+?$): the ***standard deviation of the syllables' normalized mean energy.*** According to [6], the average absolute deviation of intensity increases with sleepiness. This could be explained as a less diligent or controlled pronunciation, although a flattened intonation might also have the opposite effect.

(18) syl-rel-energy_mean ($-$): a ***medium-term estimate of the relative energy*** (computed for energy normalization purposes [9] from up to 15 neighbouring syllables, taking into account phoneme-intrinsic properties), averaged across syllables. Muscular relaxation might also lead to reduced loudness.

(19) syl-energy-slope_mean ($-$): the ***average energy slope within syllables*** which we expect to fall with sleepiness, due to flattened intonation.

prosody: duration

(20) syl-rel-duration_mean ($+$): ***medium-term estimates of the syllables' relative durations***, averaged across syllables. Slowed cognitive processing reduces speech planning, which might lead to a reduced speech rate [29, 27, 28] and thus to increased durations.

(21-22) syl-pauses_mean, syl-filled-pauses_mean ($+$): the ***average duration of silent and of filled pauses between syllables.*** Along with segment durations, pause length is expected to increase with sleepiness, too [28].

prosody: rhythm

(23) %V ($-$): ***Ramus' %V, the percentage of vocalic intervals*** [11] is expected to fall because the relative frequency of voicing decreases with sleepiness [30].

(24-25) nPVI_V, nPVI_C ($-?$): ***Grabe's normalized pairwise variability index nPVI*** [10], a rate-of-speech-normalized measure of local durational variability, computed separately for vocalic and consonantal segments, is expected to fall with monotonicity (although more disfluencies could also lead to a rise).

(26-27) varco_V, varco_C ($-?$): ***Dellwo's variation coefficients*** [12] are a measure of global durational variability (rate-of-speech-normalized standard deviations of the duration of vocalic and consonantal segments). Again, we expect a decrease, although disfluencies could have the opposite effect.

4. Experiments and Results

4.1. Analysis of single Features

We compute Pearson's correlation coefficient r between the reference sleepiness values and the individual features of each utterance only for TEST; this guarantees strict comparability with the regression results of Section 4.2. Spearman's ρ did not differ much, so we skip it. The results are given in Table 1. These individual correlations are mostly weak; for the weakest correlations, contra-intuitive effects can arise. For instance, feature (3) is negatively correlated to sleepiness for female and male speakers separately, but positively for all speakers together – this can be due to slightly different distributions of feature range and sleepiness score for female vs. male speakers. For males, correlations are mostly stronger (average $|r|$: 0.14 vs. 0.09). Using the same database, we showed in [7] that this can mainly be attributed to females showing their sleepiness less than males do. If we disregard very weak correlations – arbitrarily defined as $|r| < 0.1$ – which might well be caused by noise, given the limited number of speakers – then only 9 out of 81 cases with 'unexpected' sign remain (typeset in italics in Table 1); thus, our predictions of Section 3 are generally corroborated. The correlations of feature (5) are negative, contradicting our expectation; a more relaxed and thus closed mouth position might outweigh the effects of breathiness. Feature (19) displays highly contradicting signs as well. Our conjecture was that flattened intonation would result in negative slopes, for both F0 and energy within syllables. This did turn out right for F0, see feature (16), but not for energy. Feature (2), an estimate of the bandwidth of formants, unexpectedly falls with increased sleepiness for male speakers. One conjecture would be interactions between changes in F0 and formant extraction; but then, these interactions should be stronger for females due to the higher distance between harmonics. Another explanation could be a stronger volitional effort to fight against sleepiness in women, which might lead to muscular tension and vocal tract hardening.

Feature (6), the ratio of the energy of voiced and unvoiced segments, unexpectedly rises with sleepiness for male speakers. An explanation could be a less diligent control of the air stream, possibly resulting in louder vowels for males who tend to show the effects of sleepiness to a higher extent than females [7]. Feature (18) unexpectedly rises for female speakers: the tendency of females to show sleepiness to a lesser extent might lead to overcompensation, resulting in louder speech with clearly articulated consonants (because of the negative sign for (6)).

4.2. Regression Experiments

For robust estimation, ridge regression [31] is used. Parameters are estimated on TRAIN, results are computed on TEST; details are given in [7]. Again, Spearman's ρ is similar to Pearson's r between predicted and reference sleepiness values, so we use only the latter. The results are given in Table 2. If all 3705 features are used, the best result is 0.41 for all speakers. Also here we see higher correlations for male speakers (0.50 vs. 0.34), consistent with the single correlations above.

For the 27 manually selected features, correlations are lower, e. g. 0.33 vs. 0.41 for all speakers, but much better than a pure random selection of 27 features, which results in a correlation of 0.15 for all speakers on average (not displayed in Table 2). When looking at spectral and prosodic features separately, there is another interesting gender effect: for men, spectral features seem to be more suited than prosodic features (0.43 vs. 0.21). It is the other way around for female speakers: here, prosodic features yield better results than spectral features (0.33

Table 1: *Manually selected features and their Pearson correlation r to sleepiness: for all speakers, females (f), and males (m). The absolute value of the correlations is illustrated by the grey level of each cell's background. Grossly unexpected correlations (different sign and $|r| \geq 0.1$) are set in italics.*

Feature	exp. sign	all	f	m
(1) g-mean(F1–4)_V_mean	−	+0.06	−0.09	−0.22
(2) mean(FBW1–4)_V_mean	+	+0.12	+0.11	*−0.21*
(3) F1_V_std * F2_V_std	−	+0.03	−0.04	−0.16
(4) F1_V_mean	−?	−0.08	−0.18	−0.24
(5) MFCC2_V_mean	+?	*−0.19*	−0.07	*−0.41*
(6) MFCC1_V_mean/MFCC1_C.	−	−0.11	−0.21	*+0.16*
(7) MFCC2_V_std	−	−0.03	+0.01	−0.23
(8) MFCC3_V_std	−	−0.18	−0.17	−0.35
(9) MFCC2_C_std	−	+0.02	+0.01	−0.01
(10) MFCC3_C_std	−	−0.14	−0.10	−0.28
(11) F0_V_mean	−?	+0.05	−0.26	−0.12
(12) F0_V_std / F0_V_mean	−?	−0.02	+0.03	−0.18
(13) syl-F0-mean_std	−?	−0.03	−0.06	+0.04
(14) syl-F0-max_mean	−?	−0.03	−0.04	+0.04
(15) syl-F0-min_mean	+?	−0.02	+0.00	−0.09
(16) syl-F0-slope_mean	−	−0.08	−0.06	−0.06
(17) syl-energy-mean-norm_std	+?	+0.02	+0.01	+0.08
(18) syl-rel-energy_mean	−	*+0.15*	*+0.14*	−0.11
(19) syl-energy-slope_mean	−	*+0.17*	*+0.22*	*+0.10*
(20) syl-rel-duration_mean	+	+0.23	+0.22	+0.23
(21) syl-pauses_mean	+	+0.01	−0.03	+0.10
(22) syl-filled-pauses_mean	+	+0.21	+0.20	+0.12
(23) %V	−	−0.05	−0.02	−0.01
(24) nPVI_V	−?	−0.07	−0.06	−0.10
(25) nPVI_C	−?	+0.03	+0.04	+0.03
(26) varco_V	−?	−0.09	−0.06	−0.17
(27) varco_C	−?	+0.01	+0.01	−0.01

vs. 0.20). As for the non-ambiguous features ('−' or '+' without '?' in Table 1), results for all speakers suffer only a little by this restriction (0.30 vs. 0.33). However, now there is hardly a difference between the performance on female and male speakers (0.29 and 0.30). This is quite different when looking at the features we just removed: Training only with the 12 ambiguous features ('−?' or '+?' in Table 1), the performance difference between male and female speakers is more pronounced than ever (0.42 vs. 0.19). A possible explanation for this could be the following: the non-ambiguous features generally model sleepiness changes based on 'physiological primitives' that cannot be controlled very well by the speaker. The ambiguous features, where we identified possible antagonistic influences, however, represent parameters where the speakers do have some choice.

For a data-driven feature selection, we use a so-called wrapper approach, together with a greedy forward search: each time that feature is added which yields the best performance when training and testing the regression system with TRAIN. Here, we discuss the comparable numbers of selected features, namely 27 and 15, respectively. Intriguingly, these yield similar performance compared to the manual feature selection: for 27 features and all speakers, 0.33; for females and males separately, correlations decrease slightly (0.30 vs. 31, and 0.35 vs. 0.40, respectively). The first 15 automatically selected features are slightly better than the 15 manual non-ambiguous features (0.32/0.30/0.35 vs. 0.30/0.29/0.30). One would normally expect data-driven selection to outperform manual selection; however, it has to cope with weak sleepiness effects, facing noisy data from a limited number of speakers, and thus the unavoidable train-test mismatch. In fact, the 9th and the 21th selected feature are our manual features (6) and (10) – a very nice outcome because the probability for this to happen by chance is very low. Generally, the data-driven and manual features are not very similar, at least when compared individually: mean pairwise Pearson correlation between data-driven and manual features is 0.04, mean absolute 0.14. Minimal correlation is -0.53; maximal correlation (apart from the two identical features) is 0.91: between the 95%-quantile of pitch estimates across vocalic frames and our manual feature (11).

Table 2: *Performance when predicting sleepiness from different features. Both male and female speakers were used in training; Pearson correlation on test is reported for all, female (f), and male speakers (m). Higher absolute correlation = darker.*

Features	all	f	m
all (3705)	0.41	0.34	0.50
manually selected (27)	0.33	0.31	0.40
– spectral (10)	0.29	0.20	0.43
– prosodic (17)	0.22	0.33	0.21
manually selected – non-ambig. (15)	0.30	0.29	0.30
– spectral (8)	0.21	0.17	0.26
– prosodic (7)	0.30	0.32	0.21
manually selected – ambiguous (12)	0.16	0.19	0.42
– spectral (2)	0.19	0.16	0.40
– prosodic (10)	0.03	0.15	0.16
data-driven selection of 27	0.33	0.30	0.35
data-driven selection of 15	0.32	0.30	0.35

5. Discussion and Concluding Remarks

Expectably, brute forcing with many features beats knowledge-based selection of features (overall performance not being too high, obviously because sleepiness can only be partly modelled by speech alone, and its indication is partly speaker-dependent/idiosyncratic). However, our knowledge-based vector is on par – and in a few cases, overlapping – with the same number of automatically selected most important features, corroborating our general hypothesis of sleepiness being a relaxation phenomenon. However, females and males display interesting and partly antagonistic tendencies: male sleepiness is mainly reflected by spectral changes towards less canonical pronunciation (centralisation, cf. the MFCC features in Table 1) whereas female sleepiness primarily implies prosodic changes such as lowered pitch (feature 11). All this is in line with our explanation in [7], cf. [32, p. 130] and [33], that women tend towards more canonical speech. Generally, the non-ambiguous 15 features seem to be more 'stable' and more uniformly used by both males and females; in contrast, the 12 ambiguous features (esp. the spectral ones) obviously offer more degrees of freedom, e.g. for females, to 'hide', and for males, to 'express' their sleepiness. Of course, these explanations are tentative and have to be corroborated with future studies and additional data.

6. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The Interspeech 2011 speaker state challenge," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 3201–3204.

[2] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech and Language, Special Issue on Affective Speech in real-life interactions*, vol. 25, no. 1, pp. 4–28, 2011.

[3] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states – A review on intoxication, sleepiness and the first challenge," *Computer Speech and Language*, vol. 27, pp. 1–30, 2013.

[4] A. Shahid and K. Wilkinson, "Karolinska sleepiness scale (KSS)," in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, Eds. Springer, 2012, pp. 209–210.

[5] J. Krajewski and B. Kroeger, "Using prosodic and spectral characteristics for sleepiness detection," in *Proc. Interspeech*, Antwerp, 2007, pp. 1841–1844.

[6] J. Krajewski, A. Batliner, and M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, pp. 795–804, 2009.

[7] F. Hönig, A. Batliner, T. Bocklet, G. Stemmer, E. Nöth, S. Schnieder, and J. Krajewski, "Are men more sleepy than women or does it only look like – automatic analysis of sleepy speech," in *ICASSP 2014, International Conference on Acoustics, Speech, and Signal Processing, May 4-9, 2014, Florence, Italy, Proceedings*, 2014, to appear.

[8] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of ICASSP*, Toulouse, 2006, pp. 325–328.

[9] A. Batliner, J. Buckow, H. Niemann, E. Nöth, and V. Warnke, "The Prosody Module," in *Verbmobil: Foundations of Speech-to-Speech Translation*, W. Wahlster, Ed. Berlin: Springer, 2000, pp. 106–121.

[10] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology VII*, C. Gussenhoven and N. Warner, Eds. Berlin: de Gruyter, 2002, pp. 515–546.

[11] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," in *Proc. Speech Prosody*, Aix-en-Provence, 2002, pp. 115–120.

[12] V. Dellwo, "Influences of speech rate on the acoustic correlates of speech rhythm. An experimental phonetic study based on acoustic and perceptual evidence," Ph.D. dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, 2010.

[13] J. Zulley, R. Wever, and J. Aschoff, "The dependence of onset and duration of sleep on the circadian rhythm of rectal temperature," *Pflügers Archiv*, vol. 391, no. 4, pp. 314–318, 1981.

[14] E. G. Richardson, *Technical Aspects of Sound: Sonic range and airborne sound*. Elsevier, 1953.

[15] F. Reif, *Fundamentals of Statistical and Thermal Physics*. Waveland, 2008.

[16] T. Ananthapadmanabha, "Aerodynamic and acoustic theory of voice production," in *Forensic speaker recognition, law enforcement and counter-terrorism*, A. Neustein and H. A. Patil, Eds. New York: Springer, 2011, pp. 309–363.

[17] B. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, pp. 195–206, 2002.

[18] D. Bratzke, B. Rolke, R. Ulrich, and M. Peters, "Central slowing during the night," *Psychological Science*, vol. 18, pp. 456–461, 2007.

[19] D. Dinges and N. Kribbs, "Performing while sleepy: Effects of experimentally-induced sleepiness," in *Sleep, Sleepiness and Performance*, T. Monk, Ed. Chicester, England: Wiley, 1991, pp. 97–128.

[20] P. Lieberman, B. G. Kanki, and A. Protopapas, "Speech production and cognitive decrements on Mount Everest," *Aviation, Space, and Environmental Medicine*, vol. 66, pp. 857–864, 1995.

[21] B. E. Kostyk and A. Putnam Rochet, "Laryngeal airway resistance in teachers with vocal fatigue: A preliminary study," *Journal of voice*, vol. 12, no. 3, pp. 287–299, 1998.

[22] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *the Journal of the Acoustical Society of America*, vol. 87, p. 820, 1990.

[23] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, and B. Schuller, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing, Special Issue "From neuron to behavior: evidence from behavioral measurements"*, vol. 84, pp. 65–75, 2012.

[24] B. Johannes, V. P. Salnitski, H.-C. Gunga, and K. Kirsch, "Voice stress monitoring in space – possibilities and limits," *Aviation, Space, and Environmental Medicine*, vol. 71, pp. A58–65, 2000.

[25] T. L. Nwe, H. Li, and M. Dong, "Analysis and detection of speech under sleep deprivation," in *Proc. of Interspeech*, 2006, pp. 17–21.

[26] R. Ruiz, P. Plantin De Hugues, and C. Legros, "Advanced voice analysis of pilots to detect fatigue and sleep inertia," *Acta Acustica united with Acustica*, vol. 96, pp. 567–579, 2010.

[27] G. O. Morris, H. L. Williams, and A. Lubin, "Misperception and disorientation during sleep deprivation," *Archive of General Psychiatry*, vol. 2, pp. 247–252, 1960.

[28] A. P. Vogel, J. Fletcher, and P. Maruff, "Acoustic analysis of the effects of sustained wakefulness on speech," *Journal of the Acoustical Society of America*, vol. 128, pp. 3747–3756, 2010.

[29] W. J. M. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production," *Journal of Behavioral and Brain Sciences*, vol. 22, pp. 1–75, 1999.

[30] L. Dhupati, S. Kar, A. Rajaguru, and A. Routray, "A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings," in *Proc. IEEE Conference on Automation Science and Engineering (CASE)*, Toronto, ON, 2010, pp. 917–921.

[31] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55–67, 1970.

[32] J. Kreiman and D. Sidtis, *Foundations of Voice Studies - An Interdisciplinary Approach to Voice Production and Perception*. Wiley, 2011.

[33] P. Trudgill, "Sex, Covert Prestige and Linguistic Change in the Urban British English of Norwich," *Language in Society*, vol. 1, pp. 175–195, 1972.

Hyperarticulation in Lombard speech: A preliminary study

Juraj Šimko[1], *Štefan Beňuš*[2,3], *Martti Vainio*[1]

[1]Insitute of Behavioural Sciences, University of Helsinki, Finland
[2]Constantine the Philosopher University, Nitra, Slovakia
[3]Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
juraj.simko@helsinki.fi, sbenus@ukf.sk, martti.vainio@helsinki.fi

Abstract

Over the last century researchers collected a considerable amount of data reflecting the properties of Lombard speech, i.e., speech in a loud environment. The documented phenomena include effects on intensity, fundamental frequency, spectral tilt, speech rate and articulation. Relatively little attention has been paid to the effects on relative extent of movement of individual articulators. In an attempt to fill in this gap we present a preliminary analysis of EMA data collected in increasing levels of babble noise. We introduce HH-index as a measure of overall relative activity of articulators. Our results indicate a non-linearity of the effect of noise on articulatory movement and quantitatively different effects on the movement extent for different groups of articulators. The effects of noise are compared with those brought out by other techniques for eliciting articulatory variation. We also discuss possible application of Lombard speech as an elicitation paradigm for studies of hyperarticulation.

Index Terms: Lombard speech, hyperarticulation, articulatory variation, Slovak, EMA recordings

1. Introduction

Speakers raise their voice when they speak in environmental noise. This adaptation of speech to noise in order to increase the signal-to-noise ratio is called the Lombard effect [1] and is realized by physiological means that have different consequences on speech acoustics. Typically, the speakers increase intensity and f_0, adjust intonational contours [2, 3, 4], and their mode of vocal fold vibration is more pressed decreasing the slope of the glottal voice-source spectrum. The loud environment also affects the duration and spectral characteristics of vowels, shifting the positions of the formants [5, 6].

Compared to the vast body of research focused on acoustic and perceptual correlates of the Lombard effect, experimental articulatory investigations are relatively sparse. Studies primarily focusing on the jaw and lip movements have confirmed that Lombard speech is realized with amplification of articulatory patterns identified in "normal" speech in silent environment, and that the extent of amplification is linked to the type of noise, its loudness and the degree to which speaker's self-monitoring feedback is compromised [7]. The amplification has been shown to involve complex, non-linear reorganization of articulatory movement patterns [8].

In this preliminary study we report the results of analysis of articulatory recording including the lip, jaw and tongue movement. We focus on relative expansion of movement trajectories induced by increasing volume of babble noise as well as on global durational correlates of the Lombard effect. Furthermore, we evaluate the relative sensitivity of individual articulators on the increasing level of noise.

In articulatory terms, the Lombard effect can be interpreted as hyperarticulation. Hyperarticulation – an increase of the extent of articulatory movement – and its hypoarticulation counterpart have been widely studied as the general source of phonetic variation (H&H variation) [9] and as an important factor underlying prosodic phenomena [10]. One of the aims of this work is to assess a possibility of eliciting hyperarticulation using the Lombard effect in a controlled, quantifiable fashion. We also include two additional conditions in our recordings – a hypospeech and a non-native speaker targeting speech – and compare the effects elicited by these paradigms with those brought up by environmental noise.

In order to evaluate the effects of elicitation conditions quantitatively, we introduce a measure of relative articulatory variation, *HH-index*, capturing a proportional increase/decrease of articulatory movement in terms of articulator trajectory.

2. Methods

This study is a part of a larger investigation of the effects of prosody on articulation and inter-articulator timing. Four Slovak stimuli sentences, each containing 17 syllables, were created for a three-way manipulation of utterance-internal boundary strength, resulting in a total of 12 stimuli. With a minimum 5 intended repetitions, each block thus contained at least 60 tokens with the order of stimuli randomised within each block. Non-linguistic manipulation of prosody included the elicitation of hyper- and hypo-articulation in the following way. A reference stimulus block (subsequently referred to as condition *0dB*) was recorded in silence and the speaker was instructed to speak naturally. Three stimuli blocks were produced with three levels of babble noise at the level of 60, 70, and 80 dB – conditions *60dB*, *70dB* and *80dB*, respectively – played over the subject's headphones. Additionally, another block was elicited with no noise, where the speaker was explicitly instructed to use relaxed, hypo-articulated speech (*0dB-r* condition). Finally, the assumed highest level of hyper-articulation was elicited with 80 dB babble noise simulating a communication with a non-native interlocutor (cf. [11]) who was present and visually interacted with the subject (condition *80dB-nn*). The blocks were recorded in the following order: *70dB*, *80dB*, *0dB*, *0dB-r*, *60dB*, *80dB-nn*. Overall, we obtained 365 sentences from one subject, a native Slovak speaker with no speech or hearing impairment.

The articulatory data come from kinematic trajectories of sensors attached to 6 active articulators – lower and upper lip (LL, UL), jaw, tongue tip (TT), tongue body (TB) and tongue dorsum (TD) – obtained using electro-magnetic articulography

(EMA, AG500, Carstens Medizinelektronik, IBS, University of Helsinki). The EMA data were post-processed using TAPAD routines [12]. Audio signal was used for automatic forced alignment using the SPHINX toolkit adjusted for Slovak [13] and the time points of speech initiation and cessation were subsequently manually corrected based on the initial amplitude increase at the beginning, and the cessation of formant structure (for vowels) or voicing (for sonorants), at the end of each sentence.

To assess the quantitative articulatory characteristics for each token we calculated the approximate length of trajectory of each sensor during the utterance. That is, for each time step (determined by sampling rate of the articulatograph, 200 Hz) we calculated the Euclidean distance between subsequent positions of the given sensor in midsagittal plane. The entire trajectory of the sensor during the utterance was then calculated as a sum of these small transitions. This measure of the articulatory activity closely corresponds to the definition of Bounded Variation norm used in [4] for investigating effects of noise on fundamental frequency in Lombard speech.

Naturally, even when no global articulatory variation is elicited, the distance covered by individual articulators varies with their anatomical properties and the segmental characteristics of the stimulus. In our data, for example, the TB sensor moved on average over approx. 340 mm per token, while the UL sensor covered on average less than 70 mm. Consequently, absolute hyperarticulation effects on the distance travelled are considerably greater for "livelier" articulators than for the more restricted ones. As we are primarily interested in relative and stimulus-independent effects on articulation we have normalized the measure defined above in the following way.

The *0dB* condition – the stimuli uttered with no background noise in a natural fashion – was used as a reference. First, for each sensor we computed the mean trajectory length of the recordings of the same stimulus in this condition. Then, in order to factor out the influence of segmental and prosodic structure of different stimuli, we divided the trajectory length of every sensor for every recording in the data-set by this *0dB*-mean value of the corresponding sensor-stimulus combination. We will call the resulting measure of relative hypo-/hyper-articutation the *HH-index* for the given articulator and token. For each token, the mean of *HH-indeces* for 5 out of 6 recorded sensors, excluding UL sensor[1], is referred to as the *overall HH-index* for the token.

Naturally, the mean values of all *HH-indeces* for *0dB* condition are all equal to 1. Greater values indicate proportionally greater articulatory movement; *HH-index* value 2 means that the given articulator (or the group thereof) covered twice the distance than in the reference condition – speaker hyperarticulated. Similarly, values less than 1 correspond to hypoarticulation.

To eliminate the influence of different segmental and prosodic structure of the 12 stimuli, analogous normalization was performed for durations: again, the duration of each token was divided by the mean duration of *0dB* tokens of the same stimulus, yielding the *normalized duration* measure.

As we only have one subject and the normalizations described above removed the influence of speech material, we used a simple ANOVA-based Tukey multiple comparisons of means (with corrected p-values) implemented in R for evaluation of the effects of noise levels as well as the other two manipulations.

[1]We had to skip this sensor as we failed to recover its correct movement for multiple tokens in *70dB*-condition during the post-processing.

3. Results

3.1. Durations

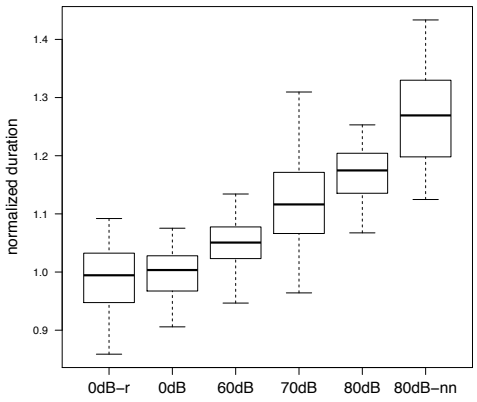


Figure 1: *Normalized durations* per condition.

Fig. 1 summarizes the effects of various conditions on duration of utterances as rendered by the subject. Tukey multiple comparisons of means showed that all differences among mean values for individual conditions are significant ($p < 0.001$), except the *0dB-r–0dB* pair ($p = 0.77$). Moreover, the pattern shown in the boxplot suggests approximately linear dependence of (non-linearly scaled) noise on the normalized duration. The differences between subsequent means, in the order captured in Fig. 1, are: 0.014, 0.050, 0.070, 0.051 and 0.096.

3.2. Overall articulatory variation

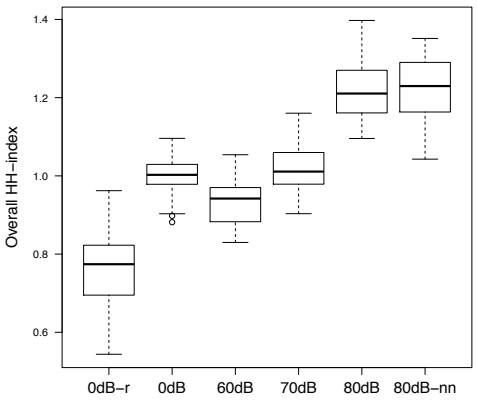


Figure 2: Overall *HH-indeces* per condition.

Fig. 2 shows the distributions of *overall HH-indeces* for individual tokens grouped by conditions. Again, most of the depicted differences are robustly significant ($p < 0.001$). The very strong effect of explicitly induced hypospeech indicates that the articulator trajectory measure as adopted in this work is a viable estimate of the magnitude of HH-variance. The difference of means between *0dB* and *0dB-r* is 0.238, the greatest among all the neighboring pairs; it means that during the "relaxed" speech the selected articulators covered almost one-quarter shorter distance than in the normal reference condition.

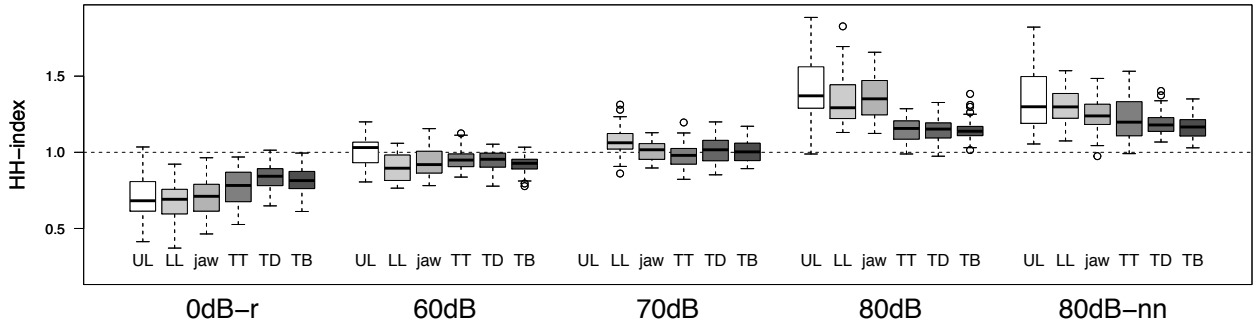


Figure 3: Sensor specific *HH-indeces* per condition.

The difference between means is not significant only in two cases. First is the difference between conditions *80dB* and *80dB-nn* ($p = 0.999$): the non-native speaker directed speech failed to elicit significantly greater articulatory movement, at least in our relatively small data-set.

Second case is the pair *0dB–70dB* ($p = 0.84$). This fact seems to be a consequence of a slightly surprising patterns present in our data. While for the three conditions with noise the mean value of *overall HH-index* significantly increases with the noise intensity, the mean value is actually significantly *lower* for *60dB* condition than for the reference quiet condition.

3.3. Sensitivity of individual articulators

We assessed the behavior of individual articulators – depicted by sensors UL, LL, jaw, TT, TD and TB – with respect to different recording conditions in two steps. First, we analyzed the effect of the conditions on each sensor separately, in a manner analogous to that in Section 3.2. Then, we compared the extents to which different sensors reacted to the influence of noise and the other two recording conditions.

Fig. 3 depicts the distributions of *HH-indeces* for individual sensors (shown in different shades of grey) and different conditions. Due to space restrictions, data for condition *0dB* are not plotted (the means for all sensors were by definition equal to 1 indicated by the dashed line). Also, the values for sensor UL are missing for *70dB*-condition for technical reasons, see Section 2.

The distributions of sensor-specific *HH-indeces* per individual conditions show, by and large, patterns very similar to that identified for the overall *HH-index* (Fig. 3.2). In general, the mean values significantly ($p < 0.001$) increase with increasing level of noise, and the comparison of *60dB* and *70dB* conditions with the quiet condition *0dB* in many cases fail to reveal a significant difference. For all sensors, the mean of *HH-indeces* for *0dB-r*-condition are significantly lower than for every other condition, and for *80dB* and *80dB-nn* conditions the means are significantly higher than for every condition with lower (or no) noise level ($p < 0.001$ in all cases).

Here we list all condition pairs for each sensor for which the difference of means of *HH-indeces* are not significantly different from 0 at $p < 0.001$ in our data set (Tukey multiple comparisons of means test).

For UL, the difference between the means for *60dB* and *0dB* is not significant ($p = 0.99$). For LL, the mean for *70dB* is significantly greater than that for *0dB* at $p < 0.01$. For both lip sensors the difference between *80dB* and *80dB-nn* is not significant ($p = 0.99, 0.86$, for UL and LL respectively).

For the jaw, the difference between *0dB* and *70dB* is not significant ($p = 0.99$), while the means for both *60dB* and *70dB* are both significantly smaller than that for *0dB* ($p < 0.05, 0.01$, respectively). The mean for *80dB* is significantly greater than that for *80dB-nn* ($p < 0.001$).

Interestingly, for TT and TD the relationship between *80dB* and *80dB-nn* is reversed compared to the jaw, the former being significantly lower that the latter ($p < 0.001, 0.05$, respectively). For TB sensor, the *80dB–80dB-nn* difference is also not significant ($p = 0.47$). For all three tongue sensors, the difference between *0dB* and *70dB* is not significant ($p = 0.63, 0.84, 0.84$, for TT, TD, TB, respectively). For TT sensor, neither are the differences between the means for *0dB* and *60dB* ($p = 0.06$) and *60dB* and *70dB* ($p = 0.83$).

Next we looked at differences in articulatory variability between individual articulators in different conditions. Table 1 summarizes the differences between the means of *HH-indeces* for all pairs of articulators (rows) organized by conditions (columns; the p-values are corrected for individual conditions). Figure 3 depicts the general trends.

First, note the very small differences between relative trajectory expansion/shrinking among tongue sensors TT, TB and

Table 1: Differences between mean values of *HH-indeces* for pairs of articulators for different conditions. Asterisks indicate significance of the difference being different from 0: ***:$p < 0.001$, **:$p < 0.01$,*:$p < 0.05$.

	0dB-r	*60dB*	*70dB*	*80dB*	*80dB-nn*
LL-UL	-0.03	-0.11^{***}		-0.10^{***}	-0.11
jaw-UL	-0.01	-0.07^{***}		-0.07^{*}	-0.17^{**}
jaw-LL	0.02	0.03	-0.07^{***}	0.03	-0.06
TT-UL	0.06	-0.05^{**}		-0.28^{***}	-0.19^{**}
TD-UL	0.13^{***}	-0.06^{***}		-0.28^{***}	-0.23^{***}
TB-UL	0.10^{***}	-0.09^{***}		-0.29^{***}	-0.25^{***}
TT-LL	0.09^{***}	0.05^{**}	-0.09^{***}	-0.18^{***}	-0.07
TD-LL	0.16^{***}	0.04^{*}	-0.06^{***}	-0.18^{***}	-0.12
TB-LL	0.13^{***}	0.02	-0.07^{***}	-0.19^{***}	-0.14^{*}
TT-jaw	0.07^{*}	0.02	-0.03	-0.21^{***}	-0.02
TD-jaw	0.13^{***}	0.01	0.01	-0.21^{***}	-0.06
TB-jaw	0.11^{***}	-0.02	0.00	-0.21^{***}	-0.09
TD-TT	0.06^{*}	-0.01	0.04	0.00	-0.04
TB-TT	0.04	-0.03	0.03	-0.01	-0.07
TB-TD	-0.03	-0.03	-0.01	0.00	-0.02

TT. The maximal difference is approximately 7 % between TT and TB for *80dB-nn* condition. Only one difference is significant (TD–TT for *0dB-r*, $p < 0.05$) in our data set.

On the other hand, there are relatively robust and in many cases significant differences between reaction of the tongue and lip-jaw articulatory systems on most conditions. The greatest differences between articulators from these groups can be seen in condition *80dB*. In this case, the lips and the jaw expanded their trajectory compared to the reference condition (*0dB*) by some 20–30 % more than the tongue articulators; all increases were significant. Smaller effect is present for *60dB* and *70dB* conditions. Interestingly, while in condition *70dB* LL trajectory expanded significantly more than tongue sensor's ones, it expanded significantly *less* than TT and TD trajectories in *60dB* condition. In the latter condition, however, UL trajectory shows significantly greater expansion of UL sensor relative to the tongue (unfortunately, data for UL in *70dB* are missing in our analysis). The jaw has not shown any significant differences compared to the tongue articulators in these two conditions.

The trend reversal suggested by LL sensor continues for the hypoarticulated condition *0dB-r*: with one exception (TT–LL), the effect is significantly greater for the tongue than for the lip-jaw articulators.

Within the lips-jaw system the results generally support the above observation of the greatest sensitivity of UL sensor compared to the LL and the jaw. The mean *HH-index* for UL is greater than both for LL (significantly so for *60dB* and *80dB*) and for the jaw (significantly for *60dB*, *80dB* and *80dB-nn*). Comparison between LL and the jaw reveals significant difference only in *70dB* condition.

Finally, for *80dB-nn* condition, the relative articulator sensitivity shows similar patterns to that of *80dB*, however, (with an exception of the UL sensor) the observed differences are generally not significant.

4. Discussion and conclusions

In the presence of a loud background noise, the utterances expand in duration and, at least in the case of 80 dB babble noise, also in the extent of articulatory movement. While temporal expansion seems to be approximately linear with logarithmic increase of the noise level, our results suggest a non-linear effect on articulatory trajectories. The overall effect on articulation seems to be negligible (or even, counterintuitively, negative) for lower noise levels, however, at the level of 80 dB the lengthening of trajectories is robust for all articulators.

Admittedly, the reported non-linearity can arise from various sources, predominantly the relatively small size of our data set limited to a single speaker. A possible source can also be the order in which the analyzed tokens were recorded: the reference, *0dB* stimuli were recorded right after the loudest *80dB* block, while *60dB* block followed the hypoarticulated *0dB-r* condition; clearly, some carryover effect could have influenced our measurements. At the same time, the lack of effect for the lower noise levels is intriguing as the subject was immersed in the noise through headphones with no self-monitoring feedback while he was wearing no headphones in quiet conditions: attenuating the external auditory feedback is to be expected to elicit a Lombard effect on its own even without the noise [14]. In any case, these initial findings warrant further investigation with additional speakers, randomized elicitation order and different ways of presenting/blocking self-monitoring feedback.

The robust effect of explicitly induced hypoarticulation (*0dB-r*) on both overall and articulatory specific *HH-indeces* in an expected direction justifies this measure as a way of quantitatively evaluating articulatory variation along HH dimension. As *HH-index* evaluates purely spatial extent of articulation and not articulatory velocity and/or duration, it can be used in a complementary fashion to other measures like the (normalized) duration used here. (Note the apparent "orthogonality" of the two measures for *0dB–0dB*-r and *80dB–80dB-nn* condition pairs.)

Furthermore, our results show that the Lombard effect is a viable methodology for eliciting global articulatory variation (more precisely, hyperarticulation) in a controlled manner: at least the loudest noise condition resulted in significant global and sensor-specific hyperarticulation patterns. In our limited data set, the other method intended to produce extra hyperarticulation – addressing non-native speaker (*80dB-nn*) – resulted in considerably longer durations but failed to elicit more overall articulatory movement compared to its nearest counterpart (*80dB*). To further evaluate and compare these two methods of triggering hyperarticulation, a condition when the subject speaks to non-native listener in quiet condition will be included in the follow up experiments.

The data analysis revealed interesting – albeit not altogether unexpected – patterns regarding relative sensitivity of articulators to conditions. Behavior of the articulators follows a plausible division to three anatomically meaningful groups: the tongue articulators, the lower lip-jaw system and the upper lip. In general, the sensors placed on the tongue exhibited mutually similar behavior as did the LL-jaw articulators, although the latter were more sensitive to noise-induced variations (at least for louder noise levels). The upper lip was still more sensitive. The greater sensitivity of the lips and the jaw is shown also for hypospeech, where the movement extent attenuated more for these articulators than for those of the tongue.

Two slightly different explanations can account for this phenomenon. It is possible that the greater extent of hyperarticulation for the lips and the jaw is specific to Lombard speech, in line with the other known correlates of the Lombard effect. Greater opening of the mouth simply contributes to better "audibility", salience in a loud environment, alongside increased loudness, pitch and spectral adjustments. The increase in motion of the visible articulators can also assist the interlocutor in parsing what has been said [15]. The observed effects on the tongue can be just a straightforward consequence of the more extensive movement of the anatomically connected jaw. Alternatively, the greater effect on the lips and the jaw compared to the tongue can be a consequence of greater freedom of the former in terms of physiological constraints and acoustic consequences of increased variation. In both aspects the tongue is more restricted than the jaw and the lips. In this case, the different sensitivity could be a hallmark of H&H variation in general, and has to be taken in consideration in research involving articulatory variation, for example, coarticulatory effects of bilabials and vowels in stressed vs. unstressed syllables. Our results provide a tentative support to both these interpretations: the differences in articulator sensitivity between *80dB* and *80db-nn* conditions for the former and the consistency of the *0dB-r* patterns with the general trend for the latter. More data and further research are required to shed more light on this issue.

5. Acknowledgements

The research leading to the results presented in this paper has received funding from the Finnish Academy, from the European Union FP7 under grant agreement n. 312382 and ERDF project RPKOM (ITMS 26240220064).

6. References

[1] E. Lombard, "Le signe de l'elevation de la voix," *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, vol. 37, no. 101-119, p. 25, 1911.

[2] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *The Journal of the Acoustical Society of America*, vol. 124, p. 3261, 2008.

[3] ——, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[4] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku, "Effect of noise type and level on focus related fundamental frequency changes," in *Proceedings of Interspeech 2012*, 2012.

[5] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, p. 917, 1988.

[6] J.-C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the lombard reflex," *Speech Communication*, vol. 20, no. 1, pp. 13–22, 1996.

[7] M. Garnier, L. Bailly, M. Dohen, P. Welby, and H. Loevenbruck, "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," in *Ninth International Conference on Spoken Language Processing*, 2006.

[8] R. Schulman, "Articulatory dynamics of loud and normal speech," *The Journal of the Acoustical Society of America*, vol. 85, p. 295, 1989.

[9] B. Lindblom, "Explaining Phonetic Variation: A Sketch of the H&H Theory," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990, pp. 403–439.

[10] K. J. de Jong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *Journal of the Acoustical Society of America*, vol. 97, no. 1, pp. 491–504, 1995.

[11] T. Cho, Y. Lee, and S. Kim, "Communicatively driven versus prosodically driven hyper-articulation in Korean," *Journal of Phonetics*, vol. 39, no. 3, pp. 344–361, 2011.

[12] P. Hoole and A. Zierdt, "Five-dimensional articulography," *Speech motor control*, pp. 331–349, 2010.

[13] S. Darjaa, M. Černák, M. Trnka, M. Rusko, and R. Sabo, "Effective triphone mapping for acoustic modeling in speech recognition." in *INTERSPEECH*, 2011, pp. 1717–1720.

[14] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *Journal of Speech, Language and Hearing Research*, vol. 53, no. 3, p. 588, 2010.

[15] J. Kim, A. Sironic, and C. Davis, "Hearing speech in noise: Seeing a loud talker is better," *Perception*, vol. 40, no. 7, pp. 853–862, 2011.

A Study of Human Perception of Intonation in Domestic Cat Meows

Susanne Schötz and Joost van de Weijer

Lund University Humanities Lab, Centre for Languages & Literature, Sweden

susanne.schotz@ling.lu.se, vdweijer@ling.lu.se

Abstract

This study examined human listeners' ability to classify domestic cat vocalisations (meows) recorded in two different contexts; during feeding time (food related meows) and while waiting to visit a veterinarian (vet related meows). A pitch analysis showed a tendency for food related meows to have rising F_0 contours, while vet related meows tended to have more falling F_0 contours. 30 listeners judged twelve meows (six of each context) in a perception test. Classification accuracy was significantly above chance, and listeners who had reported previous experience with cats performed significantly better than inexperienced listeners. Moreover, the two food related meows with the highest classification accuracy showed clear rising F_0 contours, while clear falling F_0 contours characterised the two vet related meows that received the highest classification accuracy. Listeners also reported that some meows were very easy to classify, while others were more difficult. Taken together, these results suggest that cats may use different intonation patterns in their vocal interaction with humans, and that humans are able to identify the vocalisations based on intonation.

Index Terms: Animal–Human Communication, Human Perception of Pet Prosody, Prosody of Domestic Cat Vocalisations

1. Introduction

There is much anecdotal evidence of pets – especially cats and dogs – imitating speech when interacting with humans. This is probably a learned skill used to elicit certain responses or rewards, e.g. food, from their human caretakers. Because of the position of their larynx, nonhuman mammals are able to articulate only a limited number of the vowel and consonant sounds of human language (see e.g. [1]). However, many animals can produce extensive vocal variation in duration, F_0 and sound pressure level (intensity), and should be able to adopt prosodic patterns similar to those used in human speech. Ohala [2] describes several prosodic features related to the *frequency code*, which are used in animal communication, e.g. low F_0 and resonances to signal large size and dominance.

Despite a recent increase in mammal vocalisation studies (see e.g. [3]), phonetic studies of pet vocalisations are fairly scarce, and very little is known about the prosodic aspects of pet vocalisations in pet–human communication. To what extent do pets use the frequency code when interacting with humans? Do pets learn to adopt human-like prosodic patterns, such as rising and falling intonation, when signalling different vocal messages to humans? How are prosodic patterns in pet vocalisations perceived by human listeners? Phoneticians who are pet owners can hardly avoid noticing the varied and often human-like prosodic patterns used in human-directed pet vocalisations. This study is an attempt to shed some light on these issues by examining human perception of different intonational patterns in cat vocalisations.

1.1. Cat vocalisations

The cat (Felis catus, Linneaus 1758) was domesticated 10,000 years ago, and has become one of the most popular pets of the world with more than 600 million individuals [5, 6]. Cats are social animals [4], and their interaction with humans has over a long time of living together resulted in cross-species communication that includes visual as well as vocal signals. Although there are several descriptions of the communicative social behaviour of the domestic cat (see e.g. [5, 4, 7]), the ones concerning vocalisations are scarce and often fragmented. It is still unclear how cats combine different sounds, and how they vary intonation, duration and intensity to convey or modulate a certain vocal message.

The vocal repertoire of the cat is characterised by "an indefinitely wide variation of sound and of patterning". Cat vocalisations are generally divided into three major categories: (1) sounds produced with the mouth closed (murmurs), such as the purr, the trill and the chirrup, (2) sounds produced with the mouth open(ing) and gradually closing, comprising a large variety of meows with similar [ɑːou] vowel patterns, and (3) sounds produced with the mouth held tensely open in the same position, i.e. sounds often uttered in aggressive situations, including growls, yowls, snarls, hisses, spits, and shrieks [8, 4].

1.2. The meow

In cat–human communication, the most common vocalisation is said to be the *meow* or *miaow* [9]. Nicastro [10] defines the meow as a quasi-periodic sound with at least one formant and with diphthong-like formant transitions. The duration ranges from a fraction of a second to several seconds, and the F_0 contour is generally arch-shaped with the tonal peak marking the maximum mouth opening of the opening-closing gesture. Meows can include atonal features and may be garnished with an initial or final trill or growl.

McKinley [11] divided the meow type vocalisation into four sub-patterns based on the pitch and vowels included in the sound: the mew, a high-pitched call with [i], [ɪ] or [e] quality; the squeak, a raspy nasal high-pitched mew-like call; the moan, an [o]- or [u]-like opening-closing sound; and the meow, a combination of vowels resulting in a characteristic [iau] sequence.

Cats learn to produce different meows for different purposes, e.g. to solicit feeding, to gain access to desired locations and other resources provided by humans. Each meow is believed to be "an arbitrary, learned, attention-seeking sound rather than some universal cat–human 'language'" [7]. If each cat and owner develop their own arbitrary vocal communication codes, other humans would be less able to identify meows uttered by unfamiliar cats. However, if cat vocalisations contain some kind of functional referentiality (cf. [9, 12]), i.e. that each vocalisation strongly correlates with a certain referent and also that perceiver responses correlate with the vocalisation, then ex-

perienced humans should be able to classify meows produced by unfamiliar cats fairly well.

Nicastro & Owren [9] asked naïve and experienced listeners to judge meow calls from twelve cats recorded in five different behavioural contexts (food-related, agonistic, affiliative, obstacle, and distress). Classification accuracy was modestly (but significantly) above chance, and it was suggested that meows are unspecific, negatively toned sounds that attract the attention of humans, but that humans can learn to appreciate meows as they become more experienced.

Schötz [13, 14] made a duration and F_0 analysis of 795 cat vocalisations and found that within each vocalisation type (including the meow) durations were fairly similar, but the overall F_0 variability was high, partly due to the large number of different intonation patterns.

1.3. Purpose, aims and hypotheses

The purpose of this study was to investigate human listeners' perception of domestic cat vocalisations of the same type (the meow), with similar durations, but with different intonation patterns. By asking listeners to classify a number of meows as belonging to one of two contexts: food related or vet related, our aim was to find out which intonation patterns are more often associated with food related vocalisations and which are more vet related. A larger goal was to learn more about how humans perceive prosodic cues in cat vocalisations and to increase our understanding of cat–human vocal communication.

Based on our own previous experience of these types of meows, as well as on pitch patterns used in human speech and also related to the *frequency code*, we expected the meows of both contexts to be of similar duration and mean F_0, but we expected a higher number of rising pitch patterns in the food related meows than in the vet related meows. We also hypothesised that experienced human listeners would judge the meows more often correctly than inexperienced listeners and also be more confident in their responses. Moreover, we hypothesised that meows with rising intonation patterns would more often be judged as food related meows than vet related meows.

2. Method

2.1. Material

Three young domestic cats: Donna, Rocky and Turbo (D, R and T; 1 female, 2 males, all three year old siblings from the same litter) were recorded in two different contexts: 1) in a familiar environment, i.e. in their home kitchen while waiting to be fed and 2) in an unfamiliar environment, i.e. in the waiting room (or in a car outside) of a veterinary clinic. The equipment consisted of a Sony digital HD video camera HDR-CX730 with an external shotgun microphone Sony ECM-CG50. Audio files (wav, 44.1 kHz, 16 bit, mono) were extracted with Extract Movie Soundtrack, and the vocalisations segmented, extracted and normalised for amplitude in Praat [16]. Six meows from each context produced by two of the cats (D and T) were selected as material, based on the overall recording quality and on judgements of the owner (one of the authors) of how representative the vocalisations were for each context. As one cat (R) was quiet during the recordings made in the vet context, no meows from this cat were used in the experiment. An auditive analysis of the material by one of the authors revealed that the food related meows tended to have rising tonal patterns, while veterinary related meows had slightly arched or falling intonation. In addition, we noticed slight variations in the background noise, including a few instances of background human speech, but this was judged to have a neglectable influence on the perception task.

Measures of duration and F_0 were obtained with a Praat script and manually checked. One meow was significantly shorter than the other vocalisations, but we decided to keep it in order to get a first impression of how stimulus duration would influence the perception results. The other stimuli ranged between 0.58 and 1.13 seconds in duration. All stimuli contained vowels belonging to the meow type, as described by McKinley [11], and were judged as clearly distinguishable from other common cat vocalisation types, including the purr (cf. [15]), the murmur (cf. [13]) and the chirp (cf. [14]). The longer meows were often garnished by short initial trills. Table 1 shows the duration, and the mean, minimum, and maximum F_0 values for the twelve meow stimuli. Figure 1 displays F_0 contours of the meows of the two contexts.

Table 1: *Duration (sec.) and F_0 (Hz) values for the 12 meows in two contexts (Food, Vet) by two cats (D, T).*

meow	duration	mean F_0	min F_0	max F_0	F_0 range
Food D 1	0.78	739	528	939	411
Food D 2	0.91	888	541	1003	462
Food D 3	0.27	797	782	816	34
Food T 1	1.06	532	418	582	164
Food T 2	0.85	539	423	653	230
Food T 3	1.03	567	433	640	207
Vet D 1	1.10	790	715	887	172
Vet D 2	0.80	838	764	924	160
Vet D 3	0.58	915	885	947	62
Vet T 1	1.13	510	451	589	138
Vet T 2	0.87	697	639	737	98
Vet T 3	1.02	540	487	570	83

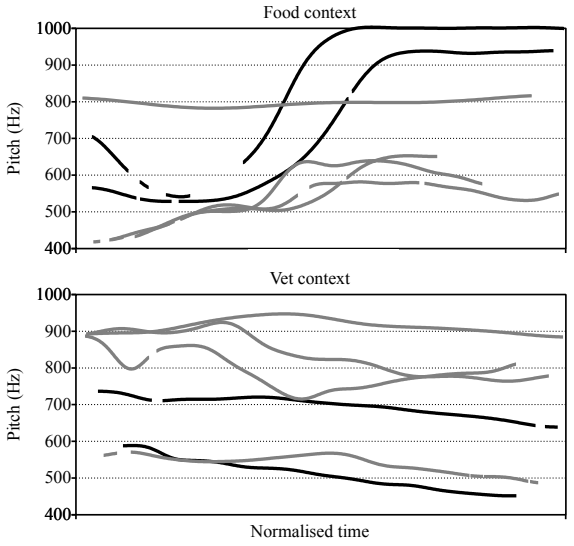


Figure 1: *Time normalised F_0 contours of the food and vet related meows. The two contours of the stimuli that received the highest proportion of correct classifications for each context in the perception experiment are drawn in black.*

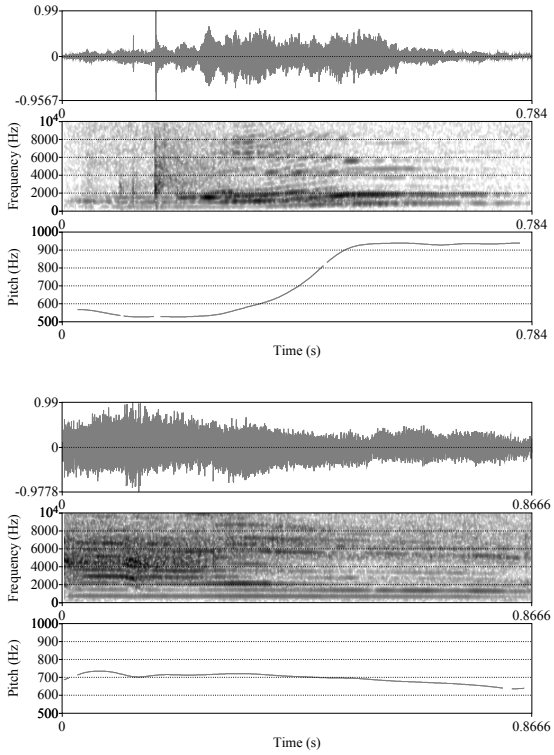


Figure 2: *Waveforms, broadband (300 Hz) spectrograms and F_0 contours of a typical food related meow (top) and a typical vet related meow (bottom).*

2.2. Procedure

The experiment was designed as a multiple forced choice identification test using the ExperimentMFC function in Praat. A group of 15 men and 15 women volunteered as participants. Their average age was 44 years (range 23 to 69 years). Of the participants, 21 reported being familiar with cats, that is, they either owned a cat at the time of testing, or they had owned a cat prior to the experiment. The time that these participants had owned a cat varied from less than one year to a maximum of 55 years (median 2.5 years). Oral and written instructions were given before the experiment, in which the task was to classify each meow as belonging to either the food context or to the vet context by clicking on the appropriate box on a computer screen. The experiment ran on an MacBook Pro computer in a quiet room. Each of the twelve meow recordings were presented three times in a randomised order through HUMP NF22A speakers or AKG K270 studio headphones at a comfortable sound level. A replay option allowed the participants to listen to each stimulus up to three times. After the test, the participants were asked to make a single judgement of the degree of certainty of their responses on a 5-point scale. Each session lasted about 3-4 minutes.

3. Results

Of all 1080 responses in the experiment 529 were food related and 551 veterinary related. In total, there were 699 correct responses (65%). The participants who reported familiarity with cats were more often correct (70%) than the participants who did not (54%).

Table 2 displays the proportions correct as well as the average reaction time for every meow stimulus. As shown in the table, there was one meow (Food D 3) that was classified incorrectly considerably more often than the other meows. This meow was exceptionally short compared to the other stimuli (cf. Table 1), and presumably contained too little information for the participants to make good judgements.

Table 2: *Percentage of correct responses and average response time (RT) for the 12 meow stimuli in the two contexts (Food, Vet) by two cats (D, T).*

meow	correct	RT (ms)
Food D 1	0.83	2342
Food D 2	0.80	2419
Food D 3	0.37	2635
Food T 1	0.54	2944
Food T 2	0.66	2673
Food T 3	0.62	2706
Vet D 1	0.63	3012
Vet D 2	0.57	2904
Vet D 3	0.68	2544
Vet T 1	0.71	2658
Vet T 2	0.71	3127
Vet T 3	0.64	3044

The F_0 contours of the two stimuli of each context category that received the highest proportion of correct classifications are the ones drawn in black in Figure 1. For the food related meows, these contours show clear rising intonation patterns, while the vet related meows that received the highest number of correct classifications generally display more falling contours.

We performed a multilevel logistic regression (with random stimulus and subject intercepts) on the results in two steps. In the first step we did not include any predictors of interest other than the intercept. The results of this analysis indicated that the overall intercept differed significantly from zero ($B = 0.7615$, $SE = 0.2529$, $z = 3.011$, $p = 0.0026$), which suggests that the overall number of correct responses was significantly above chance.

In the second step, we added the familiarity predictor to the first model. This predictor had a significant effect ($B = 0.8908$, $SE = 0.3611$, $z = 2.467$, $p = 0.0136$) and overall the second model was significantly better than the first ($\chi^2 = 5.5767$, $df = 1$, $p = 0.0182$). This suggests that the participants who were familiar with cats performed significantly better than those who were not.

We also tested whether the number of years that the participants had owned a cat was a better predictor than the familiarity, but this turned out not to be the case. In fact, number of years had a non-significant effect on the dependent variable, suggesting that participants who owned a cat for a longer period of time did not score better than those who owned a cat for a relatively short time.

The participants who were familiar with cats were not only more often correct in their answers, they were also more confident in their answers. The average confidence rating given by participants familiar with cats was 2.86, whereas that given by the other participants was 1.78. This difference was tested in a linear regression analysis, which showed that it was significant ($B = 1.0794$, $SE = 0.4133$, $t = 2.612$, $p = 0.0143$).

Finally, we examined the relation between the acoustic measurements of the stimuli shown in Table 1 and the judgements made by the participants. Given the high degree of correlation between the different F_0 variables, we used only F_0 standard deviation in combination with duration as predictors of the participant choices in a multilevel logistic regression analysis. The results showed that F_0 standard deviation was a significant predictor ($B = -0.0069$, $SE = 0.0008$, $z = -8.705$, $p = 0.0000$), while duration was not ($B = 0.3969$, $SE = 0.3502$, $z = 1.133$, $p = 0.2571$). The relation between F_0 standard deviation and the listener's judgements is visualised in Figure 3. The lower the F_0 standard deviation of the stimulus, the more often it was classified as a vet related vocalisation.

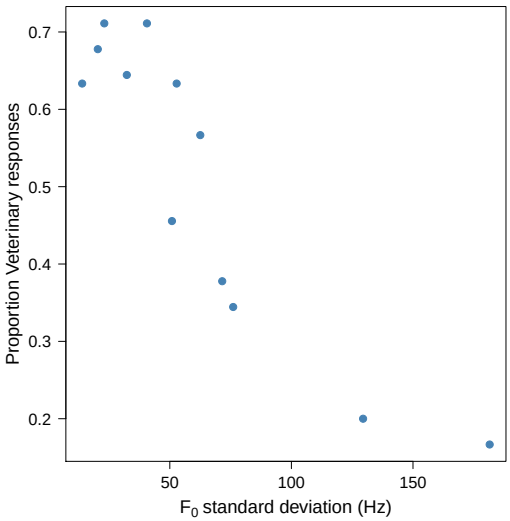


Figure 3: *Relation between F_0 standard deviation and participant choice.*

4. Discussion and future work

The results of the experiment showed that listeners were able to identify domestic cat meows from two different contexts significantly better than chance, and that experienced listeners were better judges than inexperienced ones. Moreover, there was a tendency to judge meows with rising intonation as food related, and falling intonation as vet related. Our acoustic analysis showed that the food related meows tended to have rising F_0 contours often in combination with high F_0 range, while the vet related meows often had slightly falling F_0 patterns, often accompanied by a low F_0 range. It is also possible that the listeners were influenced by these differences in F_0 range and interpreted them as expressions of different emotions; food related stimuli as happy with high F_0 range, and vet related stimuli as sad with low F_0 range.

A majority of the participants made the additional comment that some meows were quite easy to judge, while others were much more difficult. The meow with the shortest duration was often found extremely difficult to classify. Some of the listeners reported that they recognised some of the meows as similar to those of their own cats. This may suggest that different cats produce similar vocalisations in the contexts used in this study. In a future study, we will ask listeners to judge the difficulty of each individual stimulus, and also investigate the phonetic differences between vocalisations that were easy and more difficult to classify.

Several participants reported that they quickly adopted a classification strategy which they used consistently throughout the rest of the experiment even when uncertain of the success rate of this strategy. One strategy would be to listen to the intonational contours of the meows, and judge all rising patterns as belonging to one context, and all falling patterns to the other context. Another possible strategy would be to listen to the vowel quality of the meows. In this study, we did not measure formant frequencies of the vowels included in the stimuli. However, we will examine vowel quality of the cat vocalisations more carefully in future studies, and also systematically study the sound pressure level contours – including the timing of the intensity peaks – of the different meows. It is possible that we will find differences between different types or context meows.

Our study suggests that cats can learn to manipulate prosodic patterns in their vocalisations in order to better elicit the desired response from their human companions. Similarly, many humans adapt their speech or speaking style to their pets by using some kind of "pet talk" (see e.g. [17]). It is not unlikely that pets and their owners together develop a set of different prosodic patterns to improve inter-species communication. We hope to investigate this further in a future phonetic study of pet–human dialogues.

As far as we know this is one of the first phonetic studies of intonation in human-directed cat vocalisations, and there are numerous questions yet to be answered in order to better understand how cats and other pets use prosody in their vocal interaction with humans. Although this study examined a very limited number of meows from only two cats, our hypotheses that humans can judge similar cat vocalisations that differ in intonation patterns significantly better than chance and that experienced listeners perform better than inexperienced ones were confirmed. In future studies, we intend to investigate other parameters, including F_0 direction and movement, vowel quality and dynamics (diphthongisation) as well as intensity. We will also examine sounds produced to gain access to desired locations behind an obstacle (cf. [9]), and additional vocalisation types, such as the murmur and the trill, which are common in cat–human communication [13]. We will also try to include cats of a variety of breeds and cats from different countries in order to learn more about the geographical and dialectal variation in cat vocalisations.

5. Acknowledgements

The authors gratefully acknowledge support from the Linnaeus environment Thinking in Time: Cognition, Communication and Learning, financed by the Swedish Research Council, grant no. 349-2007-8695. A warm thanks to the cats Donna, Rocky and Turbo for their patience during the recordings of the material for this study. We are also very grateful to all the participants in our experiment.

6. References

[1] Fitch, W. T., "The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals", Phonetica, 57, 205—2185, 2000.

[2] Ohala, J., "An ethological perspective on common cross-language utilization of F0 of voice", Phonetica, 41:1–16, 1984.

[3] Håkansson, G. and Westander, J., "Communication in humans and other animals", Amsterdam: John Benjamins, 2013.

[4] Crowell-Davis S. L., Curtis, T. M. and Knowles, R. J., "Social organization in the cat: a modern understanding", Journal of Feline Medicine and Surgery 61:19–28., 2004.

[5] Turner, D. C. and Bateson, P. (eds.), "The domestic cat: the biology of its behaviour", Cambridge: Cambridge University Press, 2000 (2nd edition).

[6] Driscoll, C. A., Clutton-Brock, J., Kitchener, A. C. and O'Brien, S. J., "The taming of the cat", Scientific American, June 2009, 68–75, 2009.

[7] Bradshaw, J., "Cat Sense: The Feline Enigma Revealed", London: Allen Lane, 2013.

[8] Moelk, M., "Vocalizing in the House-Cat; A Phonetic and Functional Study", The American Journal of Psychology 57(2):184–205, 1944.

[9] Nicastro, N. and Owren, M. J.,"Classification of domestic cat (Felis catus) vocalizations by naïve and experienced human listeners", Journal of Comparative Psychology 117:44–52, 2003.

[10] Nicastro, N.,"Perceptual and Acoustic Evidence for Species-Level Differences in Meow Vocalizations by Domestic Cats (Felis catus) and African Wild Cats (Felis silvestris lybica", Journal of Comparative Psychology, 118(3):287–96, 2004.

[11] McKinley, P. E., "Cluster analysis of the domestic cat's vocal repertoire". Unpublished doctorial dissertation. University of Maryland, College Park, 1982.

[12] Macedonia, J.M. and Evans, C.S.,"Variation among mammalian alarm call systems and the problem of meaning in animal signals", Ethology 93:177–197, 1993.

[13] Schötz, S, "A phonetic pilot study of vocalisations in three cats", Proceedings of Fonetik 2012, Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg. pp 45–48, 2012.

[14] Schötz, S., "A phonetic pilot study of chirp, chatter, tweet and tweedle in three domestic cats", Proceedings of Fonetik 2013, 14–16 June 2013, Linköping University, Sweden, pp 65–68, 2013.

[15] Schötz, S. and Eklund, R., "A comparative acoustic analysis of purring in four cats", In: Quarterly Progress and Status Report TMH-QPSR, Volume 51, 2011. Proceedings from Fonetik 2011. Royal Institute of Technology, Stockholm, Sweden, 8–10 June 2011, 9–12, 2011.

[16] Boersma, P. and Weenink, D., "Doing phonetics by computer" [Computer program], Version 5.3.56, retrieved 15 September 2013 from http://www.praat.org/, 2013.

[17] Burnham, D., Kitamura, C. and Vollmer-Conna, U., "What's new, pussycat? On talking to babies and animals.", Science, 1435–1435, 2002.

Observation of so-called "pursed-lip" and "curled-lip" utterances in Japanese, using video and MRI images

Chunyue ZHU [1], *Toshiyuki SADANOBU*[2]

[1] School of Languages and Communication, Kobe University
[2] Graduate School of Intercultural Studies, Kobe University

chunyuez@lion.kobe-u.ac.jp, sadanobu@kobe-u.ac.jp

Abstract

The Japanese language includes utterances described by the idioms "speaking with pursed lips" and "speaking with curled lips." This study employs video and MRI imaging to examine the articulatory characteristics of these utterances ("utterances P" and "utterances C", respectively) by comparing their articulation with that of "unmarked" utterances ("utterances U"). Through doing so, we arrive at the following four conclusions: (1) For the articulation of utterance P, the lips are projected outward, and rounded by expanding in the vertical direction and narrowing in the horizontal direction. (2) For the articulation of utterance C, curling the lips is not an absolute requirement. The articulation of utterance C is similar with that of utterance P in that the lips are projected outward and rounded. (3) Utterances P and C differ in two points: (a) Lips projection accompanies the lower jaw projection only in utterances P; (b) Lips in utterance P is wider than those in utterance C. (4) The shapes the lips make in utterance P, utterance C, and utterance U can be described as a circle, a horizontal rectangle, and a horizontal oval, respectively. (5) There are many facts that contradict the accepted theory that "Rounding the lips causes both lips to project outward. In reaction to this movement, the surface of the tongue is pushed toward the rear" (Koizumi 1989).

Index Terms: "pursed-lip utterances", "curled-lip utterances", video images, MRI images

1. Introduction

Although there has been considerable research on fundamental frequency, amplitude, and timing of speech, it was not until the 21st century that researchers began to conduct close investigation into voice-quality of speech ([1]). Previous studies on various voice qualities in Japanese everyday speech adopt either quantitative approach ([1]) or qualitative approach ([2][3]). This study is a first step toward the comprehensive description of Japanese voice qualities by using qualitative and quantitative approaches both.

The specific topic of study consists of three types of articulation in contemporary Japanese communication. These three types of articulation can be described in approximate terms as the articulation of speech described in the Japanese idiom "*kuchi o togarasete mono o iu"* (*"*to speak with pursed lips"), the articulation of speech described in the idiom "*kuchi o yugamete hinan suru"* (*"*to criticize with curled lips"), and the articulation of "unmarked" common speech. However, strictly speaking these descriptions differ a little from the actual facts of the matter.

The Japanese idiom "to speak with pursed lips" refers to a childish way of speaking that expresses displeasure or dissatisfaction ([4]). However, there is another way of speaking that is highly similar to this one, used by adults to express timidity. While the intuition of native speakers tells us that these differ, for purposes of describing them in print they will be treated together here. This paper will attempt to make clear the features of these articulations of "speaking with pursed lips". We will refer to these ways of speaking as "utterance P" hereinafter.

Another Japanese idiom, "to criticize with curled lips," refers to a way of speaking that expresses a feeling of contempt. For example, the phrase "'*Ore wa ee nen kedo na,' tte, ano hito konna koto iun da yo"* (*"*He says, 'I don't care'") spoken with curled lips expresses contempt for the object. However, the intuition of a native speaker tells us that the part spoken with curled lips is not the criticism portion ("*ano hito konna koto iun da yo"*) but the speech repeated as the object of contempt ("*Ore wa ee nen kedo na"*). This paper will attempt to make clear the features of this articulation. We will refer to utterance spoken in this way as "utterance C" hereinafter.

In contrast to utterance P and utterance C, we will identify as utterance U speech that contains no feeling of displeasure, dissatisfaction, timidity, or contempt.

This study will attempt to elucidate, through empirical methods, the actual state of the articulations of these three types of utterances. We used two high-definition video cameras positioned on the front and the side of the face to record the movements of the entire face, including the speech organs, in order to examine the external properties of the speech organs and MRI imaging to record the movements of articulation in order to examine the internal properties of the speech organs, such as the shape of the vocal tract, and we analyzed the results.

2. Methods

2.1. Experiment participant

At this stage, the experiment participant is one speaker of standard Japanese who resides in the Kansai region. This is because it was not easy to find speakers who could clear the strict requirements of MRI imaging.

2.2. Spoken text

As the spoken text, we chose the following three sentences, for which the style of speaking would be easy to recall when speaking them.

a. Iyaa, chotto sorewa muzukashiisu nee. (VIDEO, MRI)
'That's hard!' (utterance P, utterance U)
b. Orewa eenenkedo na. (VIDEO, MRI)
'I don't care.' (utterance C, utterance U)
c. Sonnano muzukashii yo. (VIDEO)
'That's hard!' (utterance P, utterance C, utterance U)

2.3. Equipment used

Video camera: CANON ivis HF M32 × 2

MRI: Siemens MAGNETOM Verio (3T)

2.4. Methods of imaging

2.4.1. MRI imaging

Instead of recording video images in real time, it is possible to construct video data by repeating the same utterance 96 times, using recording that synchronizes the timing of imaging and speech using two types of trigger signals: a scan signal and a noise burst ([5][6]). In this experiment, we used this video imaging method to record images of a median sagittal plane at a resolution of 1 mm × 1 mm per pixel and a thickness of 3 mm per slice, creating a video image at 60 frames per second. We recorded the speech spoken during the experiment simultaneously.

2.4.2. Video imaging

We used two high-definition video cameras with resolutions of 1920 × 1080 pixels, one positioned in front of the speaker and one to the side of the speaker, to record speech simultaneously.

3. Qualitative analysis of the data

3.1. Qualitative analysis of the video images

Figures 1 and 2 superimpose tracings of utterance P, utterance C, and utterance U from the images recorded using the front and side video cameras, respectively. (Solid lines: utterance U, dotted lines: utterance P, dashed lines: utterance C.)

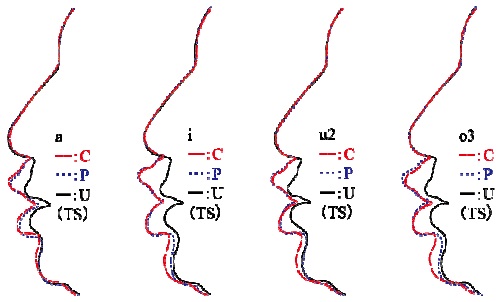


Figure 1: *Lip formations (side) of utterance P, C, and U*

Compared with utterance U, utterance P is characterized as prominent projection of lips, and utterance C of lower jaw.

From Figure 1, one can see the degree to which the lips project in comparison to utterance U and the changes in the position of the lower jaw when pronouncing each vowel /a//i//u//o/. Observation of vowels were conducted on their constant and steady region, and this is also the case with quantitative analysis.

We were able to confirm that while the lips project further in utterance P than in utterance U, the lower jaw projects only slightly. For this reason, the articulation of utterance P can be said to come mainly from projecting the lips.

While in utterance C the lips project to about the same degree as in utterance P, the lower jaw projects markedly forward.

Figure 2 is one example of a tracing of the outline of the lips during the steady-state parts of vowels in each utterance, recorded from the video camera set up in front of the speaker. The shaded area in the center of each image represents the opening of the lips as seen from the front (the narrowest part of the vocal tract opening), while the second line from the outside is the boundary between the lips and the mucous membranes inside the oral cavity that can be seen when the mouth is shut at rest. The properties of the lip formation when pronouncing each of the vowel sounds /a//i//u//o/ observed from these images are outlined below.

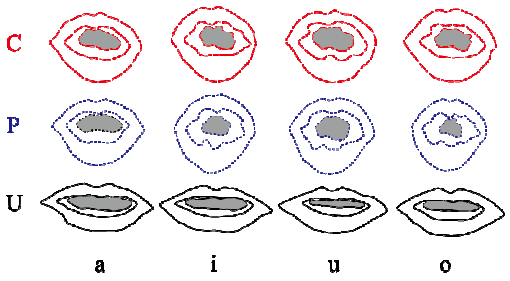


Figure 2: *Lip formations (front) of utterances P, C, and U*

The lips form, respectively, a circle, a horizontal rectangle, and a horizontal oval.

For utterance U, the lips cannot be described as being opened very wide in the vertical direction and the opening in the horizontal direction is unchanged from when the mouth is shut at rest. Using utterance U as the basis for articulation, in both utterance P and utterance C both the narrowness of the lips in the horizontal direction and the large opening in the vertical direction are prominent.

A look at the shape of the lip opening shows that in general the property of the lips rounding to form a circle is pronounced in the case of utterance P, while in utterance C the lips both round and extend in the horizontal direction.

In sum, for utterance P, utterance C, and utterance U the lips form, respectively, a circle, a horizontal rectangle, and a horizontal oval.

3.2. Qualitative analysis of the MRI images

Figure 3 uses MRI images to compare U and P pronunciation. It overlaps two tracings based on the part that

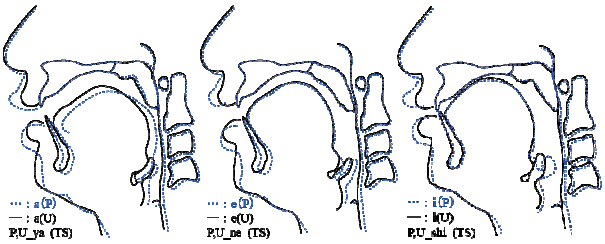


Figure 3: *Form of vocal tract in utterances U and P (MRI)*

Compared with utterance U and P is characterized as the forward projection of the lips. The mouth is openes in the vertical direction in the left and middle drawings, but not in right drawing. In utterance P, the lower jaw also often opens. The tongue doesnot show any movement toward the read in the middle and right drawings.

does not move during pronunciation (the front palate, including the upper jawbone).

From this figure, the following three points are clear.

(1) While the forward projection of the lips stands out more in utterance P than in utterance U, in some cases this accompanies the opening of the mouth in the vertical direction (as in the left and middle drawings) while in other cases it is due to projection alone (as in the right drawing).

(2) In utterance P, in many cases the lower jaw also opens. In such cases, the jaw-opening method basically is through a hinged movement with the temporomandibular joint serving as the axis.

(3) While the accepted theory holds that "Rounding the lips causes both lips to project outward. In reaction to this movement, the surface of the tongue is pushed toward the rear" ([7]), actually many facts contradict this accepted theory, since as seen in the middle and right pictures of Figure 3 in many cases the tongue itself does not show any movement.

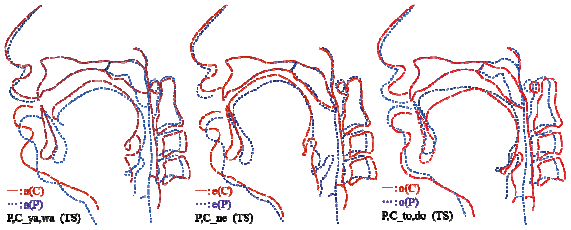


Figure 4: *Shapes of the vocal tract during utterances P and C (MRI)*
Utterances P and C have the projection of the lips in common. In utterance P the lower jaw mainly simply opens through a hinged movement of the temporomandibular joint (change in angle). In utterance C the jaw lowering involves projection of the lower joint forward. In utterance C the lips often appear to thin.

Figure 4 compares P and C pronunciation by overlapping their MRI tracings in the same way as Figure 3. As we saw in Figure 1, while utterance P and utterance C have the projection of the lips in common, they differ considerably in the way the lower jaw moves. While in the former case the lower jaw mainly simply opens through a hinged movement of the temporomandibular joint (change in angle), the latter case involves projection of the lower joint forward. In addition, in many cases in utterance C the lips appear to thin. This probably is because in the case of utterance C the lips not only project forward but also elongate in the horizontal direction. (Also see Figure 2.) Tongue back of vowels /a/ and /e/ is raised much higher in utterance C than in utterance P.

4. Quantitative analysis of the data

We measured the angles and vertical and horizontal widths of the lips in the video images taken from the front and side views and compared and analyzed the findings.

4.1. Measurement method

In our quantitative analysis of the video data filmed from the front, we measured and recorded the figures for the following five points.

(1) Horizontal width of the outer perimeter of the lips (AB in Figure 5)

(2) Height of the outer perimeter of the lips (CD in Figure 5)

(3) Horizontal width of the lip opening (narrowest part) (EF in Figure 5)

(4) Height of the lip opening (narrowest part) (GH in Figure 5)

(5) Horizontal angle of the lips (JK in Figure 5)

We also measured the following point from the video filmed from the side.

(6) Projection of lips from the face surface (MN in Figure 5)

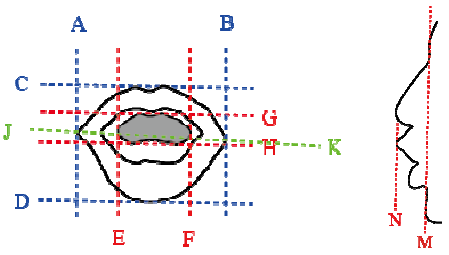


Figure 5: *Measurement of each part of the lips*

4.2. Measurement results and analysis

Selecting still images (144 in total) of the four vowel sounds /a//i//u//o/ from the video recorded from two cameras, one in front and one on the side, of the phrase *sonnano muzukashii yo* ("that's hard!") spoken six times each as utterance P, utterance C, and utterance U, we measured the six items listed under 4.1 above and recorded the resulting values. We took the average values for each of the four vowels in the three groups utterance P, utterance C, and utterance U and studied each parameter as described below.

4.2.1. Projection of lips

Identifying a line connecting the two points of the forehead and the jaw (M in Figure 5) to be the surface of the face, we measured the shortest distance from that line to the tip of the projected lips (N in Figure 5) as the projection of the lips. As shown in Figure 6, for each vowel sound the ranking from highest to lowest, of the extent of projection of the lips was as follows: utterance P > utterance C > utterance U. Since as we saw in Figures 1 and 4 the jaw moves forward in a horizontal direction in utterance C, its figure for pure projection is slightly lower than that for utterance P.

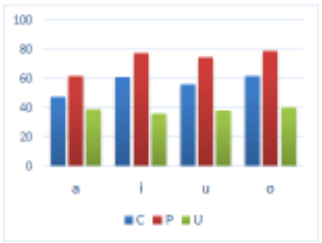


Figure 6: *Projection of lips*
The extent of projection of the lips was: utterance P > utterance C > utterance U.

4.2.2. Lip height and horizontal width

Figure 7 shows the ratio of the width of the lips' outer perimeter ((1) under 4.1) to the height ((2) under 4.1), while Figure 8 shows the ratio of the width of the lip opening at its narrowest spot ((3) under 4.1) to its height ((4) under 4.1), Since the scale on the horizontal axis shows the size of the figure for the horizontal width as a multiple of the height, the higher this number the more the lips will have a wide shape, and likewise the closer this number is to 1 the more the lips will have a rounded shape.

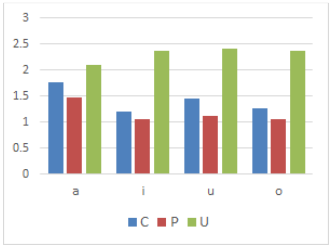


Figure 7: *W/H ratio of lip outer perimeter*

Utterance U shows the flattest shape and utterance P the closest to a circle. Although the shape for utterance C is close to that of a circle, the lips are wider than they are for utterance P.

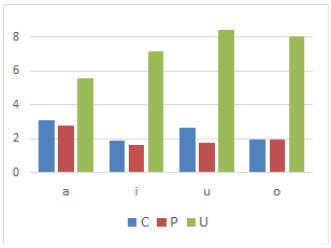


Figure 8: *W/H ratio of lip opening*

As well as Figure 7, utterance U shows the flattest shape and utterance P the closest to a circle. Although Utterance C is close to utterance P, the lips are wider than they are for utterance P.

As shown in Figures 7 and 8, it is clear that to one degree or another the vowels all show similar tendencies. That is, the shapes the lips form are similar in both outer perimeter and opening, with utterance U showing the flattest shape and utterance P the closest to a circle. Although the shape for utterance C is close to that of a circle, the lips are wider than they are for utterance P.

4.2.3. Lip angle

Figure 9 shows the average angle of the lips ((5) under 4.1) for each utterance. The scale on the horizontal axis is degree of inclination, and the fact that this value is negative for all utterances in each group means that the lips are inclined downward by 3° or more for each utterance.

As seen in Figure 9, in a so-called curled-lip utterance (utterance C), the lips actually are more curled than the others only for the vowel /a/, while for each of the vowels /i//u//o/ the lips are more curled in a pursed-lip utterance (utterance P). This suggests that curling the angle of the mouth is not an absolute requirement of utterance C.

5. Conclusions

From the studies described under Section 3 and Section 4, we reached the following five conclusions.

(1) For the articulation of utterance P, the lips are projected outward, and rounded by expanding in the vertical direction and narrowing in the horizontal direction.

(2) For the articulation of utterance C, curling the lips is not an absolute requirement. The articulation of utterance C is similar with that of utterance P in that the lips are projected outward and rounded.

(3) Utterances P and C differ in two points: (a) Lips projection accompanies the lower jaw projection only in utterances P; (b) Lips in utterance P is wider than those in utterance C.

(4) The shapes the lips make in utterance P, utterance C, and utterance U can be described as a circle, a horizontal rectangle, and a horizontal oval, respectively.

(5) There are many facts that contradict the accepted theory that "Rounding the lips causes both lips to project outward. In reaction to this movement, the surface of the tongue is pushed toward the rear" ([7]).

6. Future considerations

While there are two types of utterance P —child and adult — in natural conversation this study has not taken into consideration the differences between the two. Although we have recorded video and MRI images of all four utterances of utterance P (children), utterance P (adult), utterance C, and utterance U, the measurement process for these recordings still is underway and could not be completed in time for this paper. We would like to address the results at the next opportunity.

7. Acknowledgements

We thank our "Onsei Bunpou Kenkyukai" colleagues, especially late Miyoko Sugito for extensive advice and for helping in many other ways.

This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 23320087.

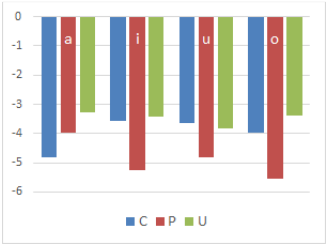


Figure 9: *Lip angle*

The lips of uutterance C are more curled than the others only for the vowel /a/, while for each of the vowels /i//u//o/ the lips are more curled in utterance P.

8. References

[1] Campbell, N., and Mokhtari, P., "Voice quality: the 4th prosodic dimension", in Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03), Barcelona, Spain, 2417-2420, 2003.

[2] Sadanobu, T., "A natural history of Japanese pressed voice", *Journal of the Phonetic Society of Japan*, 8-1: 29-44, 2004.

[3] Sadanobu, T. *Sasayaku Koibito, Rikimu Repootaa: Kuchi nonakano Bunka*. Tokyo: Iwanami, 2005.

[4] *Jitsuyou Nihongo Hyougen Jiten.* http://www.practical-japanese.com/

[5] Masaki, S., Tiede, M. Honda, K., Shimada, Y., Fujimoto, I., Nakamura, Y., and Ninomiya, N., "MRI-based speech production study using a synchronized sampling method", *Journal of the Acoustic Society of Japan. (E),* 20: 375-379, 1999.

[6] Honda, K., "MRI niyoru hatsuwakikan no kansoku", Onsei Bunpou Kenkyukai (ed.), *Bunpou to Onsei*, 5, Tokyo: Kurosio, 47-58, 2006.

[7] Koizumi, T., "Onsei to on'in", Sugito M. (ed.), *Kouza Nihongo to Nihongo Kyouiku 2: Nihongo no Onsei, On'in*, Book 1, Tokyo: Meiji Shoin, 1-20, 1989.

15 Friday 1

Prosodic Cues to Monolingual versus Code-switching Sentences in English and Spanish

Page Elizabeth Piccinini[1], *Marc Garellek*[1]

[1]Department of Linguistics, University of California, San Diego, La Jolla, California, USA

ppiccinini@ucsd.edu, mgarellek@ucsd.edu

Abstract

Code-switching offers an interesting methodology to examine what happens when two linguistic systems come into contact. In the present study, two experiments were conducted to see if (1) listeners were able to anticipate code-switches in speech-in-noise, and (2) prosodic cues were present in the signal as potential cues to an upcoming code-switch. A speech-in-noise perception experiment with early Spanish-English bilinguals found that listeners were indeed able to accurately identify words in code-switching sentences with the same accuracy as in monolingual sentences, even in highly-degraded listening conditions. We then analyzed the stimuli used in the perception experiment, and found that the speaker used different prosodic contours for code-switching productions compared to monolingual productions. We propose that listeners use specific code-switching prosody to anticipate code-switches, and thus ease processing costs in word identification.

Index Terms: prosodic contours, bilingualism, code-switching

1. Introduction

Code-switching – where two languages are used in a single utterance, sometimes switching mid sentences – is a common practice among high-proficiency bilinguals [1]. Code-switching is well documented with regards to sociolinguistic practices [2, 3, 4] and its syntactic and morphological aspects for several language pairings [5, 6, 7, 8, 9].

Psycholinguists have also studied code-switching, by examining how bilinguals deal with processing one language directly after another. Most switching studies have used an artificial form of switching, namely picture naming – where bilinguals had to switch the language in which they named a picture. [10, 11, 12, among others]. One consistent result across these studies is that bilinguals take longer to name pictures in switch trials compared to non-switch trials, regardless of the direction of the switch. A cost has also been found for processing code-switches in electrophysiological studies, both for single lexical items [13] and when comparing utterances that are either monolingual or code-switching [14, 15]. Given these costs, it seems odd that code-switching would be such a productive practice for bilinguals.

However, most of this work examined processing of words in isolation. Further, in studies that did provide full sentences, the stimuli were presented orthographically. Yet, it is possible that acoustic correlates to code-switches exist in the speech stream, and that listeners use them as cues to an upcoming switch; these cues may then reduce some of these processing costs. For example, it is well documented that English and Spanish, besides having different phonetic inventories, also have different intonation patterns. Notably, English declaratives typically have shallow-rise prenuclear and nuclear pitch accents (e.g., H* in the ToBI labeling conventions), whereas (Mexican) Spanish uses scooping rises with delayed peaks (L+>H*) in prenuclear position, but usually low-f0 targets (L*) as its nuclear pitch accent. [16, 17, 18]. In code-switching, it is possible that speakers will produce intonation patterns that do not directly map on to either language. Listeners could use this information to their advantage to anticipate upcoming language switches. For example, work on Spanish-English code-switching and narrow focus has found that while code-switching cues narrow focus, it is not enough for the sentence to include a code-switch; the sentence must also have the correct prosody for the sentence to be consistently understood as signaling narrow focus [19].

The present study asks two main questions. First, do Spanish-English bilinguals have trouble identifying words in code-switches sentences because code-switching induces high processing demands? To examine this question, we conduced a speech-in-noise study using both monolingual and code-switching utterances. If there truly is a processing deficit due to code-switches' being unexpected, then listeners should preform worse on the code-switching sentences compared to the monolingual sentences. However, if they perform equally well on code-switching and monolingual sentences, then it is possible that there are phonetic cues that allow listeners to anticipate a code-switch. As already discussed, one known difference between English and Spanish is pitch accent realization. Thus the second question of the study is, are there systematic prosodic differences between English, Spanish, and code-switching (both English to Spanish and Spanish to English) sentences? This will be investigated by examining the F0 contours of the full sentences used as stimuli in the speech-in-noise study, as well as the F0 realization of the pitch-accented words in each utterance.

2. Part 1: Perception experiment

2.1. Methods

2.1.1. Listeners

Eight early Spanish-English bilinguals (seven female, one male) of Mexican-American heritage participated in the perception experiment. The average age was 19.6 (standard deviation 1 year). Average age of acquisition of Spanish was 1.3 years (standard deviation 1.1 years); that of English was 3.1 years (standard deviation 1.8 years). We administered the Bilingual Dominance Scale [20] to all participants. The scale computes a score for how dominant the participant is in one language or another: a score of 0 means that the participant is a balanced bilingual; a high positive score mean the participant is heavily

English dominant; a high negative score means heavily Spanish dominant. The average score for the participants was 6.9 (standard deviation 7.7), suggesting that listeners were English dominant, despite the fact that their first language was Spanish. All participants were undergraduates at the University of California, San Diego, and received course credit in exchange for their participation in the study.

2.1.2. Stimuli

Sentences were modified versions of the Bamford-Kowal-Bench (BKB) sentences [21], which have been used in several past speech-in-noise studies [22, 23, 24]. The sentences were produced by a bilingual female speaker of American English and Mexican Spanish for use as English, Spanish, and code-switching sentences (code-switching English to Spanish (henceforth, **CS-ES**) and code-switching Spanish to English (**CS-SE**)). More information about the speaker's language background is provided in 3.1.1. Sentences were translated for use in the Spanish and code-switching sentences. The original BKB sentence list was created to include short sentences with vocabulary commonly used by partially-hearing children ages eight to 15. Similarly, the Spanish translations were modified as needed to include high-frequency words in Spanish, as well as be culturally appropriate for the dialect of Spanish spoken by the bilingual participants (viz., Mexican Spanish). Similar sentence structures were maintained for both the Spanish translations and the code-switching sentences. Code-switch location within a sentence was counterbalanced for syntactic position. This was done to reflect the grammatical rules of code-switching as discussed in previous research on Spanish-English code-switching [5, 6, 7]. A native Spanish-English bilingual reviewed the code-switching sentences for grammaticality and naturalness. The BKB sentences are blocked into lists of 16 sentences, each containing 50 keywords. Keywords were salient words in the sentence, generally nouns and verbs. Eight lists were used for the experiment, and half of an additional list was used for an initial practice session. Thus, there was a total of 32 test items per language context. Sentences were mixed with white noise at four signal-to-noise ratios (SNRs), -6, -3, +0, and +3 dB in Praat. There were two lists per SNR.

2.1.3. Procedure

Listeners completed a speech-in-noise task, during which they wrote down as many words as possible for each sentence presented. Listeners began with a practice session consisting of eight sentences (two from each language context) with feedback. After the practice session, listeners were presented with two blocks of four lists of sentences, one list per SNR per block. Within a given list, there was an equal number of sentences from each language context (four English, four Spanish, four CS-ES, four CS-SE). Stimuli were randomized within list and the list order was randomized within block.

2.2. Results

Regarding SNR, listeners do well in all four contexts, and perform above chance (50% correct) even at the lowest SNR of -6 dB. A logistic mixed-effects regression model was run to test for the effect of SNR on keyword identification. Keyword identification ('correct' or 'incorrect') was the dependent variable, SNR was the independent variable, and participant was included as a random slope by SNR. As expected, listeners identified words significantly better as the SNR increased [-6 vs -3: β = 0.26, z = 1.66, p = 0.10; -6 vs +0: β = 0.84, z = 4.66, $p < 0.001$; -6 vs +3: β = 1.81, z = 6.32, $p < 0.001$]. Since listeners were largely at ceiling for the higher SNR levels, a model was also run on the -6 SNR ('high noise') data alone, in order to examine more closely the effects of language context. Keyword identification ('correct' or 'incorrect') was the dependent variable, language context was the independent variable, and participant was included as a random slope by language context. These results are presented in Figure 1. Listeners were significantly better at Spanish than the three other language contexts [Spanish vs English: β = -1.53, z = -4.16, $p < 0.001$; Spanish vs CS-ES: β = -1.12, z = -3.17, $p < 0.01$; Spanish vs CS-SE: β = -1.03, z = -2.42, $p < 0.05$]; there were no significant differences among any of the other contexts.

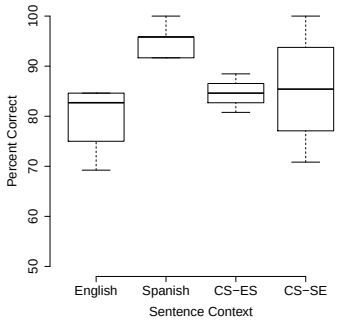


Figure 1: *Percent correct for keywords identified in each context at an SNR of -6 dB (the 'high noise' context).*

2.3. Interim discussion

The results of the perception experiment demonstrate that listeners are good at identifying words in code-switching sentences, even in listening conditions with high noise. Listeners performed best at the Spanish context. This is an interesting finding given that they are English dominant. However, this may be an effect of Spanish being their L1, thus demonstrating that current dominance is not always the most important factor. Listeners performed equally well on code-switching sentences as English sentences (their dominant language). This result seems to run counter to previous research which found a processing deficit in code-switching. This may be because previous experiments used only orthographic or pictorial switches and did not provide audio. If indeed phonetic cues in the signal can be used by listeners to anticipate a code-switch, then it is understandable that processing code-switches becomes difficult in the absence of audio information, which should result in longer processing times. It should also be noted that past studies that found a processing deficit for code-switching tested online processing, while the present study does not. However, this study does provide at least initial data suggesting that there is ***not*** a processing deficit for code-switching. To test if listeners in Part 1 performed well on code-switching stimuli because of potential phonetic cues to code-switching, in Part 2 we examine the prosodic structure of the stimuli. We hypothesize that prosodic differences – namely, differences in F0 contours over the utterance and in pitch accent realizations – may be present in code-switching contexts.

3. Part 2: F0 analysis of stimuli in Part 1

3.1. Methods

3.1.1. Speaker

One speaker produced all stimuli. She is a 22 year-old early Spanish-English bilingual having learned first Spanish at home and then English around age 6. She is a speaker of both Mexican and Peninsular Spanish, although she was instructed to produce the Spanish and code-switching sentences with a Mexican Spanish accent. She is a native speaker of Southern Californian English. She had a score of 1 on the Bilingual Dominance Scale suggesting she is a balanced bilingual. At the time of recording, she was naive to goals of the experiment. A larger-scale analysis with more bilingual speakers is underway.

3.1.2. Stimuli

Sentences were the test stimuli used in Experiment 1.

3.1.3. Method of analysis

For each sentence type (English, Spanish, CS-ES, and CS-SE), the F0 contours were analyzed both across the entire sentence and at specific points in the sentence. For the entire sentence, F0 values were extracted at 1% increments, starting at 0% into the sentence up to 100% into the sentence, resulting in a total of 101 measurements per sentence. Because content words always bore pitch accents, we also segmented each content word's stressed vowel and any adjacent sonorants to the stressed vowel within the syllable. Each sentence was uttered as one Intonation Phrase, with no IP-medial intermediate phrase break. F0 values were taken at 5% increments, starting at 0% into the stressed syllable up to 100%, resulting in a total of 21 measurements per stressed syllable. Syllables were coded as being the first stressed syllable (and thus, first pitch accent) in the Intonation Phrase, an Intonation Phrase-medial stressed syllable (sometimes there were up to three medial stressed syllables for a given Intonation Phrase), or the final stressed syllable (i.e., the nuclear-pitch-accent syllable) in an intermediate phrase (and thus, also the Intonation Phrase). Once the F0 values were obtained, outliers (based on visual inspection of the F0 tracks) were manually removed.

3.2. Results

The F0 contours across the entire sentence for the four contexts are shown in Figure 2. Visual inspection suggests that there are only two contours in use. One contour type, with a more extreme but delayed initial rise, is used in the English and CS-SE contexts. A second contour type, with a shallower and earlier initial rise, is used in the Spanish and CS-ES contexts. Towards the end of the code-switching sentences, the F0 generally maps onto the target language for the end of the sentence (i.e. CS-ES maps onto the contour for the Spanish sentence), with some deviation.

Figures 3, 4, and 5 present the contours for the first, middle, and final stressed syllables (which were also pitch-accented) in each context. To test for differences between contours, ANOVAs were conducted at 0%, 50%, and 100% for each syllable grouping. The dependent variable was F0 (in Hz) at the given time point, and the independent variable was language context. For the first stressed syllables in the sentence, there were significant effects of context at all three time points [0%: $F(3, 110) = 3.29$, $p < 0.05$; 50%: $F(3, 120) = 6.22$, $p < 0.001$; 100%: $F(3, 121) = 9.39$, $p < 0.001$]. Post-hoc Tukey tests were conducted to determine which specific contexts differed significantly from each other. English and Spanish pitch-accented vowels significantly differed from each other at the vowel's midpoint [$p < 0.001$] and end [$p < 0.001$], with F0 being higher at both time points in the English context. The F0 of CS-ES was significantly lower than that of English at the beginning of pitch-accented vowels [$p < 0.05$], but was significantly higher than that of Spanish at the vowels' end points [$p < 0.001$]. CS-SE had a significantly lower F0 than that of English midway through the vowels [$p < 0.05$].

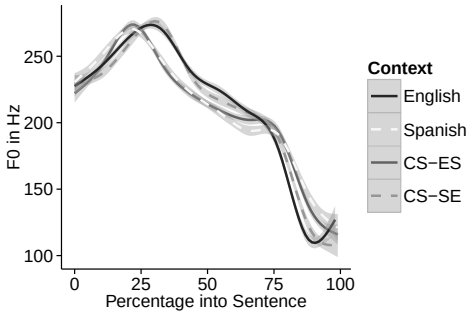


Figure 2: *Normalized F0 contour for full sentence by context. Error bars indicate one standard error.*

For the middle stressed syllables, there were significant effects of context at all three time points [0%: $F(3, 176) = 7.35$, $p < 0.001$; 50%: $F(3, 188) = 5.34$, $p < 0.01$; $F(3, 184) = 9.03$, $p < 0.001$]. Post-hoc Tukey tests were conducted to check which contexts differed from each other. English and Spanish significantly differed from each other at all three time points [0%: $p < 0.001$; 50%: $p < 0.001$; 100%: $p < 0.001$], with English having a consistently higher F0 than Spanish. CS-ES had a significantly lower F0 than English at the beginning of the vowels [$p < 0.05$], and a significantly higher F0 than Spanish at the end points [$p < 0.05$]. CS-SE also significantly differed from English at the vowel end points [$p < 0.01$], where it had a lower F0 than English.

For the final stressed syllables, there were significant effects of context at 0% and 100% [0%: $F(3, 105) = 4.32$, $p < 0.01$; 100%: $F(3, 99) = 7.65$, $p < 0.001$]. Post-hoc Tukey tests were conducted to determine which contexts differed from each other. English had a significantly higher F0 than Spanish at both beginnings and ends of vowels [0%: $p < 0.05$; 100%: $p < 0.001$]. CS-ES significantly differed from English at the end points [$p < 0.05$], where it had a lower F0 than English. CS-SE has a significantly higher F0 than Spanish at the beginning and end points [0%: $p < 0.05$; 100%: $p < 0.01$].

In sum, the stressed (and pitch-accented) syllables in the utterance differed depending on the language context. English stressed syllables generally had a higher F0 than Spanish ones. On the other hand, stressed syllables in the code-switching contexts usually had an intermediate F0 – not as high as English, but higher than Spanish.

4. General discussion

The goals of the present study were to see if (1) listeners are able to anticipate code-switches in speech-in-noise, and (2) prosodic cues are present in the signal to warn of an upcoming code-switch. The results of the perception experiment suggest

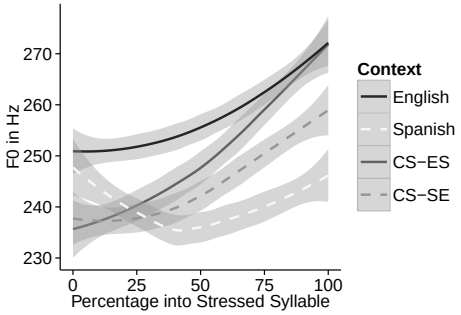


Figure 3: *Normalized F0 contour for first stressed syllable in sentence by context. Error bars indicate one standard error.*

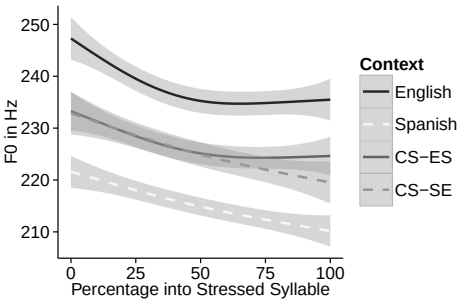


Figure 4: *Normalized F0 contour for middle stressed syllables in sentence by context. Error bars indicate one standard error.*

that listeners are able to easily process code-switches, even in a detrimental environment such as high noise. Listeners performed equally well at code-switches compared to monolingual sentences in English. This suggests that there could be something in the signal that cues listeners of an upcoming switch, which eases processing costs later on. One possibility is that listeners use syntactic and semantic information to anticipate a code-switch. This conclusion is unlikely here, as all sentences were derived from English monolingual sentences and thus could be completed in either language. Conversely, there may well be phonetic cues, either segmental or suprasegmental. The aim of Part 2 was to determine if there were prosodic cues (in the form of F0 differences) that listeners may use to anticipate a code-switch.

The results of Part 2 found that the speaker who produced the stimuli in Part 1 did in fact produce different F0 contours for code-switching sentences compared to monolingual sentences. English stressed syllables had an overall higher F0 than Spanish ones, consistent with the fact that Spanish has low nuclear pitch accents and scooping rises (with delayed high F0) in prenuclear position [17, 18]. On the other hand, English pitch accents tend to be high-toned, even in the nuclear-pitch-accented position [16]. However, in the English-to-Spanish code-switch, the speaker does not simply use English-like tone realization for the first half of the sentence and then immediately switch to Spanish-like tones for the second half of the sentence. Instead, the speaker used intonational patterns from both languages throughout her code-switching sentences. When examining the contours for the entire sentence, the speaker produced the beginning of code-switching sentences with F0 patterns that

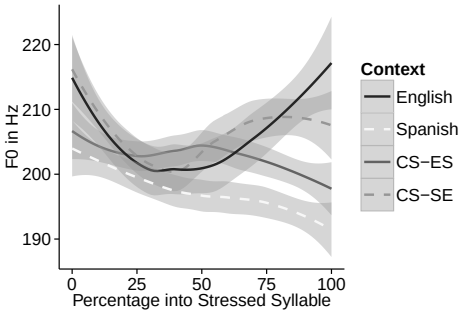


Figure 5: *Normalized F0 contour for final stressed syllable in sentence by context. Error bars indicate one standard error.*

were more similar to the post-switch language: e.g., English-to-Spanish code-switching sentences began with more Spanish-like intonation, rather with English-like intonation that became progressively more Spanish-like. This may serve as a cue to the listener of an impending code-switch. When stressed syllables of content words were examined, results were generally mixed: the speaker produced code-switching contours that never fully mapped on to either language. This can be seen most clearly in Figure 4. This is in contrast to the monolingual contours, which were consistently different from each other at almost all time points analyzed. Thus, while the speaker maintains different prosodic patterns for English and Spanish, this does not necessarily map directly onto her code-switching contours. Instead, code-switching F0 contours are in a category of their own, regardless of the direction of the switch (English-to-Spanish vs Spanish-to-English). However, another possible explanation for the differences in F0 is the segmental make-up of the words in each context, particularly the segmental make-up of stressed syllables, which could affect the timing and maximum of F0 peaks. To address this possibility, future analyses will include a phonological analysis of the sentences, to see how contexts differ in terms of tone realization and alignment. Future work will also expand upon the production study by recording additional speakers, to see if the effects shown here hold for the population at large.

5. Conclusion

The present study demonstrated that listeners are able to anticipate code-switches and thus ease processing costs in speech-in-noise word identification. An analysis of the stimuli suggested that this anticipation may in part be due to the F0 contours of the sentences as a whole, as well as those of pitch-accented words, which differ in code-switching contexts compared to monolingual contexts. Additional work on the intonation patterns of naturalistic code-switching, both in production and perception, will test whether code-switching utterances generally differ prosodically from monolingual ones in the ways found here, and whether these differences are used as cues in real-life language processing.

6. Acknowledgements

We thank our speaker for volunteering to record the stimuli and the members of the Phonetics Lab for comments and advice on this project.

7. References

[1] Gumperz, J. J., "The sociolinguistic significance of conversational code-switching", RELC Journal, 8(2):1–34, 1977.

[2] Milroy, L., and Gordon, M., "Style-shifting and code-switching", in, Sociolinguistics: Method and Interpretation, 198222, Blackwell Publishing Ltd, 2003.

[3] Woolard, K. A., "Codeswitching", in A. Duranti [Ed], A Companion to Linguistic Anthropology, 73–94, Blackwell Publishing Ltd, 2004.

[4] Lipski, J. M., "Code-switching or borrowing? No sé so no puedo decir, you know", in L. Sayahi and M. Westmoreland [Eds], Selected Proceedings of the Second Workshop on Spanish Sociolinguistics, 1–15, Cascadilla Proceedings Project, 2005.

[5] Pfaff, C. W., "Constraints on language mixing: Intrasentential code-switching and borrowing in Spanish/English", Language, 55(2):291–318, 1979.

[6] Poplack, P., "Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching", Linguistics, 18(7-8):581–618, 1980.

[7] Poplack, P., "Contrasting patterns of code-switching in two communities", in E. Wande, J. Anward, B. Nordberg, L. Steensland and M. Thelander [Eds], Aspects of Multilingualism, 51–77, Uppsala, 1987.

[8] Myers-Scotton, C., "Language contact: Why outsider system morphemes resist transfer", Journal of Language Contact - TEMA, 2:21–41, 2008.

[9] Poplack, S., Zentz, L., and Dion, N., "Phrase-final prepositions in Quebec French: An empirical study of contact, code-switching and resistance to convergence", Bilingualism: Language and Cognition, 15(2):203–225, 2011.

[10] Meuter, R. F. I., and Allport, A., "Bilingual language switching in naming: Asymmetrical costs of language selection", Journal of Memory and Language, 40(1):25–40, 1999.

[11] Costa, A., and Santesteban, M., "Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners", Journal of Memory and Language, 50(4):491–511, 2004.

[12] Schwieter, J. W., and Sunderman, G., "Language switching in bilingual speech production: In search of the language-specific selection mechanism", The Mental Lexicon, 3(2):214–238, 2008.

[13] Chauncey, K., Grainger, J., and Holcomb, P. J., "Code-switching effects in bilingual word recognition: a masked priming study with event-related potentials", Brain and Language, 105(3):161-74, 2008.

[14] Moreno, E. M., Federmeier, K. D., and Kutas, M., "Switching languages, switching palabras (words): An electrophysiological study of code switching", Brain and Language, 80(2):188-207, 2002.

[15] Proverbio, A. M., Leoni, G., and Zani, A., "Language switching mechanisms in simultaneous interpreters: An ERP study", Neuropsychologia, 42(12):1636–1656, 2004.

[16] Beckman, M. E., and Elam, G. A., "Guidelines for ToBI labeling", The Ohio State University Research Foundation, 2007.

[17] Beckman, M. E., Díaz-Campos, M., McGory, J. T. and Morgan, T. A., "Intonation across Spanish, in the Tones and Break Indices framework", Probus, 14:9–36, 2002.

[18] De-la-Mota, C., Butragueño, M., and Prieto, P., "Mexican Spanish Intonation", in P. Prieto and P. Roseano [Eds], Transcription of Intonation of the Spanish Language, 319–350, Lincom Europa, 2010.

[19] Olson, D., and Ortego-Llebaria, M., "The perceptual relevance of code switching and intonation in creating narrow focus", in M. Ortego-Llebaria [Ed], Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology, Cascadilla Proceedings Project, 57–68, 2010.

[20] Dunn, A. L., and Fox Tree, J. E., "A quick, gradient Bilingual Dominance Scale", Bilingualism: Language and Cognition, 12(3):273–289, 2009.

[21] Bench, J., Kowal, A., and Bamford, J., "The BKB (Bamford-Kowal-Bench) sentence lists for partially hearing children", British Journal of Audiology, 13(3):108–112, 1979.

[22] Bent, T., and Bradlow, A. R., "The interlanguage speech intelligibility benefit", Journal of the Acoustical Society of America, 114(3):1600–1610, 2003.

[23] Bradlow, A. R., Kraus, N., and Hayes, E., "Speaking clearly for children with learning disabilities: Sentence perception in noise", Journal of Speech, Language, and Hearing Research, 46:80–97, 2003.

[24] Van Engen, Kristin J., "Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble", Speech Communication, 52:943–953, 2010.

Speakers modulate noise-induced pitch according to intonational context

Simon Ritter, Timo B. Roettger

I*f*L Phonetik, University of Cologne

{simon.ritter; timo.roettger}@uni-koeln.de

Abstract

Recent studies have shown that speakers systematically modulate properties of voiceless segments according to intonational context. More specifically, in the absence of fundamental frequency (F0), speakers appear to adjust the Center of Gravity (CoG) and the intensity of voiceless fricatives to convey the impression of pitch. In line with these findings, the present production study extends earlier work and investigates noise-induced properties of fricatives, modulated by the intonational context. It is shown for German that the mean CoG and intensity of intended contours with a high boundary tone are higher than those produced for intended contours with a low boundary tone. Furthermore, looking at the development of CoG and intensity over the time course of the fricative, the trajectories corresponding to the boundary tones differ in intercept (CoG and intensity) and slope (intensity), i.e. reveal a steeper fall in case of a corresponding falling tone.

Index Terms: noise-induced pitch, intonation, boundary tone, truncation

1. Introduction

Segmental and suprasegmental properties of the speech signal have traditionally been described as two separate levels. More specifically, intonation has been mainly associated with the acoustic parameter of fundamental frequency (F0), which is superimposed on the segmental string to bear different communicative functions. In intonation research, segments themselves are traditionally considered to constitute a potential perturbation of F0. In particular, voiceless parts of the signal have been regarded as communicatively irrelevant for the interpretation of the meaning conveyed by the F0 contour. However, intonation contours are frequently interrupted by the lack of voiced segments. As a consequence, speakers realize F0 movements on the segmental material available, i.e. they might truncate the contour (e.g. [1]). In the case of truncation, the contour simply ends earlier. Work on German phrase-final intonation patterns provides evidence for truncation in nuclear contours consisting of a high peak followed by a fall [2]: If the voiced material available at the end of the phrase was limited (e.g. in words like "Schiff" /ʃɪf/), the falling intonation movement is not realized completely (truncated). In some cases, the fall is entirely absent. Despite missing F0 information, speakers appear to have no problem with understanding each other's communicative intentions. In fact, [2] notes that even if the final fall is missing in the F0, German listeners perceive the "word as having 'falling pitch'" [2:140].

So why can speakers understand each other even though communicative relevant F0 movements may be entirely absent? One possible answer to this question is that the signal is over specified: acoustically different parts of the signal can serve as cues for certain meanings. In particular, other parts of the signal might function as acoustic cues which correspond to the intended meanings.

Experiments with whispered speech have shown that segmental cues can be used to convey meanings that would otherwise be encoded in the F0. Listeners of Mandarin compensate for the lack of F0 in whispered speech e.g. by using duration of the syllable as an auditory cue for the contrast between lexical tones [3]. Such a usage of acoustic substitutes for F0 is not restricted to whispered speech: Recent experiments have shown that speakers consistently adjust noise-induced differences to intonational contexts in normal speech. Several experiments demonstrated that frication and aspiration noise of phrase medial and final voiceless obstruents corresponds to the expected modulation of F0 [4,5,6,7]. In particular, voiceless parts of the signal corresponding to high/rising tones exhibit higher mean Center of Gravity (CoG) and higher mean intensity values than their counterparts corresponding to low/falling tones.

Niebuhr [7] investigated polar questions and statements in German. German polar questions typically end in a rise, whereas statements typically end in a fall (in Autosegmental-Metrical terms H-% and L-%, respectively [8]). In his study, target words ending in voiceless fricatives were placed at the end of the utterance. Acoustic measurements revealed that fricatives obtained higher CoG and intensity means at the end of questions than at the end of statements. Niebuhr interpreted these differences as noise-induced correlates of the intended boundary tone. This conclusion, however, is limited to some degree: The contours not only differ in their boundary tones, they are characterized by the combination of a boundary tone and a particular pitch accent type preceding the boundary tone. In the case of the question it is a low pitch accent (L*), in the case of the statement it is either a high (H*), a rising (L+H*) or a falling (H+L*, H+!H*) pitch accent. Thus, the factor boundary tone is confounded with the pitch accent type, making a conclusion towards a direct causal link between missing F0 information and spectral differences difficult.

This is important because a considerable amount of the global contour differences between statements and questions is manifested through the pitch accent preceding the boundary tone, i.e. the F0 movement of the pitch accent is realized within the vowel. This leaves the listeners with relevant information of the tonal movement encoded directly in the F0 to distinguish sentence modalities. In turn, modulation of spectral characteristics of the fricatives might be less required to signal the contrast. To shed further light on the relation between boundary tone and noise-induced characteristics of voiceless sounds, we report on a production study on German boundary tones extending Niebuhr's findings. In this study, we keep the pitch accent type constant and vary only the boundary tone.

A further contribution of this paper is the nature of our dependent variable. Tonal events are inherently dynamic, i.e. a function of F0 developing over time. Previous work only looked at summary measures, which reflect the averages over the whole segment, i.e. mean CoG and intensity. However, CoG and intensity may change dynamically throughout the duration of a segment. Any differences of CoG or intensity found for the arithmetic mean may reflect (a) a baseline

difference of CoG (an overall difference with a similar trajectory), (b) a trajectory difference starting at the same intercept or (c) a combination of an intercept difference and a trajectory difference. A first indication of dynamic differences of such noise-induced cues was reported by [4]. CoG and intensity were measured at the beginning and the end of the aspiration of /t/. The results showed that intonational contexts with a final F0 fall exhibited steeper slopes of the highest spectral energy in the aspiration. In the present study, we explore the development of CoG and intensity over the time course of the fricative in addition to the static mean.

2. Methodology

2.1 Reading material

Six monosyllabic nouns with CVC structures served as target words with three words for each of the target fricatives (listed in table 1). The target sounds were the postalveolar and uvular voiceless fricatives (/ʃ/ and /χ/). Following [7]'s findings, those sounds were chosen in order to elicit the strongest segmental pitch effects.

/ʃ/	Fisch ('fish')	/fɪʃ/
	Tisch ('table')	/tɪʃ/
	Busch ('bush')	/bʊʃ/
/χ/	Koch ('cook')	/kɔχ/
	Loch ('hole')	/lɔχ/
	Bach ('stream')	/baχ/

Table 1: *Target words sorted by fricative*

The vowels in syllable nucleus position were phonologically short. Short vowels were chosen in order to reduce the voiced material available to realize the F0 movements before the voiceless fricatives at the end of the target words.

In addition to the six target words, we used two control words. Both contained voiced segments only. One was monosyllabic (*See* 'lake' /zeː/) and one was trisyllabic (*Brombeere* 'blackberry' /ˈbʀɔmbeːʀə/). We included these control words to elicit undisturbed F0 contours for the contexts under scrutiny.

The target words were embedded in short dialogues on everyday topics. The dialogues comprised two to three turns per speaker. Each target word occurred in two different contexts: In context A, the target word was the first noun in an *enumeration* of three or more nouns. In context B, the target word was utterance final in a *contrastive focus* statement (corrective contrast: "is it X?" "No it is *Y."*).

The syntactic structures of the critical utterances as well as the semantic-pragmatic context frames set by the preceding utterances were designed in such a way that they elicited fundamentally different types of edge contours: Context A mainly elicited a high nuclear pitch accent (L+H* or H*) on the target followed by a high boundary (H- or H-%) resulting in a plateau. Context B mainly elicited a high nuclear pitch accent (L+H* or H*) on the target with a terminal falling utterance-final movement (L-%). This contour is typical for a contrastive focus statement in German.

2.2 Participants and recording procedure

Twelve Participants (mean age = 22; 6 men; 6 women) were seated in front of a computer screen together with one of the experimenters and read aloud the mini-dialogues at a time. They were instructed to read each dialog silently first. After the silent reading they read the dialog together with the experimenter. They were instructed to read the contexts as naturally as possible.

2.3 Analyses

The recordings were digitized at a sampling rate of 44.1 kHz (16bit). All acoustic material was manually annotated. For the acoustic analysis of the segments, we identified segmental boundaries of the target word using a waveform and a wide-band spectrogram. All segmental boundaries of vowels and consonants were labeled at abrupt changes in the spectra. Intonation contours were labeled according to the GToBI annotation system [8]. Those productions of the target utterances that could not be counted as instances of one of the two contours described in §2.1 were excluded from the analysis (n = 29).

Based on the labels for the acoustic boundaries, we measured the fricative *Center of Gravity* (CoG) and *intensity*. The CoG measurements were taken on the basis of spectral slices in Praat [9]. The slices resulted from a 20 ms Hamming window and were shifted in 5 ms steps across the fricative. Of each interval, CoG measurements were taken within a frequency range of 0.5-10 kHz. The frequency range covers the main spectral characteristics of /ʃ/ and /χ/ and excludes potential F0 residuals as well as high-frequency ambient noise. Using a fixed window width resulted in a different number of data points for different segment durations (e.g. longer fricatives yield more windows). To get an equal number of data points for all tokens, we normalized the data by calculating nine normalized time points of the CoG trajectory over the fricative for each token separately. In line with [5,7,10], we assume that CoG is a suitable estimate of perceived fricative pitch. For intensity, we extracted ten normalized time points of the intensity trajectory over the fricative for each token separately (due to strong perturbations of the intensity measurements at the vowel-consonant transition, the first time point was excluded from the subsequent analyses, resulting in nine time points, analogously to the CoG measurements). Measuring the CoG and intensity at different points in time throughout the segment enabled us to analyze the time course of the spectral and intensity changes.

All data were analyzed with generalized linear mixed models using *R* [11] and the package *lme4* [12]. For CoG and intensity mean (mean of the intervals), we used a Gaussian error distribution (assuming normality). We adhered to the random effect specification principles outlined in [13] including a term for random intercepts for speakers and words, which quantifies by-speaker and by-words variability. The critical fixed effects in question were BOUNDARY TONE (i.e. H vs. L) and FRICATIVE (i.e. /ʃ/ vs. /x/), and for these fixed effects, we included random slopes for speakers and words (this quantifies by-speaker and by-word variability in the effects of BOUNDARY TONE and FRICATIVE). We tested whether the inclusion of the fixed effects BOUNDARY TONE and FRICATIVE did improve the model's prediction significantly for CoG and intensity mean via likelihood ratio tests (LRT).

To test whether the actual trajectories of CoG and intensity development throughout the fricative differed as a function of time, we performed a *Growth Curve Analysis* [GCA, 14]. GCA is a multilevel regression technique designed for analysis of time series data. It fits trajectories to multilevel polynomial curves and allows for comparison of such curves. We decided

to model CoG and intensity trajectories as second order polynomials (parabola shaped curves). Thus, for both CoG and intensity, the nine time steps entered the analysis as a second order orthogonal polynomial fixed effect (including first order polynomial). The crucial effect of interest was the interaction of BOUNDARY TONE with the FIRST and SECOND ORDER POLYNOMIAL. We included a term for random intercepts for speakers and words as well as random slopes for speakers and words for the FIRST and SECOND ORDER POLYNOMIAL interaction with BOUNDARY TONE. For all models, p-values were generated using likelihood ratio tests.

3. Results and Discussion

Figure 1 shows representative contours for both conditions, as produced by one male speaker. In the upper panel, the contours over the utterance "Wir hatten Brombeere" ('We had blackberry') are displayed. Here the difference between the conditions is clearly visible on the target word "Brombeere": A high accent on the initial syllable (marked in grey) is followed by a *low tone* in the case of the contrastive statement (left), or by a *high tone* in the case of the enumeration (right). In the lower panel, where the same contours are produced over the sentence "Wir brauchen 'nen Tisch" ('We need a table'), it can be seen that the contours on the target word "Tisch" are severely truncated. This observation is in line with [2], i.e. the intended boundary tone (H-% vs. L-%) is at least highly impoverished. In this context, in which F0 is drastically reduced, we expect spectral cues to retain the contrast between the two intonation contours and their meanings.

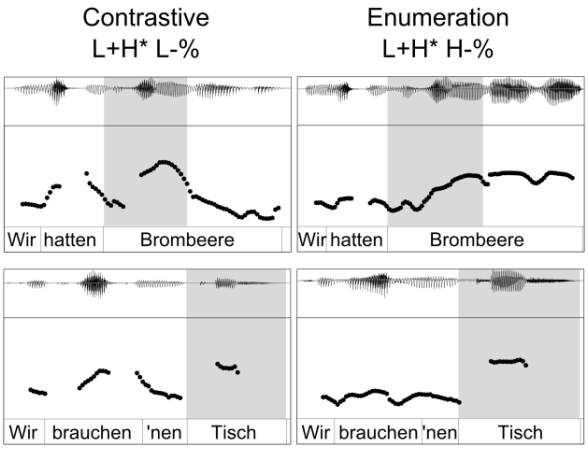


Figure 1: *F0 contours for the utterances containing the control word "Brombeere" (top) and the target word "Tisch" (bottom) in the two conditions Contrastive Focus (left) and Enumeration (right). The accented syllable is in grey.*

Mean CoG and intensity values are displayed in Figure 2. For both intensity and CoG mean, there was a significant effect of fricative, such that uvular fricatives /x/ had a 7.4 dB lower intensity (β=7.2 dB, SE=1.7, $\chi^2(1)$=11.5, p<0.0007) and a 1612 Hz lower CoG mean (β=1629.3 Hz, SE=241.3, $\chi^2(1)$=16.43, p<0.0001) than postalveolar fricatives /ʃ/. Crucially, there was a significant effect of BOUNDARY TONE on mean CoG and intensity, such that H tones elicited 4.0 dB higher mean intensities (β=4 dB, SE=1.2, $\chi^2(1)$= 8.2, p=0.00414) and 366.3 Hz higher CoG means (β=315.6 Hz, SE=120.2, $\chi^2(1)$= 5.66, p=0.0174) than L tones.

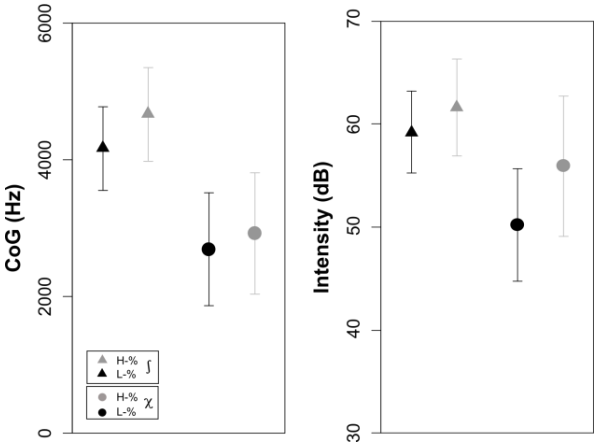


Figure 2: *CoG and intensity means and standard deviations for both fricatives and boundary tones (black = low boundary; grey = high boundary).*

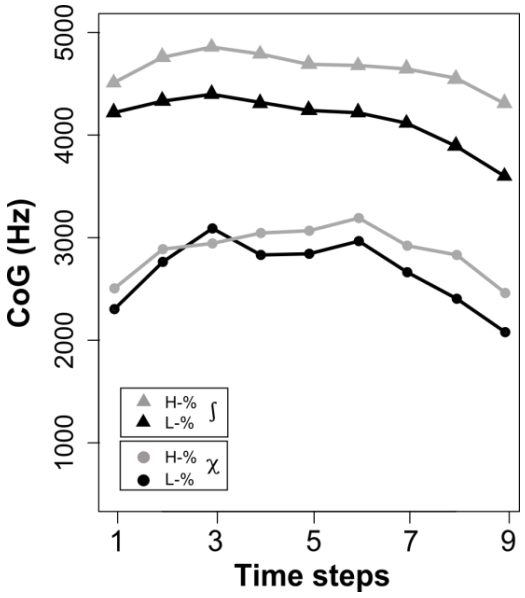


Figure 3: *CoG development as a function of time for both fricatives and boundary tones. (black = low boundary; grey = high boundary).*

To investigate the development of the measurements over time we performed growth curve analyses. As can be seen in Figure 3 and 4, for both CoG and intensity there are intercept differences, i.e. CoG and intensity start lower in the case of L-%. This difference appears to become stronger over time for intensity, that is, intensity trajectories have slightly steeper slopes for L-%. Looking at the CoG trajectories, there appear to be only small slope differences between L-% and H-%, mainly manifested in the last three time steps.

This is reflected in a significant interaction effect of BOUNDARY TONE with the FIRST ORDER POLYNOMIAL (the linear component of the models) for intensity ($\chi^2(1)=14.2$, $p=0.0002$) such that L tones corresponded to intensity trajectories with a steeper negative slope. This interaction is not significant for CoG ($\chi^2(1)=2.2$, $p=0.14$), although numerical trends point towards a comparable slope difference. Even though the SECOND ORDER POLYNOMIAL (the square components of the models) ($\chi^2(1)<0.35$, $p>0.55$) did not significantly interact with BOUNDARY TONE there were numeric tendencies suggesting that the trajectories elicited by H-% are slightly less curved than the trajectories elicited by L-%, or in other words: flatter. It is important to note that even though we found significant differences between the conditions, CoG and intensity decreases over time for both low and high boundary tones.

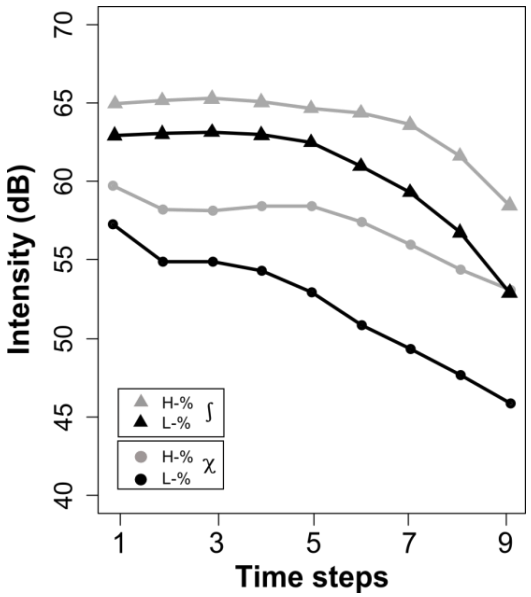


Figure 4: *Intensity development as a function of time for both fricatives and boundary tones. (black = low boundary; grey = high boundary).*

The present study has demonstrated that German speakers systematically modulate properties of voiceless segments to convey the meaning encoded in the intonation contours. Specifically, we have found that the spectral properties of voiceless fricatives, as reflected in the measure of CoG and intensity, are modulated in different intonational contexts: Higher CoG and intensity means are found for fricatives at the end of phrases ending with a high boundary tone. Thus, this study replicates earlier findings reported by [7], circumventing the confounding of boundary tone and pitch accents, and confirms that German speakers produce noise-induced correlates of intended boundary tones [4,5,6,7].

Crucially, the obtained mean differences can be ascribed to differences in development over time, at least for intensity: Intensity of high boundary tones starts higher (higher intercept) and remains flatter (flatter slope, less curved) than those of low boundary tones. CoG differences appear to be mainly due to intercept differences, i.e. CoG for high boundary tones starts higher than those of low boundary tones with a comparable development of the trajectory over time.

4. General Discussion

Grabe [2] noted that German native listeners hear truncated contours as falling even if an explicit fall in the F0 is missing. So listeners infer communicative intention even though relevant F0 movements may be entirely absent. This might be possible due to a highly over specified signal. The present study demonstrates that voiceless parts of the signal, which are not able to convey F0 information, bear their own acoustic dimensions which correspond to the missing F0 information. The question arises as to whether these acoustic differences have any communicative function – in other words, whether speakers are able to use these cues to distinguish e.g. sentence modalities. A semantic differential task performed by [4] has demonstrated that noise-induced cues of /t/ aspiration were able to shift the attitudinal meaning of the stimuli towards the meaning profile of the respective intonational context. This is a first indication that speakers, indeed, might be able to use such subtle cues communicatively. Future research is needed to further elaborate the impact of noise-induced pitch on communication.

Generally, research into this phenomenon would benefit from cross-linguistic investigations. It is important to note that the acoustic differences reported here are very subtle and prone to variation necessarily limiting its examination. Other languages could in fact be better suited for such investigations. For example, Tashlhiyt Berber, an Afroasiatic language spoken in Morocco, can have whole utterances with neither a vowel nor a voiced segment. As a result, the phonetic opportunity it affords for the execution of intonational pitch movements is exceptionally limited. In fact, it has been reported that entire complex tonal movements (Rise-Fall) can be missing in certain phonotactic environments [15,16]. Speakers of such languages might rely heavily on other cues than F0 to capture the meaning encoded in the intonation contour.

To conclude, the present findings suggest that the traditionally separated levels of analysis – segmental and suprasegmental – are strongly intertwined. Voiceless segments have been regarded as irrelevant for the interpretation of the F0 contour. They have been treated as elements to be ignored. The present findings, however, demonstrate that these parts of the signal contain acoustic dimensions that potentially contribute to the perception of the intonation contour. Thus, the results may provide an answer to the question why German listeners perceive a truncated contour as falling although there is no fall in the F0 contour [2].

5. Acknowledgements

We would like to thank Oliver Niebuhr for his valuable comments and suggestions and Bodo Winter for his statistical advice on our analyses.

6. References

[1] Erikson, Y. and M. Alstermark, "Fundamental Frequency correlates of the grave word accent in Swedish: the effect of vowel duration", Speech Transmission Laboratory, Quarterly Progress and Status Report, 2-3, KTH, Sweden, 1972.

[2] Grabe, E., "Pitch accent realisation in English and German", Journal of Phonetics, 26: 129-144, 1998.

[3] Liu, S. and A.G. Samuel, "Perception of Mandarin Lexical Tones when F0 Information is Neutralized", Language and Speech, 47: 109-138, 2004.

[4] Niebuhr, O., "Coding of intonational meanings beyond F0: evidence from utterance-final /t/ aspiration in German", Journal of the Acoustical Society of America, 124: 1252–1263, 2008.
[5] Niebuhr, O., "Intonation segments and segmental intonations", Proc. of the 10th Interspeech conference, Brighton, UK, 2435-2438, 2009.
[6] Niebuhr, O., C. Lill., J. Neuschulz, "At the segment-prosody divide: The interplay of intonation, sibilant pitch and sibilant assimilation", Proc. of the 17th ICPhS, Hong Kong, China, 1478-1481, 2011.
[7] Niebuhr, O., "At the edge of intonation: the interplay of utterance-final F0 movements and voiceless fricative sounds", Phonetica, 69: 7–27, 2012.
[8] Grice, M. and S. Baumann, "Deutsche Intonation und GToBI", Linguistische Berichte, 191: 267–298, 2002.
[9] Boersma, P, "Praat, a system for doing phonetics by computer", Glot International, 5, 341-345. 2002.
[10] Traunmüller, H. "Some aspects of the sound of speech sounds", in M. E. Schouten [Ed.], The psychophysics of speech perception, 293–305, Nijhoff: Dordrecht, 1987.
[11] R Core Team, "R: A Language and Environment for Statistical Computing", http://www.R-project.org/, 2012.
[12] Bates, D., M. Maechler, B. Bolker, "lme4: Linear mixed- effects modelsm using S4 classes", R package version: 0.999999-0, 2012.
[13] Barr, D. J., R. Levy, C. Scheepers, H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal", Journal of Memory and Language, 68: 255-278, 2013.
[14] Mirman, D., J.A. Dixon, J.S. Magnuson, "Statistical and computational models of the visual world paradigm: Growth curves and individual differences", Journal of Memory and Language, 59, 475-494, 2008.
[15] Röttger, T. B., R. Ridouane, M. Grice, "Sonority and syllable weight determine tonal association in Tashlhiyt Berber", Proc. of 6th International Conference on Speech Prosody, Shanghai, 2012.
[16] Röttger, T. B., R. Ridouane, M. Grice, "Phonetic alignment and phonological association in Tashlhiyt Berber", *Journal of Acoustical Society of America*, 133: 3572, 2013.

US English attitudinal prosody performances in L1 and L2 speakers

Albert Rilliard [1], *Donna Erickson* [2], *Takaaki Shochi* [3], *João Antônio de Moraes* [4]

[1] LIMSI-CNRS, France, [2] Kanazawa Medical University & Sophia University, Japan
[3] CLLE-ERSSàB UMR 5263, France [4] Laboratório de Fonética Acústica, FL/UFRJ/CNPq, Brazil
rilliard@limsi.fr, ericksondonna2000@gmail.com, shochi38@gmail.com, jamoraes3@gmail.com

Abstract

Expressive behavior linked to paralinguistic meanings finds grounds in codes proposed as universals, as well as in culture-specific conventions. This study observes performances in such kinds of attitudinal prosody for USA English, produced by L1 and L2 speakers. The results show that the observed variance is linked to individual competence, to the linguistic context, and to the cultural background of the speakers. They also show that the code used to express a given speech act, code learned in the L1 language by L2 speakers of English, may be used in their L2 language. For some of these expressions, L2 speakers received higher scores than L1 speakers, suggesting that expressions conventionalized in a foreign language, are adequately fulfilling not-conventionalized expressions in the L1 culture.

Index Terms: prosody, attitude, cross-cultural, first and second language

1. Introduction

Prosodic performances help a speaker to express a position about her/his speech and also about the addressee [1].They help to actualize speech acts (e.g. to convey doubt or authority) in a non-verbal way, thus allowing a smoother interaction – it is easier to retract from a position that has not been verbalized. Such attitudinal prosody has been theorized in pragmatic terms as intermediate between emotional expressions and illocutions [2,3]. Differing from emotions, they are intentionally produced and controlled by the speaker; unlike e.g. wh-questions, attitudinal prosody may not have a one-to-one relationship with the meaning of the intended speech act. Attitudinal prosody is rather part of a contextualized communication process, and speakers may use many tools, including prosody, to reach their goals.,

Most studies of attitudinal prosody focus on performance in a given language, either proposing inventories of the existing clichés observed in a language [4,5,6,7], or studying the various means used to express a given speech act in a given language (e.g. politeness in [8,9]). Some studies are interested in the cross-cultural comparison of attitudinal prosody perception by listeners of various linguistic origins [10, 11]. To that aim, they take samples of social affects from a given language's inventory to present stimuli to listeners from various languages, studying the biases induced by cultural backgrounds in the reception of prosodic attitudes. In all these works, attitudes are obtained and labeled according to the cultural specificities of a given language; that is they are interested in e.g. the performance of Catalan politeness, of French irony, etc.

But, as advocated by Wierzbicka [12,13,14] for emotions, the label of an attitude in a given language (e.g. French *ironie*, USA English *irony* or Brazilian *ironia*) does not necessarily cover the exact same concept: these three ironies are not produced in similar social context, or with a similar communicative goal. It is interesting to study these conceptual differences, but we'll here concentrate on how prosodic performances may vary across languages in identical situations – trying to freeze such conceptual mismatches (i.e., what would have been the prosodic performances of Brazilian Portuguese speakers expressing the speech act corresponding to French "*ironie*"?) The aim of this ongoing research is precisely to record speakers from various linguistic and cultural origins, and to compare their prosodic performances in the same situations of communication. Speakers are not asked to produce a sentence with e.g. irony or authority, but rather to behave in situations conducive of these attitudes, as exemplified below:

- *Authority*: Speaker A is a custom agent; speaker B is a traveler. B is in front of A, requesting permission to enter the country; A needs to impose his authority; the scene is at a custom counter at the airport.
- *Irony*: A & B are friends, same age; A is going to Boston to see an important baseball game, and B, who is living in Boston calls A. Unfortunately, the weather in Boston is rainy and B says its wonderful; the scene is at an airport.

The interaction contexts as well as the communication goals of speakers are the same, whatever the language of the speakers. Speakers may be L1 or L2 speakers of the studied language. Several speakers from the same linguistic origin, and of both genders, are recorded to enable a comparison of the cross-speaker, cross-gender, and cross-cultural influence on the prosodic performances. Among the hypotheses of the work are the codes that have been proposed in the literature as universal of prosodic expressivity – namely the frequency code [15], and the effort and production codes [16]. These codes would predict for example the use of higher pitch for expressions where the speaker is in an inferior position, and of lower pitch in the case of dominant expressions. Another research aim is to observe the possible influence of the linguistic and cultural backgrounds on the performances of speakers, for expressions corresponding to complex social settings. All interaction contexts do not necessarily correspond to a prototypical prosodic attitude in each culture, but this is precisely part of what is under investigation here. Will speakers use their L1 prosodic typology speaking in L2? If a situation is (or nor) conventionalized in the speaker's L1, will it help (or restrain) the performances of her/his productions, in L1 and in L2?

This paper focuses on the results obtained for USA English for L1 speakers, compared to L2 speakers whose L1 language is either Japanese (L2JP) or French (L2FR). The aims of the study include observations of prosodic performances across languages and cultural backgrounds, in comparable communication situations, and examination of the perception of these prosodic changes by L1 listeners of these languages. We will here review the evaluation by L1 listeners of USA English of the performances of L1 and L2 speakers in expressing speech acts that correspond to sixteen interaction situations.

2. Corpus

Sixteen communication situations have been designed, where the recorded speakers interact with an experimenter (L1 speaker of the target language) in order to elicit the corresponding attitudes. These attitudes are performed on two target sentences ("A banana" and "Mary was dancing") that are produced in the end of a small dialogue picturing a situation where these sentences may be produced with the intended attitudes. A complete description of the recording setting may be found in [17]. The sixteen situations correspond roughly to expressions described by the following labels: admiration (ADMI), arrogance (ARRO), authority (AUTH), contempt (CONT), doubt (DOUB), irony (IRON), irritation (IRRI), neutral declarative sentence (DECL), neutral question (QUES), obviousness (OBVI), politeness (POLI), seduction (SEDU), sincerity (SINC), surprise (SURP), uncertainty (UNCE), "Walking On Eggs" (WOEG), often referred to as "walking on eggshells" or "walking on thin ice."

Labels "sincerity" and "walking on eggs" correspond to situations typical of the Japanese society, where the speaker is supposed to behave in a specific way because of the hierarchical relationship with the interlocutor and the intended speech act. For sincerity, the speaker asserts the sincerity of her/his utterance to a higher-level interlocutor. The WOEG situation is close to the Japanese concept of *kyoshuku*, defined by [18] as "*corresponding to a mixture of suffering ashamedness and embarrassment, which comes from the speaker's consciousness of the fact his/her utterance of request imposes a burden to the hearer*" (p. 34). Expressions labeled "irony" and "seduction" correspond to attitude types frequently used in the U.S.A., where two speakers have mutually positive, friendly attitudes. Irony may often be used to help lighten an otherwise negative situation. Other types of "negative" irony, intending to hurt the other person, are not addressed in this study. The attitude of "seduction" might best be defined here as where the speaker interacts with her/his interlocutor with the intention of being attractive, fascinating, inviting to the listener. This type of attitude may be akin to that frequently seen in Hollywood movies, but without necessarily being sexually provocative.

Among the speakers recorded for this study on USA English, eight (5 females and 3 males) are L1 speakers; six (3 females and 3 males) are L2 speakers with Japanese (Tokyo variety) as their L1; five (3 males and 2 females) are L2 speakers with French (standard French variety) as their L1. Most speakers are university students. To select L2 speakers – and to ensure a basic level in the target language-selected L2 speakers all spent one year studying in the USA. All speakers were audio-visually recorded performing the two sentences in the 16 situations. The video were hand segmented, resulting in 256+192+160=608 stimuli.

3. Performance evaluation

3.1. Experimental paradigm

These audio-visual performances were presented to listeners, whose L1 is USA English, who had to judge the performances on a raw 1 to 9 scale. Three different groups of listeners evaluated each of the three groups of speakers (L1, L2 Japanese, L2 French). Subjects were first presented with the 16 situations, explained showing the pictures used to elicit the 16 attitudes for "banana." The stimuli were then presented in isolation, by speaker (in order to give subjects a better idea of the expressive habit of this speaker) for both sentences and were randomized. Prior to the presentation of a stimulus, the name of the targeted attitude was displayed, then the stimuli was presented (only once), and listeners had 10 seconds to use the keyboard to give a performance score; then the next stimuli was presented. In cases of no answer, a 0 score is given to the stimulus (it represents less than 0.1% of the number of answers) – these 0 scores were latter removed from the data. A test session lasts typically 30 to 45 minutes. To move to the next speaker, subjects were instructed to press the "enter key."

3.2. Subjects

17 subjects (7 females, mean age 25) evaluated the L1 speakers' performance; 16 subjects (6 females, mean age 21) evaluated the L2 speakers with Japanese L1; 35 subjects (26 females, mean age 24) evaluated the L2 speakers with French L1. The evaluations for all three groups were conducted in Midwest U.S.A., (the prevailing dialect is Midwestern English). The first two evaluations were conducted in Ohio and South Dakota, the third group, in New Mexico.

3.3. Performance judgments analysis

The performance judgments received by each stimulus were analyzed using a mixed-effects model [19] based on the lme4 library [20] of the R software [21]. The performance scores were normalized using a z-score for each subject to avoid individual differences in the use of the answer scale. These performance measures were fitted as a function of the subjects and the speakers – as random effects; the latter being nested into the group of linguistic origin factor, and the linguistic group factor interacts with the targeted attitude and the sentence (the last three being fixed factors).

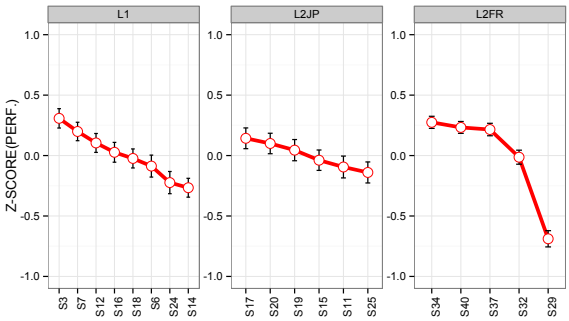


Figure 1: *Mean performances (expressed in z-scores) obtained by individual speakers, nested in the linguistic groups (L1 speakers, L2 Japanese, L2 French); the error bars represent the confidence intervals at 5%.*

Note that, because each of the three groups of listeners evaluate one of the three groups of speakers, the raw performances between speakers' group cannot be directly compared – thus the use of z-scores. But z-scores, since they are standardized, also prevent us from comparing the average performances of two groups of speaker. The present analysis focuses on the differences inside each group of speakers, and compares their relative performances in the 16 situations.

An ANOVA based on the mixed-effects model presented above shows that the linguistic group and the subject factors

did not have a significant main effect (this result is obvious, after the preceding comment on z-scores), but the linguistic group shows a significant interaction with the attitude ($\chi^2_{(30)}$=156.2, p<2.2e-16), with the sentence ($\chi^2_{(2)}$=22.1, p=1.6e-5), and the triple interaction was also significant ($\chi^2_{(30)}$=80.9, p=1.4e-6). The main effects of targeted attitudes and support sentences are also significant (respectively: $\chi^2_{(15)}$=1105.5, p<2.2e-16; $\chi^2_{(1)}$=7.3, p=0.007). The speaker factor also has a significant effect on the performances (effect tested using the method presented in [19:242]; comparing this model with another model, it is identical except it did not have the speaker factor – $\chi^2_{(1)}$=940.1, p<2.2e-16).

3.3.1. *Speakers performances*

The mean performances reached by each speaker, in each linguistic group, are displayed in figure 1. There are important variations across speakers. These variations are mostly linear in the L1 and L2 Japanese groups; one speaker (S29) departs from the others in the L2 French group. Performance variations are more restricted (the group of speakers is more homogeneous) in the L2 Japanese groups, compared to the two other groups. The following analyses consider performances averaged across all speakers of each of the three groups.

3.3.2. *Effect of sentences*

Targeted expressions were produced on two sentences ("A banana" and "Mary was dancing"), chosen for their syntactic simplicity and absence of affective meaning. These individual sentences did have a small – yet significant – effect (cf. Figure 2) on the performances. This effect is different in the case of the L1 group of speakers, as compared to the two L2 groups: L1 speakers received higher scores for their performances on the "Mary" sentence than on the "banana" one, while the reverse is observed for both groups of L2 speakers.

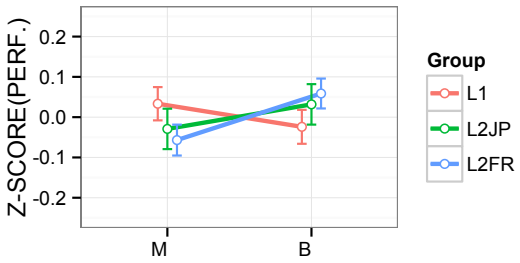


Figure 2: *Mean performances (in z-scores) for both sentences ("Banana" and "Mary was dancing"), for each linguistic group (L1, L2 Japanese, L2 French); error bars represent the confidence intervals at 5%.*

3.3.3. *Effect of attitudes*

Expressing each of the 16 individual attitudes is not achieved by speakers with the same accuracy. The significant main effect of attitude on the performance is depicted in figure 3, and shows important differences (of about one standard deviation) between the best-elicited expression (the situations of surprise) and the worst one (the situation of irony).

The tendency observed for the speakers of all linguistic groups for the two best (surprise and doubt) and the two worst (irony and seduction) expressions holds for the L1 speakers, while it seems culture-specific competences or constraints induce varying patterns of performances for the two L2 groups of speakers.

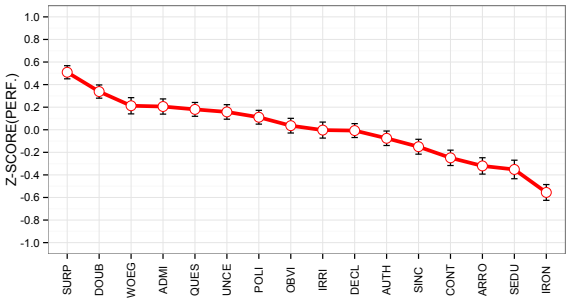


Figure 3: *Mean performances (in z-scores) for each attitude (see text for labels), all linguistic groups averaged; the error bars represent the confidence intervals at 5%. Attitudes are sorted in descending level of performances.*

These cultural-specific patterns, as well as some cross-cultural similarities, are displayed in figure 4, where attitudes are sorted in descending order of performances for the L1-speaker group. The expression of surprise received the highest performance scores whatever the linguistic origin of speakers; the expression of doubt is also ranked amongst the best performances for all speaker groups (if only the third best for the L2 Japanese speakers).

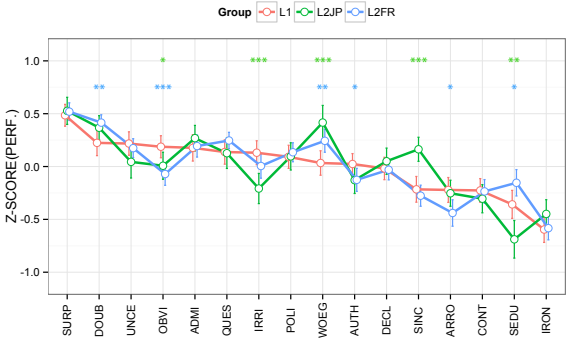


Figure 4: *Mean performances (in z-scores) for each attitude (see text for labels), and for each linguistic group (L1, L2 Japanese, L2 French); the error bars represent the confidence intervals at 5%. Attitudes are sorted in descending level of performances for the L1 speaker group. Stars above attitudes indicate significant differences between the L1-speaker group and L2 Japanese group (top line of stars) or L2 French group (bottom line of stars).*

More differences are observed amongst the performances of the lowest ranked attitudes. The expression of irony seems to be difficult to elicit for speakers of all linguistic origins, but it is not necessarily the most problematic one: L2 Japanese speakers received their lowest performances scores for their behavior in the situation of seduction. The expression of seduction received also the second lowest scores for L1 speakers, while it received only the fifth lowest score for L2 French speakers (close to the mean performance level of this group). Other important differences are observed between the L1 speaker group and the two L2 speaker groups.

In order to focus on the most important differences between language groups, T-tests were run to test significant

differences between performances for a given attitude of the L1-speaker group and each L2-speaker groups. Significant differences are indicated on figure 4 by stars above the significant differences observed for a given attitude (*: $p<0.05$; **: $p<0.01$; ***: $p<0.001$).

The L2 Japanese speakers received significantly different scores from the L1 group for five interaction situations. For irritation, obviousness, and seduction, Japanese speakers received significantly lower scores than L1 speakers; conversely, Japanese speakers outperformed L1 speakers in the cases of sincerity and WOEG expressions.

The L2 French speakers received significantly different scores from the L1 group for six interaction situations. For arrogance, authority, and obviousness, they received lower scores than the L1 speakers, while for the expressions of doubt, seduction, and WOEG, L2 French speakers outperformed the L1 speakers.

4. Discussion & conclusions

Speakers, depending on their personalities among other things, performed differently. The interaction situations were prepared to ease the expression of the targeted attitude in a laboratory environment – but still, in the laboratory as in real life, individuals behave differently, and express themselves with more or less change in voice and face; thus the observed variation in the performances. The focus of this paper is not on individual differences. In order to have a better idea of performance differences at the linguistic and cultural levels, several speakers have been recorded in each language group. It would of course be of a great interest to enter the details of these differences and study the influence of individual characters on expressivity (cf. [22, 23, 24] on this topic). Along this line, it is interesting to note that some speakers received low mean scores for their performances, but may have received the best score for their elicitation of a given situation. It is typically the case for L1 speaker S24, whose mean performance scores are rather low, but who was rated amongst the best for expressing surprise.

Also, the variations induced in performances by the sentences may be related to the type of elicitation contexts used for each of the sentences. L1 speakers tended to perform better for the "Mary was dancing" sentence, which is introduced by a written dialogue, while the L2 speakers received higher scores in the "banana" situations, which are presented using iconic presentations based on images. The language level, and the linguistic complexity may be an explanation for the observed differences. Another explanation of this observation (not necessarily exclusive of the first one) could be linked with the greater linguistic complexity of the "Mary was dancing" sentence, compared to "Banana".

Most of the explained variance is linked with the varying performances of the 16 attitudes, and with the differences between linguistic groups in the performance of individual attitudes. A first remark about these differences is that they concern only a minority of the communication contexts: in most cases, L1 and L2 speakers do perform comparably and this result is in itself interesting. This supports literature reporting few differences found cross-culturally e.g. in [25]. The differences observed between L1 and L2 speakers can be explained by different factors.

- A first kind of differences is the one where L1 speakers outperform one (and only one) group of L2 speakers in performing a social attitude. It is the case for irritation with L2 Japanese speakers, and for arrogance and authority with L2 French speakers. These situations are conventionalized in the considered cultures. Whether they are based on different codes may be answered by comparing the productions of both L1 and L2 groups of speaker.
- In the situation leading to expressions of obviousness (a propositional attitude, see e.g., [26]), both groups of L2 speakers are outperformed by L1 speakers. Although the expression of obviousness may be conventionalized in each of these cultures, propositional attitudes address the linguistic content of the utterances and thus are subjected to more variations in their pitch contours [26]. One can suppose that the L2 speakers did not catch the typical prosodic patterns used by L1 speakers.
- The situation demanding speaker to have a seductive behavior shows an interesting partition, with L2 Japanese speakers receiving lower score than the L1 speakers, while L2 French speakers received higher scores. In this case, the social interaction is not conventionalized in the Japanese society, while it is – if under different conventions in the French culture. The lack of – or presence of – a conventionalization seems to be a critical point in the measured performance of speakers.
- In the case of the situation leading to the expression of sincerity and WOEG, the L2 Japanese speakers outperformed the L1 speakers. These expressions correspond to common situations in the Japanese culture, while they are not conventionalized in the culture of the USA, where the society is less based on hierarchical relations. This lack of conventionalization in the L1 culture may also, as in the previous case, be a key point in the observed differences in performance.
- The L2 French outperformed L1 speakers in the WOEG situation to a lesser degree, although it is not a conventional situation in the French culture. A possible explanation may lie in the more hierarchical nature of social relationships in France, as compared to the USA, that may play role for an unconventional expression, but not for conventional ones such as arrogance or authority.

A question remains. In the cases where L2 speakers outperformed L1 speakers, (perhaps because the L2 culture conventionalization "trained" the L2 speaker how to behave in these situations), why were these performances evaluated so highly by L1 speakers of USA English who did not know the conventions. Why did they give high ratings to L2 French speakers' seduction and L2 Japanese speakers' WOEG and sincerity? Could it be possible that the judges do have stereotypic representations of peoples from these cultures, and thus relate Japanese behaviors more easily to politeness expressions, and French to seductive behavior? It could be possible to test for such stereotypes by running performance judgments of the same expressions with native listeners of the two L2 cultures less biased by such stereotypes.

5. Acknowledgements

This work was supported by the following grants: ANR PADE, JSPS Grants A #25240026 and A #23320087, and PEPS IDEX MAVOIX. The authors warmly thanks M. Kondo & S. Detey from Waseda University for their help; they also thank all the speakers and listeners for their participation.

6. References

[1] Moraes, J. A., "The pitch accents in Brazilian Portuguese: Analysis by synthesis", in Proceedings of Speech Prosody 2008, Campinas, 389–397, 2008.

[2] Wichmann, A., "The attitudinal effects of prosody, and how they relate to emotion", in Proceedings of the ISCA Workshop on Speech and Emotion, Newcastle, 143–148, 2000.

[3] Wichmann, A. "Attitudinal intonation and the inferential process", In Speech Prosody 2002, 11-16, 2002.

[4] Martins-Baltar, M., "De l'énoncé à l'énonciation: une approche des fonctions intonatives", Paris: Didier, 1977.

[5] Fujisaki, H. & Hirose, K., "Analysis and perception of intonation expressing paralinguistic information in spoken Japanese", Proceedings of the ESCA Workshop on Prosody, 254-257, Lund, Sweden, 1993.

[6] Morlec, Y., Bailly, G. & Aubergé, V., "Generating prosodic attitudes in French: Data, model and evaluation", Speech Communication, 33(4):357–371, 2001.

[7] Gu, W., Zhang, T. & Fujisaki, H., "Prosodic Analysis and Perception of Mandarin Utterances Conveying Attitudes", Proceedings of Interspeech, Firenze, Italy, 1069-1072, 2011.

[8] Wichmann, A., "The intonation of Please-requests: a corpus-based study", Journal of Pragmatics, 36: 1521–1549, 2004.

[9] Nadeu, M. & Prieto, P., "Pitch range, gestural information, and perceived politeness in Catalan", Journal of Pragmatics, 43(3): 841-854, 2011.

[10] Scherer, K. R., Brosch, T., "Culture-Specific Appraisal Biases Contribute to Emotion Dispositions", European Journal of Personality, 23: 265–288, 2009.

[11] Shochi, T., Rilliard, A., Aubergé, V. & Erickson, D., "Intercultural perception of English, French and Japanese social affective prosody", in S. Hancil [Ed.], The role of prosody in affective speech, Linguistic Insights 97, Bern: Peter Lang, AG, Bern, 31-59, 2009.

[12] Wierzbicka, A., "A semantic metalanguage for a cross-cultural comparison of speech acts and speech genres", Language in Society 14(4): 491-513, 1985.

[13] Wierzbicka, A., "Defining Emotion Concepts", Cognitive Science 16:539-581,1992.

[14] Wierzbicka, A., "Empirical Universals of Language as a Basis for the Study of Other Human Universals and as a Tool for Exploring Cross-Cultural Differences", Ethos 33(2):256–291, 2005.

[15] Ohala, J.J., "An ethological perspective on common cross-language utilization of F0 of voice", Phonetica, 41:1-16, 1984.

[16] Gussenhoven, C., "The Phonology of Tone and Intonation", Cambridge: Cambridge University Press, 2004.

[17] Rilliard, A., Erickson, D., Shochi, T., Moraes, J., "A. Social face to face communication – American English attitudinal prosody", in Proceedings of Interspeech, Lyon, 1648-1652, 2013.

[18] Sadanobu, T., "A natural history of Japanese pressed voice", Journal of the Phonetic Society of Japan 8(1): 29-44, 2004.

[19] Baayen, R. H., Davidson, D. J., Bates, D. M., "Mixed-effects modeling with crossed random effects for subjects and items", Journal of Memory and Language, 59: 390–412, 2008.

[20] Bates, D., Maechler, M., Bolker, B. & Walker, S., "lme4: Linear mixed-effects models using Eigen and S4". R package version 1.0-5. 2013. http://CRAN.R-project.org/package=lme4

[21] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[22] Scherer, K. R., "Personality markers in speech", in K.R. Scherer & H. Giles [Eds.]. Social markers in speech. Cambridge: Cambridge University Press, 147-209, 1979.

[23] Sadanobu, T., "Nihon shakai, Nozoki Chara-kuri - Kaotsuki, Karadatsuki, kotobatsuki-", Sanseido Publisher, 2011.

[24] Sadanobu, T. & Luo, M., "Bumpo, Para-gengojouhou, Character ni motozuku Nihongo meishisei bunsetsu no tougoteki kijyutsu", Journal CAJLE (ISSN 1481-5168), 12: 77-95, 2011.

[25] Rilliard, A.; Erickson, D.; Moraes, J. A.; Shochi, T., "Cross-Cultural Perception of some Japanese Expressions of Politeness and Impoliteness", in F. Baider, G. Cislariu [Eds.] Linguistic approaches to emotions in context. Amsterdam: John Benjamins, 251-276, 2014.

[26] Moraes, J. A., Rilliard, A., "Illocution, Attitudes and Prosody", In T. Raso et al. [Eds.], Spoken Corpora and Linguistic Studies, Amsterdam: John Benjamins, to appear.

An Automatic Hierarchical Multiple Level Phrase Segmentation Approach for Spontaneous Speech

András Beke[1], *György Szaszák*[2], *Viola Váradi*[3]

[1]Research Institute of Linguistics, Hungarian Academy of Sciences, Budapest, Hungary
[2]Idiap Research Institute, Martigny, Switzerland
[3]Dept. of Phonetics, Eötvös Loránd University, Budapest, Hungary
beke.andras@nytud.mta.hu, gyorgy.szaszak@idiap.ch

Abstract

The present paper investigates automatic prosodic phrasing of spontaneous speech: a two-step segmentation technique is presented, based on unsupervised learning. In the first step, the Intonational Phrases (IP) are detected automatically based on speech energy, spectral centroid and a double-thresholding technique. In the second step, Phonological Phrases (PP) are identified within the IPs. As acoustic features, F0, overall energy and vowel duration are investigated. An adaptive thresholding method is used based on Kullback-Leibler divergence computed in an autocorrelative manner for the feature streams. For Hungarian spontaneous speech, a phrasing accuracy of over 80% can be reached when comparing to a hand-labelled reference phrasing. It is found that in Hungarian sponatenous speech, F0 and energy play an essential role in IP level phrasing, whereas PP level phrasing is most effective using F0 related features alone. Vowel durations are shown not to contribute to prosodic phrasing in Hungarian. Although the evaluation targets the Hungarian language, the applied method is universal and can be easily adapted for other languages.

Index Terms: phrasing, spontaneous speech, hierarchical

1. Introduction

Prosodic phrasing and/or prosodic boundary detection is an important research topic and several phrasing or boundary detection approaches have been developed and analysed for read and slightly spontaneous speech (such as semi-formal speech used in information retrieval systems) [1], [2]. When using a supervised approach, machine learning can be applied on labelled data, which will end in producing a classifier or detector capable of predicting prosodic boundaries based on the associated acoustic-prosodic features. This approach also implies that the entities to be classified or detected (boundary or break and their types) are a priori known and annotated in the training corpus. More recently, unsupervised modelling of prosody [3] or adaptation of seed prosody models in an unsupervised manner has also received attention [4]. Such approaches are of primary interest when dealing with spontaneous speech, as due to its extreme variabilty (disfluencies, atypical or non-canonical realizations in terms of acoustic correlates), prosodic entities worth modelling can be problematic to be identified or clustered even by human experts.

This paper focuses on exploring the prosodic structure of spontaneous speech, work is done on a Hungarian spontaneous speech database. Earlier efforts for prosodic event detection and automatic phrasing in Hungarian read speech showed that based partly on the fixed stress of Hungarian, robust stress detection was possible [5] and a phonological phrase alignment approach was proposed [6], where modelling of F0 and energy contours of phonological phrases was used in a supervised machine learning approach to perform automatic phonological phrase alignment and based on this, a partial, but powerful recovery of the prosodic and, to a lesser extent, of the syntactic structure in read speech. However, not surprisingly, when trying to adapt this automatic approach for spontaneous speech, recall rates fall by approx. 20-30% from around 80%. An effort to try to identify and cluster characteristic prosodic entities or phrase types in Hungarian spontaneous speech by using an unsupervised approach lead only to partial success [7].

However, supposing a hierarchical structure of prosody as described by Selkirk [8], the automatic phrasing implemented for read speech in [6] was able to yield a reliable (accuracies close to 80%) phrasing down to the phonological phrase level, and also to separate intonational phrase level from the underlying phonological phrase level. This approach required clustering of a number of phonological phrase prototypes, which were then modelled by HMM/GMM based on acoustic-prosodic features. As already mentioned, clustering of such characteristic prototypes in sponatenous speech failed. However, our hypothesis is that upon the intonational phrase level, also the phonological phrase level should be identifiable based on prosody in spontaneous speech as well. Moreover, cues for the perception of phonological phrasing are supposed to be simple enough, in order to allow the listener to concentrate also on other modalities (attitudes, emotions, also dialogue management functions) of the complex and rich information transmitted during a spontaneous conversation. Therefore, in this paper an attempt is made to detect intonational and phonological phrases in spontaneous speech with a two-step method, able to separate these two levels according to the prosodic hierarchy.

This paper is organized as follows: First, the spontaneous database is presented, than the intonational and phonological phrase segmentation approaches are described. Phonological phrase detection is implemented and evaluated in the subsequent section, and finally conclusions are drawn.

2. Data and methods

The BEA (BEszélt nyelvi Adatbázis: spoken language database [9]) spontaneous speech database was used in this research. BEA is a multi-purpose database of Hungarian spontaneous speech. 8 spontaneous narratives were selected (4 male and 4 female) from the database. The subcorpus was manually annotated by two different phoneticians, although the annotation will be exclusively used as reference for evaluation. The

annotation contained three levels: Intonational Phrases (IP), Phonological Phrases (PP) and also involved a word level transcription. The intonational phrase can be thought of being a part of speech forming a unity in terms of stress and intonation contour, and is found often between two pauses. The IPs can be further divided into phonological phrases based on intonation and stress pattern. A PP is a unity characterized by its own stress and intonation contour, but this latter can be unterminated (continued in next PP). The corpus contained 398 IPs and 751 PPs in total from the 8 speakers.

2.1. Intonational Phrase segmentation

In the spontaneous speech of the BEA database, turns can usually be further segmented into separate utterance units. However, there is no consensus on how to define an utterance unit [10]. The manner in which speakers segment their speech into intonational phrases undoubtedly plays a major role in its definition. Intonational phrase endings can be signalled through variations in the pitch contour, segmental lengthening and pauses. Cruttenden [11] in his theory uses external criteria for identifying intonation groups defined by *potential* boundaries. One of them is a potential pause following an intonation group, however, pauses are not obligatory boundary markers and may occur within a group. According to Shriberg and his colleagues [12] important cues to boundaries between semantic units, such as sentences or topics, are breaks in prosodic continuity, including pauses. In their system for sentence segmentation, the pause model used by the recognizer was trained as an individual phone. The result showed that this pause model was among the ones with the highest influence in sentence segmentation task. Pauses are without a doubt the most expressive instruments for marking of strong boundaries [13]. For Hungarian language, Gósy's research results showed [14] that pauses were among the most important features for utterance segmentation in human perception. In our database, in most cases the IP boundaries are bounded to pauses. A number of filled pauses and unfinished words were perceived as separate IPs as well by the annotators. These findings suggest using two features: a feature which plays an important role in speech detection such as speech energy or speech centroid; and a second feature which is sensitive to speech intonation such as fundamental frequency. In IP segmentation, the task is not only to find silent regions in continuous speech, but to detect the strong intonation changes in the acoustic signal. There are many solutions to detect silence and speech in the audio signal. In this research we choose a very simple and fast algorithm to segment pauses, created by Giannakopoulos [15] and implemented in MATLAB. In this system, a fundamental frequency estimation algorithm is also implemented.

This algorithm first extracts three features: signal energy, spectral centroid and fundamental frequency. These features are extracted for every frame using a 50 ms long window. The signal energy is higher in case of a speech segment than a silent segment. The spectral centroid is a spectral position and this expresses which frequency region contains the most part of intensity in the spectrum. A higher value of the spectral centroid usually corresponds to a speech segment in the audio signal. F0 can refer to the speech signal, and provides a good representation of intonation movements. On both features, 5 point median filtering is applied to smooth the signal. After the feature extraction, a threshold is calculated to each feature stream. In order to determine the threshold, first the smoothed histogram of the feature is used. The following step is to find local maxima of the histogram. In this step we deviate from the original method. To find the two most frequent values in the histogram we use k-means unsupervised learning algorithm. K-means clustering [16] is one of the simplest and oldest unsupervised learning algorithms. Given a set of data (consisting of n different, d-dimensional observations) and the desired number of clusters (k), this algorithm clusters iteratively the data around the so called centroids. For bootstrapping, k data can be randomly chosen as centroids, then each observation is clustered to the nearest centroid. The nearest centroid is computed using some distance measure, such as the Euclidean distance or sum of squares, for example. Centroids are iteratively updated to the mean of the belonging observations, until a specified level of convergence is reached. The main drawback of the algorithm is that the number of clusters (k) has to be determined prior to the clustering itself. In this process we used two clusters: pause and speech.

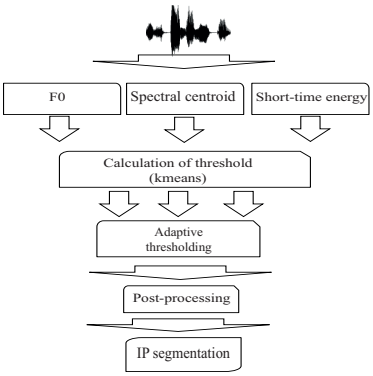


Figure 1: Block diagram of IPs detection

The clustering problem can be formulated as minimizing:

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} D(x_j, \mu_i) \tag{1}$$

for clusters $C_1, C_2, ...C_k$, given n observations ($k < n$), using a $D(.)$ distance function to evaluate the distance between centroid means (μ_i) and observations (x_j). Since this problem is NP hard, usually heuristic approximation is used to solve the problem. If the two cluster centers are available, we calculate the threshold using the following equation:

$$T = \frac{W * M_1 + M_2}{W + 1} \tag{2}$$

where W is a user-defined parameter (set to $W = 0.5$ in our case). Once the threshold is calculated, we apply it to the feature streams. The last step is the post-processing step. In this step the overlapping segments are merged using a large window (usually of a length of about $250ms$, see Figure 1).

2.2. Phonological phrase segmentation

Segmentation for PPs is a harder task than detecting IPs in the continuous audio signal. As PPs are embedded into IPs, the output of the segmentation for IPs constitutes the input of the segmentation for PPs (Figure 2), leading to a two-level, hierarchical approach.

As PPs can be described as phrases with own (proper) intonational and stress patterns, we used fundamental frequency

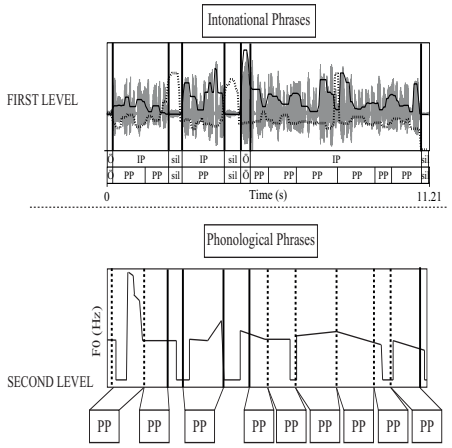


Figure 2: Hierarchical PPs detection

(F0), mid-term speech energy and vowel's duration for their detection. PP boundary detection is carried out as described in the next subsection.

2.2.1. Feature extraction

We use the three basic acoustic correlates of prosody as features: fundamental frequency, energy and duration (tempo).

Fundamental frequency (F_0) is extracted by ESPS method using a 25 ms long window, by a frame rate of 10 ms. The obtained F_0 contour is first filtered with an anti-octave jump tool. This is followed by a smoothing with a 5 point mean filter. F_0 is linearly interpolated in log domain. The interpolation is omitted for voiceless sections longer than 150 ms and also for F_0-rises higher than 110% after an unvoiced part.

Energy is extracted with a 150 ms window by 10 ms frame rate and then a further 5 point mean filtering is applied. As duration features, vowel lengths are used. In order to make the feature extraction automatic, an HMM-GMM based broad phoneme classifier is used. Broad phoneme classes cover vowels, nasal and approximant consonants, plosives, affricates and fricatives. The phoneme classifier uses standard MFCC features as input (with first and second order deltas) and produces a phoneme class level alignment at its output. Resulting vowel length is normalized per speaker, made continuous by a 10 ms frame rate and smoothed in order to obtain a vowel duration contour, called tempo feature. The reason for using a broad phoneme classifier instead of an ASR is twofold: we would like to keep the system generalizable to untranscribed, highly spontaneous speech with no proper language model coverage and hence use a phoneme-class loop grammar, and the phoneme classifier is more accurate in the required pure phoneme-class loop recognition task.

Although the phoneme classifier itself works by some incertaincy, confusions between classes are not crucial if they are systematic, whereby they can even reflect important prosodic information (such as in case of utterance final positions, where vowels are often confused due to low energy and irregularity). For each feature we calculate the first and second order deltas as differential and acceleration coefficients. The dimensionality of the feature vector is hence 9. After the feature extraction we normalise features to 1.

2.2.2. Segmentation using symmetrical Kullback-Leibler distance

The Kullback-Leibler (KL) distance is one of the most commonly used algorithms to measure the dissimilarity between two distributions [17]. The KL distance has been used for various tasks like speaker diarisation, speaker recognition, speech recognition, voice activity detection, etc. The KL distance can be used in speech segmentation and music segmentation as well. In this study, we apply the KL distance to detect phonological phrase boundaries in spontaneous speech. Matthew et al. [18] showed that the symmetric Kullback–Leibler distance is an effective distance metric to facilitate the detection of long-term statistical differences in speech signals. The mathematical background is as follows: let us assume X and Y are two random distributions, and KL is the dissimilarity between these two distributions. The distance $KL(X;Y)$ between X and Y can be calculated as:

$$KL(X;Y) = E_X(log(\frac{P_X}{P_Y}), \tag{3}$$

where E_X stands for the expected value of the probability density function of X. If distributions are modelled by Gaussians, the above equation becomes:

$$\begin{aligned} KL(X;Y) = \frac{1}{2}tr[(\Sigma_X - \Sigma_Y)(\Sigma_Y^- 1 - \Sigma_X - 1)] + \\ + \frac{1}{2}tr[(\Sigma_Y^- 1 - \Sigma_X^- 1)(\mu_X - \mu_Y)(\mu_X - \mu_Y)^T], \end{aligned} \tag{4}$$

where Σ refers to covariance matrices and μ to mean vectors of the respective distributions.

As this function is asymmetric we can symmetrise it with the following method:

$$KL2(X;Y) = KL(X;Y) + KL(Y;X) \tag{5}$$

As said before, if both distributions are considered to be Gaussian, a closed form solution exists for the $KL2$ symmetric KL distance.

In this work, the $KL2$ distance was calculated between two consecutive parts of the signal, corresponding to 4 frames (40 ms) length each. The window step was 1 frame (10 ms). The following task is to find the peaks in $KL2$ value curve. In $KL2$ value a high distance value indicates a possible acoustic change, whereas a low value indicates that the two compared regions of the signal are acoustically similar. From the point of view of the peak detection, it is very important to choose the right window length. For this reason, we used various window sizes from 25 frame (250 ms) up to 70 frame (700 ms) on the $KL2$ curve.

The second problem is the threshold estimation. To detect the changing point, we use two adapted thresholds (thr_A and thr_B). The first is computed as the mean of a window around the given point, multiplied by a constant:

$$thr_A = \alpha \frac{1}{2N_1} \sum(F). \tag{6}$$

where F is the feature vector, N_1 is the length of windows, and α is a constant.

However, in order to be detected as PP boundary, the given value must also be greater than thr_B, which is calculated by:

$$thr_B = \sigma_F + \beta \frac{1}{2N_1} \sum(F) \tag{7}$$

where σ_F is the standard deviation over the windowed area, β is the size of the window. The first threshold ensures that the

given value is greater than the surrounding area, calculated over a small window. The second threshold is calculated over a larger window, and ensures that the change takes into account the general trend of the data changes. The window sizes are currently set to 3 and 4 seconds respectively. Use of these thresholds enables us to reduce false positive rate, and to return only the highest value at any possible PP boundary.

3. Evaluation

The techniques described in this section allow us to measure the performance of any segmentation algorithm. For evaluation method we use Brandts GLR method [19]. This method proposes three common measures which show the performance of the automatic segmentation. The first is the insertion (Ins), which means that there an extra boundary (event) in automatic segmentation to the reference segmentation. Omission (Oms) value means that there are left boundaries (missed events) in the automatic segmentation compared to the reference segmentation. The accuracy (Acc) is calculated using the number of correctly matched boundaries ($Corr$) – if the distance between the automatic label and manual label is within a pre-set tolerance –, the insertion value and the omission value:

$$Acc = \frac{Corr - (Ins + Oms)}{All} \quad (8)$$

The accuracy can measure the performance of the segmentation algorithm. As reference, the PP hand labelling is used. The result depends on the tolerance value, therefore we tried various tolerance values between 25 ms and 100 ms.

4. Results

First, the automatic IP boundary segmentation algorithm is evaluated. Based on speech energy, speech centroid and F0, accuracy of IP boundary detection was 83.1% in spontaneous speech. Leaving one of the features out considerably lowered performance. Most of the errors are caused by the IPs starting or ending by filled pauses. The second aim was to test the automatic PP boundary segmentation algorithm. In our research we tried several features an their combination to detect the PP boundary as well. We compared results obtained from five combinations of the three features. In our first experiment, we focused on the impact of the window size (used for similarity measure calculation in KL2) on PP boundary detection. Window length for the KL2 similarity measure ranges between 100 ms and 400 ms. The result shows the accuracy of the segmentation depending on window length for KL2 and the type of features as well. The best result is obtained by using F0 alone and a 400 ms long window for KL2 (Table 1).

Table 1: *Accuracy of PP segmentation depending on window length.*

Windows length (ms)	100	200	300	400
F0	68.71	75.83	76.33	**80.18**
Tempo	55.33	56.67	56.91	57.38
F0+energy	68.45	74.75	74.25	79.04
F0+tempo	68.79	73.15	72.33	78.22
F0+energy+tempo	68.86	72.60	71.84	77.05

We especially focused on speech tempo in PP boundary detection, as it was found unhelpful for read Hungarian [5]. When we use only the speech tempo for IP boundary detection the accuracy is very poor. However, the accuracy improves when the speech tempo is combined with F0, but it is still lower than using only F0. The accuracy also decreases if speech tempo is combined with F0 and energy. The result shows that the accuracy increases if speech tempo is discarded. Speech tempo does not correlate well with F0 (R=-0.06) or energy (R=-0.04). Based on these results, its contribution to the perception of PPs is likely to be negligible in Hungarian.

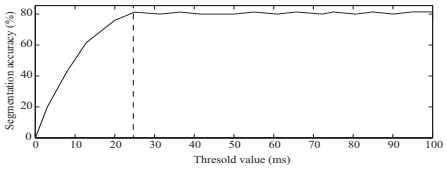


Figure 3: Accuracy of segmentation depending on threshold value

Next, the performance of the segmentation (accuracy) is shown, depending on the tolerance value. We tested our segmentation algorithm using various tolerance values (Figure 3). The result showed that if the tolerance is 25 ms the segmentation accuracy is 80.2%. By further augmenting the tolerance, there is no additional gain, accuracy saturates. This means that the segmentation algorithm is quite precise in time.

These results also suggest that F0 is the basic acoustic cue in PP boundary perception, whereas energy is also important in the perception of upper level unit boundaries is spontaneous Hungarian. The relatively powerful detectability of such boundaries in spontaneous speech strengthens the hypothesis that this kind of segmentation plays a key role in human perception as well.

5. Conclusions

The aim of this research was to segment spontaneous speech based on an unsupervised learning technique and a sophisticated peak detection approach. Phonological phrase segmentation was implemented in two hierarchical steps: first intonational phrase segmentation was performed using k-means clustering. Thereafter, intonational phrases were further segmented for phonological phrases, based on prosodic event detection exploiting symmetric Kullback–Leibler distance in an autocorrelation like approach. KL-distance features were themselves used as derived features, and peak detection was carried out on these. In our research we tried several features an their combination to detect the PP boundary as well. The results showed that fundamental frequency can be clearly associated with the phonological phrase level, as the best PP segmentation result is yielded by using F0 features alone by allowing only 25 ms time deviation between the detected PP boundary and the reference one. In this case, the accuracy was 80.2%. Results showed that speech tempo had no identifiable role in PP boundary detection. Regarding energy-based features, they seem to be associated to the IP level and did not improve PP detection based on F0 alone. However, some lack of robustness is supposed for the energy cues despite using normalization, due to the high degree of channel variability in the corpora (for example the speech recorded with a single microphone in a two-party conversation). These results are comparable to results seen for read speech [5], especially by the low uncertainty of the PP boundary detection regarding accurate placement in time.

6. References

[1] N. M. Veilleux and M. Ostendorf, "Prosody/parse scoring and its application in atis," in *Proceedings of the workshop on Human Language Technology*, ser. HLT '93. Stroudsburg, PA, USA: Association for Computational Linguistics, 1993, pp. 335–340.

[2] F. Gallwitz, H. Niemann, E. Nöth, and W. Warnke, "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication*, vol. 36, pp. 81–95, 2002.

[3] C. Chiang, S. Chen, H. Yu, and Y. Wang, "Unsupervised joint prosody labeling and modeling for mandarin speech," *Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1164–1183, 2009.

[4] A. S. and N. S., "Automatic detection of disfluency boundaries in spontaneous speech of children using audiovisual information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 138–149, 2009.

[5] K. Vicsi and G. Szaszák, "Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features," *International Journal of Speech Technology*, vol. 8, no. 4, pp. 363–370, 2005.

[6] G. Szaszák and A. Beke, "Exploiting Prosody for Automatic Syntactic Phrase Boundary Detection in Speech," *Journal of Language Modeling*, vol. 1, pp. 143–172, 2012.

[7] A. Beke and G. Szaszák, "Unsupervised clustering of prosodic patterns in spontaneous speech," in *Lecture Notes in Computer Science*. Springer, 2012, pp. 648–655.

[8] E. Selkirk, "The syntax-phonology interface," in *International Encyclopaedia of the Social and Behavioural Sciences*, N. Smelser and P. Baltes, Eds. Oxford: Pergamon, 2001, pp. 15 407–15 412.

[9] M. Gósy, "BEA - A multifunctional Hungarian spoken language database," *PHONETICIAN 105-106*, 2012, pp. 50–61.

[10] P. A. H. Traum, David R., "Utterance units in spoken dialogue," in *Dialogue Processing in Spoken Language Systems — ECAI-96 Workshop, Lecture Notes in Artificial Intelligence*, pp. 125–140, 1997.

[11] A. Cruttenden, *Intonation*. Cambridge University Press, 1997.

[12] D. H.-T. G. T. E. Shriberg, A. Stolcke, "Prosody-based automatic segmentation of speech into sentences and topics," pp. 127–154, 2000.

[13] P. Hansson, *Prosodic phrasing in spontaneous Swedish*. *Travaux de l'institut de linguistique de lund 43*, Lund University, 2003.

[14] M. Gósy, "Vitrual sentencse in spontaneous speech," in *Speechresearch 2003*. 2003, Budapest, Hungary, pp. 19-43.

[15] T. Giannakopoulos, "Study and application of acoustic information for the detection of harmful content, and fusion with visual information," Ph.D. dissertation, Dpt of Informatics and Telecommunications,University of Athens, Greece, 2009.

[16] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1967, pp. 281–297.

[17] C. L. Boite Jean Marc, "Speaker tracking in broadcast audio material in the frame work of the THISL project," in *Proceedings of the ESCA ETRW workshop Accessing Information in Spoken Audio*, pp. 84–89, 1999.

[18] B. R.-R. S. Matthew Siegler, Uday Jain, "Automatic segmentation, classification, and clustering of broadcast news audio," in *Proceeding of DARPA Speech Recognition Workshop*,pp. 97–99, 1997.

[19] D. P. S. Jarifi and O. Rosec, "Brandts GLR method and refined HMM segmentation for tts synthesis application," in *Proceeding of European Signal Processing Conference, EUSIPCO2005*, pp. 23–33, 2005.

Effects of auditory, visual and gestural input on the perceptual learning of tones

Katelyn Eng, Beverly Hannah, Keith Leung, Yue Wang

Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada

kse3@sfu.ca, beverlyw@sfu.ca, kwl23@sfu.ca, yuew@sfu.ca

Abstract

Research has shown that audio-visual speech information facilitates second language (L2) speech learning, yet multiple input modalities including co-speech gestures show mixed results. While L2 learners may benefit from additional channels of input for processing challenging L2 sounds, multiple resources may also be inhibitory if learners experience excessive cognitive load. The present study examines the use of metaphoric hand gestures in training English perceivers to identify Mandarin tones. Native Mandarin speakers produced tonal stimuli with simultaneous hand gestures mimicking pitch contours in space. The English participants were trained to identify Mandarin tones in one of four modalities: audio-only (AO, speaker voice only), audio-visual (AV, speaker voice and face), audio-gesture (AG, speaker voice and hand gestures) and audio-visual-gesture (AVG). Results show significant improvements in tone identification from pre- to post-training tests across all four training groups, demonstrating that gestural as well as visual articulatory information may facilitate tone perception. However, further analyses with individual tones reveal some group differences. Most noticeably, the AVG group had a slower learning curve during training compared to the other trainee groups for Tone 4, the least accurately identified tone, indicating a negative effect of multiple input modalities on the perception of difficult L2 sounds. In contrast, for Tones 2 and 3, the AG group revealed slower learning effects compared to the AV group, presumably because of the similar gestural trajectories for these two tones, which made the gestural input less distinct. Overall, the results suggest a positive role of gestures in tone identification, one that may also be constrained by phonetic and cognitive demands.

Index Terms: auditory, visual and gestural speech perception, Mandarin tone, L2 speech learning

1. Introduction

Research on multimodal speech processing has indicated that integration of auditory and visual articulatory information can enhance native and non-native speech perception [1], [2]. Additionally, co-speech hand gestures have been shown to facilitate native speech perception [3]. However, research has been inconclusive to the amount of gain from gestures in L2 speech learning [4], [5], [6]. For example, while beat gestures can aid L2 learners in parsing words into syllables [7], they are not as effective in discriminating durational differences [6]. Indeed, when integrated effectively, simultaneous auditory, visual and gestural information may have a combinatory effect in aiding speech learning [7], [8]. Conversely, the addition of gestural input may also be inhibitory as learners may experience excessive cognitive load, especially when phonetic demands are high [5], [6]. This discrepancy in the role of gestures motivates the present research.

In this study, native speakers of Canadian English were trained to perceive Mandarin Chinese lexical tones with one of four input modalities: audio-only (AO), audio-visual (AV), audio-gestural (AG), and audio-visual-gestural (AVG). The gesture used here is the metaphorical gesture, which traces an imaginary tone as it changes in pitch along the dimension of time (duration) and height (pitch), as is commonly used in Chinese tone teaching environments. This type of gestures has been shown to help pitch learning in contexts such musical training [9]. Training follows previously established high variability perceptual training procedures which involve various phonetic and speaker voice contexts to expose trainees to a variety of exemplars of the non-native speech categories [10], [11]. Trainees' performance was assessed by a pre- and a post-training tone identification test along with three intersession tests during training. Comparisons of the training effects with AG vs. AO or AV groups would determine whether and to what extent gestural information is beneficial in tone perception. On the other hand, if the addition of gestural input caused information overload, we would expect AVG training to be less effective than AG or AV training.

2. Methods

2.1. Participants

Four native Mandarin-speaking instructors (2 male, 2 female) were recorded to provide the training stimuli. They were chosen because of their familiarity with training students and knowledge of the Mandarin tones. Two additional native Mandarin speakers (1 male, 1 female) produced the pre- and post-test stimuli.

The trainees were 57 native Canadian English young adults. They had no prior experience with any tonal languages and no extensive experience with music either (with fewer than five years of musical training [12]). They were randomly assigned to one of the training groups (AO, AV, AG, or AVG), with 16 in each group (8 male, 8 female), except for the AO group, which had 9 participants (2 male, 7 female).

2.2. Stimuli

The training word list contained 80 Mandarin monosyllabic real words (20 syllables x 4 tones, Tone 1: high-level pitch, Tone 2: mid-high-rising pitch, Tone 3: low-falling-rising pitch, Tone 4: high-falling pitch), which was derived from those used in [11] and [13]. Audio-visual recordings of these words were made twice for all speakers, with and then without hand gestures. The stimuli used for the pre- and post-tests were the 60 additional Mandarin monosyllabic real words (15 tone quadruplets) used in [11].

Training stimuli speakers were instructed to simultaneously speak a word and trace an acetate graph

representation [14] of the corresponding tone on the feedback screen of the digital camera with their right index finger so that the gestures would be standard across speakers. The video was then mirrored so that the tone contour was presented in the correct direction for the trainees. For the AG condition, the speaker face was blacked out so that the only area visible was the arm including the hand and finger during tone tracing. Six additional native Mandarin speakers then participated in a stimuli goodness evaluation task. All the stimuli used in training were correctly identified and rated as good tokens of the Mandarin syllables by the native Mandarin speakers. Figure 1 displays sample images of (a) tone contour tracing used as gestural input in training, and (b-e) the four training conditions (AO, AV, AG, and AVG).

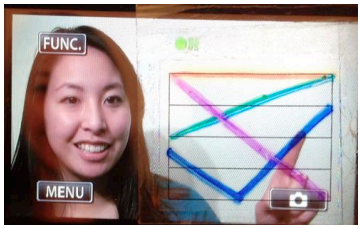


Figure 1-a: *Sample image of a speaker tracing the tone contour (pitch) on an acetate graph attached to the camera feedback screen.*

Figure 1-b: *Audio-only (AO) training presentation. No visual information is given in this condition.*

Figure 1-c: *Audio-visual (AV) training presentation with speaker face and voice.*

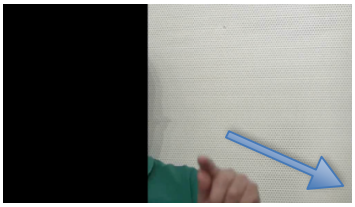


Figure 1-d: *Audio-gestural (AG) training presentation. The arrow indicates where the speaker will trace the contour of Tone 4.*

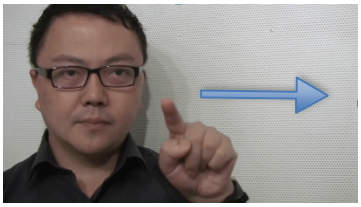


Figure 1-e: *Audio-visual-gestural (AVG) training presentation. The arrow indicates where the speaker will trace the contour of Tone 1.*

2.3. Procedures

Trainees were set up in a sound-treated booth in the Language and Brain Lab (Simon Fraser University, SFU) wearing AKG-brand circumaural headphones to hear the stimuli. They were first familiarized with the Mandarin tones by listening to a tone quadruplet and learned to associate each tone with its tonal label. They repeated this familiarization at the beginning of subsequent training sessions to review the task and hear an example before starting.

Prior to and after training, participants were tested with auditorily presented tone words described above. The identical pre- and post-tests employed a four-alternative forced choice task with no feedback provided. Identification was made using corresponding keys on the keyboard, with the labels "LEVEL" (for Tone 1), "RISING" (for Tone 2), "DIPPING" (for Tone 3) and "FALLING" (for Tone 4). Trainees were familiarized with these terms before the pre-test. After the pre-test, participant scores were calculated to determine if their percent-correct score was suitable for our inclusion criteria: If the participant scored between 25-80% correct, they were permitted to continue with the training. If the participants' scores fell outside that range, they were excluded and did not continue training. This was done to exclude those who had extreme scores due to hitting floor or ceiling.

Training took place during a two-week period with six sessions of 40 minutes each. Each session contained 40 words balanced across tones and syllables and produced by four speakers. Training stimuli were presented in AO, AV, AG or AVG, depending on the condition (as shown in Figure 1, b-e). The video was presented on a 12" (H) x 15" (W) display computer monitor, with the trainee's face roughly 18" away from the monitor screen.

Each trial started with presentation of a stimulus, followed by the trainee's task to identify the tone, and end with the feedback along with stimulus replay. Three intersession tests were administered after every second training session. These tests were used to track learning trajectory during training. They were presented in audio-only format, similar to the pre/posttest. The 80 intersession test stimuli (5 syllables × 4 tones × 4 speakers) employed selected stimuli used in training.

3. Results

3.1. Overall performance

The participants' percent correct identification scores were first analyzed using a three-way repeated measures ANOVA with Training Group (AO, AV, AG, AVG) as the between-subject factor, and Test (pre-test, intersession test 1 [Int1], intersession test 2 [Int2], intersession test 3 [Int3], post-test) and Tone (Tone1, Tone2, Tone3, Tone4) as the within-subject factors.

Significant main effects of Test [$F(4,212)=198.3$, $p<.001$] and Tone [$F(3,159)=31.6$, $p<.001$] were found. Bonferroni adjusted post hoc pairwise comparisons among tests reveal that across groups and tones, the mean pre-test score (48%) was significantly lower than all the intersession and post-test scores (Int1: 79%, Int2: 85%, Int3: 87%, post-test: 83%, $ps \leq .001$), and additionally, Int1 score was lower than Int2 and 3 scores ($ps \leq .001$), indicating significant

improvements with training. Post hoc tone comparisons show that, across tests and groups, perception of Tone 4 was significantly less accurate (68%) than Tones 1 (76%) and 2 (75%), which were in turn less accurate than Tone 3 (87%) ($ps \leq .038$), revealing Tone 4 as the most challenging tone.

The ANOVA also yielded significant interactions of Test x Tone [$F(12,636)=17.6$, $p<.001$] and Test x Tone x Group [$F(36,636)=1.6$, $p=.017$].

3.2. Individual tones and groups

Further analyses were performed based on the above interactions to identify possible Group differences as a function of Test and Tone. These involve sets of two-way Test x Group ANOVAs for each Tone, followed by further one-way ANOVAs for each Tone and Group with Test as a factor. Group comparisons of the five Tests for each Tone are displayed in Figure 2 (a-d).

First, only for Tone 4 (Figure 2-d) did the two-way ANOVAs show a significant interaction of Test and Group [$F(3,53)=4.0$, $p=.012$], along with a significant main effect of Test [$F(1,53)=337.4$, $p<.001$]. Bonferroni adjusted pairwise comparisons among the five tests for Tone 4 are consistent with the overall patterns. Across groups, the pre-test score (25%) was lower than all other test scores (Int1: 72%, Int2: 81%, Int3: 82%, posttest: 80%, $ps<.001$ for all). Furthermore, performance at Int1 was significantly poorer than that at Int2 and Int3 ($ps \leq .001$). Subsequent one-way ANOVAs with individual groups revealed that the aforementioned differences among intersession tests only occurred with the AVG group, with Int1 (67%) scoring marginally less well than in Int2 (81%, $p=.050$) and significantly lower than Int3 (86%, $p=.005$) and post-test (83%, $p=.007$), indicating a slower learning curve (during training) and a lack of generalization (to new stimuli at posttest) when training involved all three input modalities.

Consistently, for Tones 1-3, the two-way Test x Group ANOVAs also revealed significant main effects of Test. Though the ANOVAs yielded no significant interactions of Test and Group for these tones, the multiple levels of variables may obscure any possible difference in Test scores for each Tone and Group. Therefore, one-way ANOVAs were still run for each Group and Tone.

For Tone 1 (Figure 2-a), the two-way ANOVA yielded a significant effect of Test [$F(1,53)=125.5$, $p<.001$], with the scores being significantly lower at pre-test (45%) than at all the other tests which scored equally high (Int1: 81%, Int2: 85%, Int3: 85%; post-test: 82%, $ps<.001$ for all). Further one-way ANOVAs for each Group did not reveal any different patterns either. Thus all four training groups improved equally and retained the level of performance observed at Int1.

For Tone 2 (Figure 2-b), there was also a significant effect of Test [$F(1,53)=102.3$, $p<.001$], with pre-test (52%) being significantly lower than the other tests across groups (Int1: 75%, Int2: 83%, Int3: 86%, post-test: 76%, $ps<.001$ for all). Additionally, Int1 score was significantly lower than Int2 and Int3 scores ($p \leq .014$), both of which scored higher than post-test ($ps \leq .005$). Subsequent analyses with each trainee group revealed some group-specific patterns. First, the overall result of a higher-intersession-than-post-test performance was only exhibited in AO and AG groups, whose Int3 (AO=92%, AG=88%) score was higher than the post-test (AO=72%, AG=77%, $ps \leq .008$). Moreover, the overall gradual learning pattern was particularly true for AG, whose Int1 (75%) score did not differ significantly from the pre-test (51%, $p=.163$). Finally, only for AO, the pre- (50%) and post-test (72%) scores did not show significant improvement ($p=.074$). These results show different learning trajectories across groups for Tone 2.

For Tone 3 (Figure 2-c), a significant effect of Test was again found [$F(1,53)= 61.0$, $p<.001$], with the pre-test (69%) being lower than the subsequent tests across groups (Int1: 89%, Int2: 90%, Int3: 92%, posttest: 93%, $ps<.001$ for all). Subsequent individual group analyses revealed that only the AG group showed exceptions to this general pattern, in that their pre-test score (65%) was not significantly different from their Int1 (82%, $p=.068$) and Int2 (81%, $p=.138$) scores, showing a slower learning effect.

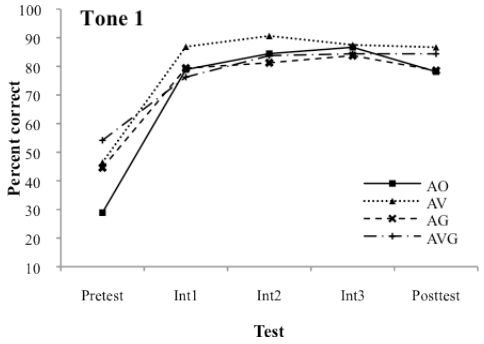


Figure 2-a: *Percent correct identification scores for each group in five tests for Tone 1.*

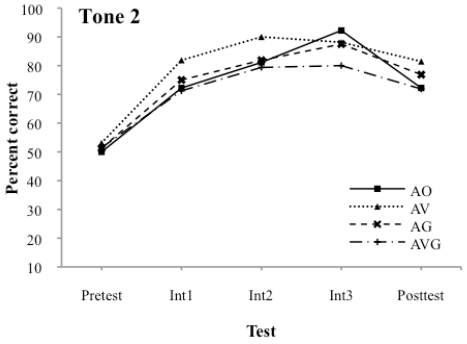


Figure 2-b: *Percent correct identification scores for each group in five tests for Tone 2.*

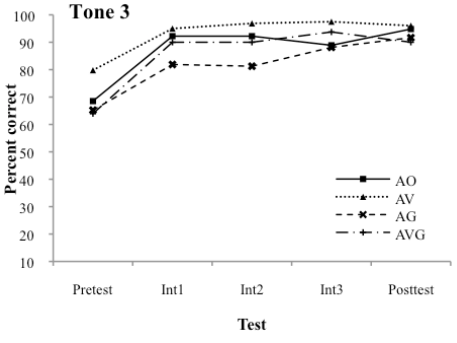


Figure 2-c: *Percent correct identification scores for each group in five tests for Tone 3.*

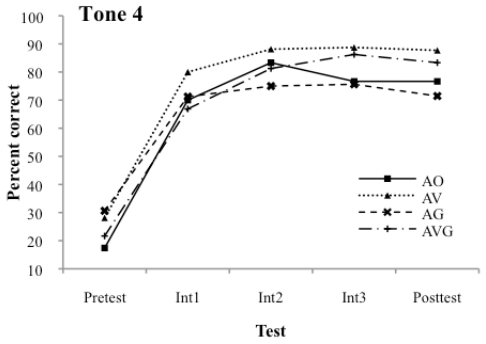


Figure 2-d: *Percent correct identification scores for each group in five tests for Tone 4.*

3.3. Summary

The results demonstrate that the performance for all four groups improved through training. The perception for each tone was slightly different, with Tone 4 being the most difficult tone to identify. Moreover, the results for individual tone and group showed robust differences in Tone 4 learning, where only the AVG group exhibited a delay in improvement during training and a failure in generalization at post-test. Similar gradual learning patterns were also observed for Tone 2 and Tone 3 with the AG group. Furthermore, for Tone 2, the AO group did not improve significantly from the pre-test to the post-test.

4. Discussion and concluding remarks

Overall, the results showed substantial improvement after training for all groups, which is in keeping with previous research indicating that speech perception from multimodal presentation benefits L2 speech learning [2], [7], [8], [15].

However, further analysis with individual tones did reveal differences in the extent of improvement among the four training groups. Most noticeably, compared to the other three trainee groups, the AVG group exhibited a slower learning curve during training and a lack of generalization after training for Tone 4 (with high-falling pitch), also shown as the most difficult tone to identify. It is possible that the cognitive resources for the identification of Tone 4, the most challenging tone, were overtaxed by too many channels of input, which in turn resulted in lower performance of the AVG group compared to the other trainees. This supports the "information overload" hypothesis, being in line with previous claims that gestural input in addition to auditory and visual input may increase the cognitive load, resulting in an inhibitory effect in learning [6], particularly when phonetic demands are high [5].

Results for the learning trajectory of Tone 2 (with rising pitch) showed an increased identification from the pre- to post-training tests for AV and AG but not for AO, suggesting that the addition of either visual or gestural input could result in a significant improvement in tone identification. This is in line with the results for Tone 4, where similar levels of improvement were found in both AV and AG groups. The positive effects of visual input support the previous findings that visual articulatory movements involving the head, mouth, and eyebrows provide robust correlates to tone perception [16], [17], [18]. Unlike visual speech information present in a speaker's face, which has anticipated and fixed articulatory configurations for the resultant speech sounds, the gestures used in the current study involve spatial changes which are not directly bound to speech. However, the AG group's improvements demonstrate that the trainees were able to make the spatial-auditory association embodied by the hand gestures tracing pitch trajectories and utilize that to aid their learning, just as how similar gestures could guide musical pitch processing [9] and aid L2 prosodic perception [7].

Nonetheless, it is also worth noting that visual and gestural modalities are not always equal in terms of degree of improvement. For both Tone 2 and Tone 3, the increase in tone identification accuracy during training for the AG group was more gradual than that for AV. One possible reason for the slower improvement of the AG group may be due to the similar gestural trajectories of the two tones (both involving a long rising pitch contour), which may have made the two gestures less distinct. Thus it may have taken the AG group longer to effectively integrate the gestural information as compared to the AV group whose input involved more predictable articulatory-auditory correspondence.

Taken together, the differing patterns observed within these results demonstrate the complex role of multimodal input in L2 speech perception. The most promising finding is that co-speech gestural as well as visual articulatory information can aid L2 speech learning. However, facilitation from multiple input domains may not be additive and may be constrained by phonetic and cognitive demands. When perceiving phonetically challenging sounds, learners (particularly those at the elementary level) may find multiple input resources distracting, as they may not be able to simultaneously focus on all these input domains and effectively integrate them into a single percept. As such, training with fewer input domains may reduce the cognitive load required to attend to multiple channels [6], [19]. Moreover, the different patterns for individual tones and training groups suggest that multimodal facilitative effects may take place in a complimentary manner. Learners could be trained to selectively focus on those domains that can most readily and reliably aid their learning. Further research may delve into these avenues to better understand the complex relationship between L2 prosody learning and gestures.

5. Acknowledgements

All authors made equal contributions to this study. We thank Drs. Yukari Hirata and Spencer Kelly (Colgate University) and Drs. Allard Jongman and Joan Sereno (University of Kansas) for their valuable comments. We also thank Anthony Chor, Mathieu Dovan, Courtney Lawrence, and Lindsay Leong (SFU) for their assistance in data collection and analysis. This research has been funded by research grants from SFU Vice President Academic and the Social Sciences and Humanities Research Council of Canada to YW.

6. References

[1] Davis, C., & Kim, J., "Audio-visual interactions with intact clearly audible speech," Q. J. Exp. Psychol. 57A, 1103–1121, 2004.

[2] Hazan, V., Sennema, A., Faulkner, A., & Ortega-Llebaria, M., "The use of visual cues in the perception of non-native consonant contrasts," J. Acoust. Soc. Am. 119, 1740–1751, 2006.

[3] Krahmer, E. & Swerts, M., "The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception," J. Memory Lang. 57, 396–414, 2007.

[4] Kelly, S. D., Manning, S. M., & Rodak, S., “Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education,” Lang. Linguistics Compass 2, 569-588, 2004.
[5] Kelly, S., & Lee, A., “When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high,” Lang. Cognitive Proc. 2, 793-807, 2012.
[6] Hirata, Y., & Kelly, S. D., “Effects of lips and hands on auditory learning of second-language speech sounds,” J. Speech Hear. Res. 53, 298-310, 2010.
[7] McCafferty, S., “Gesture and the materialization of second language prosody,” IRAL 44, 197–209, 2006.
[8] Kelly, S. D., McDevitt, T., & Esch, M., “Brief training with co-speech gesture lends a hand to word learning in a foreign language,” Lang. Cognitive Proc. 24, 313-334, 2009.
[9] Connell, L., Cai, Z. G., & Holler, J., “Do you see what I'm singing? Visuospatial movement biases pitch perception,” Brain Cognition 81, 124-130, 2013.
[10] Lively, S. E., Logan, J. S., and Pisoni, D. B. ~1993!. ‘‘Training Japanese listeners to identify English /r/ and /l/ II: The role of phonetic environment and talker variability in learning new perceptual categories,’’ J. Acoust.Soc. Am. 94, 1242–1255
[11] Wang, Y., Spence, M., Jongman, A., & Sereno, J.A., “Training American listeners to perceive mandarin tones,” J. Acoust. Soc. Am. 106, 3649-3658, 1999.
[12] Cooper, A. & Wang, Y., “The influence of linguistic and musical experience on Cantonese word learning,” J. Acoust. Soc. Am. 131, 4756-69, 2012.
[13] Liu, S. & Samuel, A. G., “Perception of mandarin lexical tones when F0 information is neutralized,” Lang. Speech 47, 109-138, 2004.
[14] Chao, Y. R., “Mandarin Primer: An Intensive Course in Spoken Chinese,” Harvard University Press, 1948.
[15] Wang, Y., Behne, D.M., & Jiang, H. “Linguistic experience and audio-visual perception of non-native fricatives,” J. Acoust. Soc. Am. 124, 1716- 1726.
[16] Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Ciocca, V., Morris, R. H., Hill, H.,Vignali, G., Bollwerk, S., Tam, H., & Jones, C., “The perception and production of phones and tones: The role of rigid and non-rigid face and head motion,” Proc. 7th Int’l Seminar on Speech Production, pp. 185–192, 2006.
[17] Chen, T.H. & Massaro, D.W., “Seeing pitch: Visual information for lexical tones of Mandarin-Chinese,” J. Acoust. Soc. Am. 123, 2356-2366, 2008.
[18] Mixdorff, H., Hu, Y., & Burnham, D., “Visual cues in Mandarin tone perception,” Proc. InterSpeech, pp. 405–408, 2005.
[19] Hardison, D. A., “Acquisition of second-language speech: Effects of visual cues, context, and talker variability,” Appl. Psycholing. 24, 495-522, 2003.

The OMe (Octave-Median) scale: a natural scale for speech melody.

Céline De Looze[1], *Daniel Hirst*[2,3]

[1]Speech and Phonetics Laboratory, CLCS, Trinity College, Dublin, Ireland
[2]Laboratoire Parole et Langage, CNRS & Aix-Marseille Univ., France;
[3]School of Foreign Languages, Tongji University, Shanghai, China
deloozec@tcd.ie, daniel.hirst@lpl-aix.fr

Abstract

Fundamental frequency, the primary acoustic correlate of speech melody, is generally analysed and displayed using a linear scale (Hertz) or a logarithmic one, generally in semitones and usually offset to an arbitrary reference level such as 100 Hz. In this paper we argue that a more natural scale for analysing speech is the OME (Octave-MEdian) scale, using the octave (o) as the basic unit, offset to the median value of the speaker's range. We present results showing that a reasonable estimate of a speaker's neutral pitch range can be obtained directly from the median.

Index Terms: Tone, intonation, melody, pitch, octave.

1. Introduction

Although one can observe some non-linearity in the perception of the pitch of speech sounds, fundamental frequency is unquestionably the main acoustic correlate of perceived pitch height. Psycho-acoustic scales for the study of speech have also been proposed, particularly the *Mel*, *Bark* and *ERB* scales. The relevance of these scales, however, remains unproven. A recent study [25], for example, using a task of replicating pitch contours between male and female voices, showed that a logarithmic scale better reflects the performance of speakers than either a linear or a psycho-acoustic scale. The physical scale in *Hertz* (cycles per second) is very often transformed, in studies of prosody, to a logarithmic scale, most often expressed in semitones with a reference value (called C0), arbitrarily set at 16.3516 Hz [28]. This reference was chosen as being close to the lowest pitch perceptible to the average human ear and as well as corresponding to the tuning of A_4 (= 'A above middle C') to 440 Hz, in conformity with the ISO Standard ISO16 [19].

Fant [12] proposed the St (= semitone) unit defined as:

$$St = 12 \cdot \frac{ln(\frac{Hz}{100})}{ln(2)} \quad (1)$$

The semitone is, however, in no sense a natural unit of measurement. It is, in fact, the product of a complex history of Western classical music culture, corresponding to the division of the octave into 12 equal intervals, an idea that had first been described in a treatise published in China in 1584 [20]. In Europe, the scale of 12 equal semitones equal (= *equal temperament*), has been used increasingly since the 18th century to tune keyboards, replacing the *natural scale* ('just intonation') previously used, or Bach's *well-tempered scale*. All these scales were the result of a search for a compromise which would allow musicians to modulate from one scale to another without introducing major discord and without having to switch keyboards.

In different civilizations at different times, the use of different sets of notes can be observed. Practically all of these scales, however, have in common the fact that the names of the notes are generally the same, regardless of the octave. Thus, in the Western classical scale, for example, the sequence *Do Re Mi Fa Sol La Ti Do Re Mi ...* etc. can be repeated indefinitely within the physical limits of sound production.

This circularity (also known as *chromatic repetition*) seems to stem from a physiological basis of human perception [6, 5] including that of neonates [22] and also that of rhesus monkeys [27]. It was observed as early as the 60s, in an anatomical study of a cat, that the auditory thalamus is organised in stacked layers or laminae. It was suggested that this organisation may have a specific function in the processing of acoustic frequencies [24]. [23] and [18] later demonstrated that the auditory thalamus of the cat actually contains a neural chroma map, underlying an octave architecture. While the functional role of the mammalian auditory thalamus octave topography still needs to be determined, recent research has suggested that it may cause, as a side effect, the octave circularity of pitch that has been observed in the rhesus monkey as well as in humans [6]. This study investigated the effect on a musician with absolute pitch, of the neurotropic medical drug carbamazepine (CBZ), known to have a down-shift pitch side effect,in order to better understand the mechanism of octave circularity of pitch. They observed in their subject, during a pitch identification task, an internal tone-scale or chroma representation. When CBZ was taken, pitch shift was indeed observed but the pattern of tone representation remained unchanged. This suggests that the human brain may be hard-wired for octave-circular pitch perception.

In any case, it is the octave, not the semi-tone, which appears clearly as the basic unit for the natural perception of the pitch of speech sounds and music.

The use of the *semi-tone* (or in more precise studies, its subdivision the *cent* [where $1st = 100cents$) has paradoxically had the negative effect of masking the importance of the octave as a basic unit in pitch production and perception. Rereading a number of studies on pitch range with this in mind reveals a very large number of cases where authors report an interval close to an octave (= 12sts) or half-octave (= 6sts) without drawing attention to this fact, or perhaps, even, sometimes without having noticed it. In [26], for instance, the authors reported a f0 mean at the beginning of sentences produced in neutral, happy, angry and scared voices of 6.72, 12.64, 12.52 and 12.38 sts respectively. If we calculate the difference between the neutral voice mean f0 and the other voices mean f0, it reveals for each 'arousal' voice a shift of half-an-octave.

The intervals octave and half-octave may play a specific role in speech production. [4] investigated the pitch contours of utterances produced under two conditions (in a normal voice in a quiet room vs. in a louder voice when exposed to noise over headphones), and observed for instance a raising of half-an-octave for the increased loudness condition. A raise of a half-an-octave or an octave may be used to convey specific linguistic and paralinguistic functions in speech, e.g. signaling focus, topic change, turn-taking as well as expressing arousal.

2. The octave as the basic unit for the perception of pitch.

We recommend the systematic use of the octave (o) and its subdivision the millioctave (mo) for the study of pitch. For precise measurements, the (mo) gives, in fact, approximately the same degree of precision as the cent [$1mo = 1.2cents$] and has the advantage of being in conformity with the general practice of the *International System of Units: SI,* in which prefixes corresponding to an exponent divisible by 3 are generally preferred.

As a derived *SI* unit, the octave can be defined as:

$$o = log_2(s^{-1}) \quad (2)$$

where s is the duration in seconds of a period.

The second author has suggested elsewhere [13, 14] that there may also be a physiological explanation for the octave and half octave as a basis for the production of melodic intervals. [14] reported an experiment where these two intervals were observed as modal values in a task of producing varied contours on isolated syllables in French, *oui* and *non*.

In so far as the vocal folds behave like vibrating strings, the relationship between tension and frequency is governed by Mersenne's law which states that

> *The frequency of a vibrating string is proportional to the square root of its tension.*

A doubling of the tension would consequently correspond to a rise of half an octave. This might explain why the intervals - octave and half octave - seem to be frequent in the production of speech melody, even though a rise or fall of an octave on a single syllable is certainly not perceived in its entirety.

3. Corpora

Four corpora, a total of about 2 hours of speech, were selected for this study: these consisted of extracts of the *PFC* and the *CID* corpora for French and the *PCA* and *Aix-MARSEC* corpora for English. These are briefly described below.

The PFC corpus (Phonology of Contemporary French) [9].

We selected 10 French speakers, from the region of Marseille, 6 female and 4 male. We chose the recordings of their production reading aloud a passage of text, such as an extract from a regional newspaper. This corresponded to approximately 30 minutes of recording.

The CID corpus (Corpus of Interactional Data) [2].

We selected six French speakers, from the region of Marseille, 3 male and 3 female. The recordings corresponded to conversations, where speakers discuss either professional conflicts or unusual situations in which they had found themselves, a total of 30 minutes of recording.

The PCE corpus (Phonology of Contemporary English) [7].

We selected eight English speakers, four male and four female, from the North of England. We chose the the recordings of their production reading aloud a passage, such as an extract from a regional newspaper. This corresponded to approximately 25 minutes of recording.

The Aix-MARSEC corpus [1]. The recordings correspond essentially to extracts from the BBC made in the 1980's.

We selected 51 speakers, 13 women and 38 men. 11 types of production are represented: comments, newsletters, public speech, religious programs, documentaries, fiction, poetry, dialogues, propaganda, etc. This represented a total of approximately 50 minutes of recording.

4. Estimation of pitch range

In [11] we reported results based on the 4 corpora, in English and French, described in the previous section, which showed that, in the production of natural speech, the lower range of fundamental frequency systematically corresponds to half an octave below the median pitch of a speaker's voice, and the upper range generally extends between half an octave and one octave above the median.

The pitch range of a speaker (ie the tonal space actually used in an utterance) is generally measured by two parameters: its *height* (or key) and its *extent* (or span) [21]. Pitch height is generally measured by taking the mean or the median of the distribution of f_0, or by taking the mean value of points considered as representative targets. The span of the pitch range can be calculated by comparing the minimum and the maximum value produced in an utterance or the average values for high and low targets.

A measurement of pitch based on the analysis of tonal targets can be both costly and error prone, especially if the targets are annotated manually.

In this study, we calculated the value of pitch range with respect to the median of the distribution of f0 since this is more stable than the mean which is influenced by extreme values some of which may be erroneous. In addition, the median is a non-parametric measure, independent of the unit or scale of measurement. The median value, in other words, is always the same value whether it is measured on a linear scale or on a logarithmic scale or on one of the psyco-acoustic scales mentioned above.

To avoid the problems inherent in manual measurements, our calculations were carried out using the *Momel* and *INTSINT* algorithms [15].

The Momel algorithm takes as input the raw fundamental frequency curve and models it as the sum of two components, a 'macroprosodic' component on the one hand, consisting of a smooth continuous underlying function corresponding to the intonation pattern of the utterance, and a 'microprosodic' component consisting of a sequence of functions, some of which are discontinuous and which correspond to the local effect of the different individual speech sounds. For more details on these two components cf [16], for a recent attempt to model the 'microprosodic' component for speech synthesis in Arabic, cf [8].

The INTSINT algorithm takes as input the target points detected by the automatic pitch modeling algorithm (Momel) and codes these targets using an alphabet of 8 discrete symbols. The symbols **T**(op) and **B**(ottom) delineate *high* and *low* values of the speaker's pitch range, respectively, while **M**(id) codes its central tendency. Targets coded **H**(igher), **L**(ower), **S**(ame),

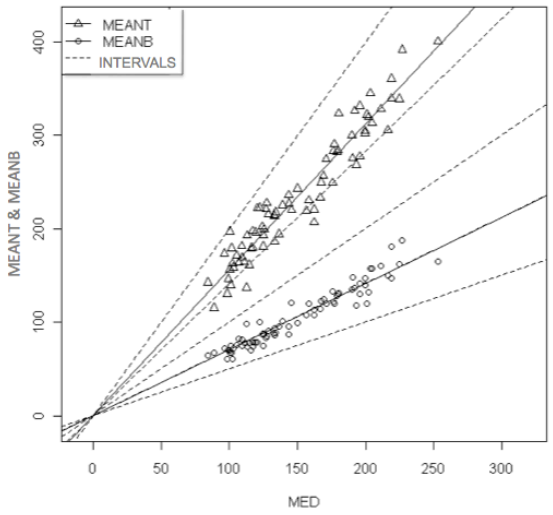


Figure 1: Linear regressions corresponding to Mean-**B** and Mean-**T** are traced in continuous lines and the dotted lines represent, from top to bottom, the intervals +octave, +half-octave, unison, -half-octave and octave compared to the median.

U(pstepped) or **D**(ownstepped) are defined not globally but locally, with respect to the value of the immediately preceding pitch target and are defined as being, respectively, *higher*, *lower*, *equal to*, *slightly higher* or *slightly lower* than the preceding target. For the precise definitions of the coding cf. [15].

In this study, we focus on the values obtained for the absolute **T** and **B** tones since all the other values are dependent on the value of the preceding target.

As we mentioned, one of the most common ways to measure a speaker's pitch range is by comparing the mean values of pitch which have been identified as corresponding to high tones and low tones. It is instructive to look at the correlations between the mean values of the low tones (**B**) and high tones (**T**), as determined by the INTSINT algorithm, and that of the *median* of the pitch distribution.

We find, in fact, two strong correlations. Affine relations are as follows:

$$B = 0.741 * median - 5.52$$

$$T = 1.537 * median + 3.75 \quad (3)$$

Significance tests of regression coefficients are highly significant $p < 2^e - 16$. The critical probabilities of the offsets are, however, not significant (p = 0.161 and 0.659). An adjustment of the model without the offset gives:

$$B = 0.706 * median$$

$$T = 1.561 * median \quad (4)$$

For **B**, we find a coefficient of determination (R^2) of 0.92 and for **T** 0.91. This means that it is possible, at least as a reasonable approximation, to predict the limits of the register of a speaker and hence its span, from the median of the distribution of f_0.

ANOVAs on the prediction of average low (**B**) and high (**T**) tones from the median showed no effect of either sex (p = 0.0917 (**B**) and 0.381 (**T**)) or language (p = 0.170 (**B**) and 0.274 (**T**)), or type of production (p = 0.134 (**B**) and 0.368 (**T**) on the slopes of linear regression. It is therefore possible, from the value of the *median*, regardless of the sex of the speaker, both in English and French and whatever the style of speech, to make a reasonably good prediction of the limits (span) of the pitch range. In fact, [10] showed that the relationship between the height and span of the pitch range is actually more complex. In the model given in 4, the range in Hz is strictly proportional to the height because the relationship between T and B is fixed. An even better correlation is obtained with the values on a logarithmic scale, corresponding to a model where the span in octaves is proportional to the height, going from one octave for a low-pitched voice to a little less than an octave and a half for a high-pitched voice.

This co-variation was pointed out by [21] in his discussion of pitch range. The author explains that the difficulty of admitting two dimensions for the register is that these two dimensions co-vary.

5. The OMe scale. A natural scale for the melody of speech

It is interesting to note in the relationships defined in (2) that the coefficient 0.706 corresponds almost exactly to half an octave($log2(0.706) = -0.502$) and the coefficient 1.561 is just slightly over half an octave ($log2(1.561) = 0,642$). We can therefore conclude that the average of the high tones and the average low tones, i.e. the limits of the range of a speaker, for unemphatic speech, usually correspond to about an octave centered on the speaker's median.

This led us to propose [11] a new scale of measurement: the *OMe* (Octave-MEdian) scale defined by the formula:

$$ome = log_2(\frac{Hz}{median}) \quad (5)$$

where *median* corresponds to the median value of f_0 for the recording.

Figure (1) gives a graphical representation of the average low tones (**B**) and the average high tones (**T**) compared to the speaker's *median* pitch. The corresponding linear regressions are plotted in solid lines and dotted lines represent the intervals + *octave*, + *half-octave*, *unison*, - *half-octave* and - *octave* with respect to the *median*. The linear regression on the mean of the low tones (**B**) coincides with the half-octave below the median so that the two lines are not distilnguishable in the figure. That of the average of the high tones (**T**) falls between half an octave and one octave above the median. These musical intervals, defined relative to the median, can therefore be used to estimate the range of a speaker with a reasonable reliability.

They also allow us to propose, as suggested above, a natural scale for the analysis and visualisation of the melody of speech defined in octaves, centered on the median, which we call the *OMe* (Octave-Median) scale.

Figure 2 shows an example of expressive speech by a radio broadcaster pronouncing the sentence 'He dra**MA**tically flourished a **CO**py of Time from nineteen fifty-**THREE**.". As can be seen, the places where the pitch goes above half an octave above the median, correspond precisely to the parts of the words which are perceived as being emphasised expressively.

Figure (3) illustrates the sentence "What can I have for dinner tonight? " read by one male and one female speaker.

The visualisation of these recordings was obtained automatically from the signal and TextGrid using the Praat plugin

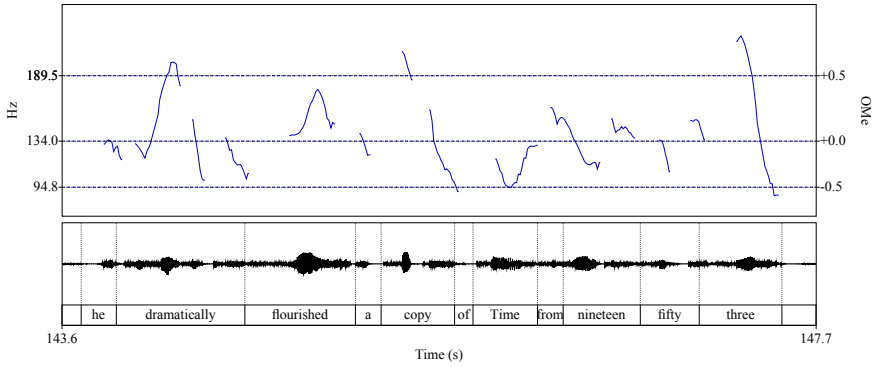


Figure 2: An extract of journalistic speech "He dra**MA**tically flourished a **CO**py of Time from nineteen fifty-**THREE**."

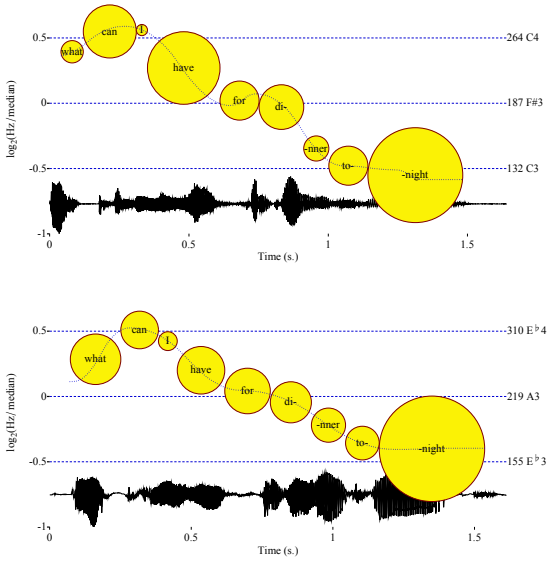


Figure 3: Graphical representation of readings of the sentence "What can I have for dinner tonight." by one male (top) and one female (bottom)speaker, displayed using the *OMe scale* (Octave-Median). The diameter of the circles corresponds to the duration of the syllables. The horizontal blue lines delimit the central octave surrounding the speaker's median pitch.

ProZed [17], which is freely downloadable from the *Speech and Language Data Repository* at: http://sldr.org/sldr000778/en.

The diameters of the yellow circles correspond to the syllable durations and the dashed blue line corresponds to the Momel curve. The horizontal dotted lines correspond to the speaker's median (middle line) and a half octave above and below the median, delimiting the speakers unemphatic pitch range corresponding to the median-centred octave. The values of the *Median* and the *Top* and *Bottom* of the central octave are given both in Hz and as musical notes as defined with respect to concert pitch at 440 Hz.

With this technique, the optimal parameters for the analysis of the fundamental frequency of the speaker are automatically determined from the median pitch.

6. Conclusions.

We propose in this paper that it is the octave, rather than the semitone, which should be considered the basic unit of a scale for natural speech prosody. This follows evidence reported in several studies based in neuronatomy, neurophysiology, behavioral studies, speech production as well as speech perception. In particular, we propose the use of the OMe scale to define an automatic display of the fundamental frequency curve. The reference (key) for such a scale is given by the median of the speakers fundamental frequency. The *Bottom* of the central octave of the speakers voice is consequently half an octave below the median while the *Top* is half an octave above.

The Bottom and Top lines of the display should not be thought of as physical obstacles for speakers. Obviously, in more spontaneous corpora we are likely to find a larger pitch range - up to two octaves has been reported in the literature. Pitch often goes beyond these lines, particularly in the case of the Top , but when it does so, it may be taken as a good sign that the speech is expressive or signalling important information.

Since the expressive use of pitch particularly concerns the top of the range, it is natural that the distribution of pitch in these cases will be skewed. It should be noted, however, that taking the median pitch rather than the mean pitch as reference value largely reduces the impact of the skewness, but this naturally remains to be tested on much more data.

Further research is also necessary on the variability of the median pitch for given speakers. As mentioned previously, the octave and half-octave intervals may be used in structuring the discourse (e.g. indicating focus, topic change, turn-taking) as well as in expressing arousal or attitudes. The relation between speakers' span and median voice may have facilitated the perception of linguistic and paralinguistic functions of pitch range in speech production across genders, languages and speaking styles and may be interesting to examine in terms of prosodic universals. It opens up the debate on the formal and functional aspects of speech prosody as a result of learning and experience or of having some basis in the operation of the auditory system.

7. References

[1] Auran, C., Bouzon, C.; Hirst, D.J. "The Aix-MARSEC Project: An Evolutive Database of Spoken British English." in Proceedings of the 2nd International Conference on Speech Prosody, Nara, Japan 2004.

[2] Bertrand, R.; Blache, P.; Espesser, R.; Ferre, G.; Meunier, C.; Priego-Valverde, B.; Rauzy, S. "Le CID -Corpus of Interactional Data-: protocoles, conventions, annotations." in Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence, 25, 25-55., 2007.

[3] Boersma, P.; Weenink, D. Praat, a system for doing phonetics by computer. http://www.praat.org [version 5.3.41, February 2013], 1992 (2013).

[4] Braun, M. "Speech mirrors norm-tones: Absolute pitch as a normal but precognitive trait." Acoustics Research Letters Online 2, 8590.

[5] Braun, M. "A retrospective study of the spectral probability of spontaneous otoacoustic emissions: Rise of octave shifted second mode after infancy." Hearing Research 215, 39-46. 2006.

[6] Braun, M.; Chaloupka, V. "Carbamazepine induced pitch shift and octave space representation.", Hearing Research 210, 85-92. 2005.

[7] Carr, P.; Durand, J.; Pukli, M. "The PAC project: Principles and Methods." Tribune des Langues Vivantes 36: 24-35. 2004.

[8] Chentir, A.; Guerti, M.; Hirst, D.J. "Extraction of standard Arabic micromelody. " Journal of Computer Science, 5(2):8689, 2009.

[9] Delais-Roussarie, E.; Durand, J. Corpus et variation en phonologie du français: méthodes et analyses. Presses Universitaires du Mirail. 2003.

[10] De Looze, C. Analyse et interprétation de l'empan temporel des variations prosodiques en anglais et en français. Doctoral thesis, January 2009. LPL and Aix-Marseille University, Aix-en-Provence, France. 2009.

[11] De Looze, C.; Hirst, D.J. "L'echelle OME (Octave-MEdiane): une échelle naturelle pour la mélodie de la parole." in Actes des XXVIIIes Journées d'Etude sur la Parole, Mons, Belgium, May 25-28 2010.

[12] Fant, G; Kruckenberg, A.; Gustafson, K.; Liljencrants, J.. "A new approach to intonation analysis and synthesis of Swedish.", in Proceedings of the First International Conference on Speech Prosody, Aix en Provence. 283-286. 2002.

[13] Hirst, D.J. "Phonological implications of a production model of English intonation.", Phonologica 1980, 195-201. 1981.

[14] Hirst, D.J. "Structures and categories in prosodic representations.", in Cutler and Ladd (eds.) 1983. Prosody : Models and Measurements (Springer, Berlin), 93-109. 1983.

[15] Hirst, D.J. "A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation.", in Proceeedings ICPhS 2007. 1233-1236. 2007.

[16] Hirst, D.J. "The analysis by synthesis of speech melody: from data to models.", Journal of Speech Sciences, 1(1): 5583, 2011.

[17] Hirst, D.J. "ProZed: A speech prosody analysis-by-synthesis tool for linguists.", in Proceedings of the 6th International Conference on Speech Prosody, Shanghai May, 15-18.

[18] Imig, T.J.;Morel, A. " Tonotopic organization in the lateral part of posterior group of thalamic nuclei in the cat." Journal of Neurophysiology 53: 836851 1985.

[19] ISO 16:1975. "Acoustics – Standard tuning frequency (Standard musical pitch).", International Organization for Standardization. http://en.wikipedia.org/wiki/ISO_16 1975.

[20] Kuttner, F. A. " Prince Chu Tsai-Yu's life and work: a re-evaluation of his contribution to equal temperament theory." Ethnomusicology, Vol. 19, No. 2 (May, 1975), pp. 163206.1975.

[21] Ladd, D.R. Intonational Phonology, (Cambridge University Press, Cambridge) 1996 [second edition 2008].

[22] Liu J.; Wang N.; Li J.; Shi B.; Wang H. "Frequency distribution of synchronized spontaneous otoacoustic emissions showing sex-dependent differences and asymmetry between ears in 2- to 4- day-old neonates." International Journal of Pediatric Otorhinolaryngology. 2009 May; 73(5):731-6 2009.

[23] Morel, A. Codage des sons dans le corps genouillé median du chat: évaluation de l'organisation tonotopique de ses différents noyaux. PhD dissertation. Université de Lausanne. Juris, Zurich 1980.

[24] Morest, D.K. "The laminar structure of the medial geniculate body of the cat." Journal of Anatomy 99: 143160 1965.

[25] Nolan, F. "Intonational equivalence: an experimental evaluation of pitch scales." in Proceedings of ICPhS 15, Barcelona, 771-774. 2003.

[26] Paeschke, A; Sendlmeier, W. F. "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements." In Proceedings of the ISCA-Workshop "On Speech and Emotion" (Belfast), SpeechEmotion-2000, 75-80. 2000.

[27] Wright, A. A. ; Rivera, J.J.; Hulse, S.H.; Shyan, M.; Neiworth,J.J. "Music perception and octave generalization in rhesus monkeys.", Joural of Experimental Psychology Gen, Sep, Vol 129 No 3, 291-307 2000.

[28] Young, R. W. "Terminology for Logarithmic Frequency Units", The Journal of the Acoustical Society of America. 11: 134. 1939.

Automatic Discovery of Simply-Composable Prosodic Elements

Nigel G. Ward

Department of Computer Science, University of Texas at El Paso
nigelward@acm.org

Abstract

As a way to discover the elements of prosody, Principal Component Analysis was applied to several dozen contextual prosodic features, sampled at 600,000 timepoints in dialog data. The resulting components are interpretable as prosodic patterns, including some which involve behaviors of both interlocutors. Examining contexts and co-occurring words, many of these have clear interpretations. This suggests that English has at least several dozen prosodic patterns, each with its own communicative function.

Index Terms: principal components analysis, prosodic elements, prosodic patterns, factors, dimensions, contours, superposition, intonation, modeling, dialog, interaction, pragmatics

1. Introduction

To understand a complex machine, one needs to identify the pieces and how they work together. Classical approaches to prosody have found many likely pieces, including targets, contours, impulses, and events, and much has been written about each. However the question of how these elements are combined has received less attention; many models of prosody are vague here. This is a problem because theories that rely on unexplained mechanisms have little predictive power: they are impossible to test rigorously and potentially falsify [1, 2].

Well-specified descriptions of how prosodic elements combine do exist, for example [3, 4, 5, 6]. However so far these have been worked out only for carefully circumscribed sets of phenomena, in datasets where every other prosody-related factor is controlled. The more general trend is, it seems, to give up on modeling prosody in terms of composable elements, instead using machine learning techniques that operate directly over raw features [7, 8]. While this can be of great practical value, the resulting models are tailored to single applications, are difficult to interpret, and do not much advance our understanding of prosody.

This paper presents a way to outflank the problem of combination mechanisms. The novel idea is to start with a composition rule, and to then use it to infer the elements; the reverse of the classical strategy. This guarantees that these elements will compose simply, with no slack in the model. Originally developed for purely practical reasons [9, 10], this method is here presented as it relates to other approaches, with new visualizations, and with more dicussion of the broader implications and prospects.

2. Principal Component Analysis for Prosody

The fundamental assumption of the method is that superposition is the main combining principle for prosodic elements. Thus the elements, whatever they are, are required to be summable, and, when summed in various combinations and weightings, to fully explain the observed reality. The discovery task is accordingly to infer the underlying elements from data. This is an underconstrained problem; however, the desire for models that minimize the number of elements and maximize their explanatory power leads us to Principal Components Analysis (PCA).

PCA can be described in several ways, but it is helpful to view it as an iterative analysis process. In each stage, PCA finds the factor that explains as much as possible of the observed variation, across many datapoints and many variables. It then subtracts out what that factor explains, finds another factor to explain much of the remaining variation, and iterates. For example, if we are interested in statistics on people, including income, wealth, family size, number of cars, age, education level, food budget, and so on, the first underlying factor may be something like socioeconomic status, the second may be related to age, the third may be gender, and so on. The observed variable values for any datapoint (person) are modeled as linear combinations of the factors, and conversely, one can go from the observed values for any datapoint to the values of the underlying factors trivially, with a simple matrix multiplication.

PCA is good for dealing with variables which are highly correlated and thus mutually partially redundant. This is commonly the situation in prosody, and PCA has indeed been used here, for identifying the prosodic and other vocal parameters relevant to emotional dimensions [11] and to levels of vocal effort [12], for categorizing glottal-flow waveforms [13], for finding the factors involved in boundaries and accents [14], for characterizing ambiance [15], and for purely practical purposes [16, 17, 18, 19]. A related method, Functional Data Analysis (FDA), has also been applied to prosody, including for identification of the key dimensions of variation in pitch contours [20, 21, 22, 23]. Despite all these precursors, our strategy, of using PCA as a way to discover prosodic elements, is something new.

3. Base Features

In our approach, the datapoints input to the PCA are points in time, and the variables are various prosodic features. PCA is then applied to discover the underlying factors, and these are the elements of prosody.

Because a prosodic value at a point in time is meaningless without context, for each datapoint we use several dozen base features to broadly represent the local prosodic context. For example, in addition to the average pitch over the past 50 milliseconds, we also use the average pitch over a 50 millisecond window centered 75 ms in the past, and over a 100 ms window centered 150 ms in the past, and so on, for both past and future windows, spanning about 6 seconds centered around the point of interest. Including such features enables the use of PCA for

time-series analysis [24, 20, 21]. Unusually, our features are not uniformly spaced, but are denser closer to the point whose context is being considered, as detailed elsewhere [9, 25].

Given that the features are from the local prosodic context, each PCA-derived factor will represent a patterning of prosodic features over that context: a 'phonological entity with a distinct time course' [1]. For example, one factor has a region of speech with a slowly dropping pitch, followed by a region that is quiet and slow in rate, followed by a second region of speech that has an early fast region, but then turns slower and lower. The bottom half of Figure 1 shows a visualization of this.

Mathematically a factor can be present either with a positive weight or a negative weight. However it can be difficult to intuitively understand the contributions of a factor when it has a negative weight for a given datapoint. Accordingly the discussion below will focus on one or the other of the two poles of each factor (dimension), discussing either the pattern characterizing points that are high on the dimension or the one for low points. For example the pattern in Figure 1 is the high side of dimension 6.

For PCA to work, the base features should be continuous-valued, on scales for which summation is meaningful. The features we have used approach this ideal but imperfectly. For loudness we use log energy normalized per track to correct for different recording conditions and different speakers. For pitch height we use percentile in the distribution of pitch seen for that track, thus again normalizing for speaker. For pitch range, we similarly use the number of percentiles between the highest pitch point in the window and the lowest. For windows without voiced frames, the mean pitch values and ranges are used instead. For rate, we use a simple frame-by-frame energy-shimmer measure. These features are from our standard inventory, chosen for utility for modeling turn-taking prosody and for language modeling. Better choices could certainly be made, but fortunately PCA is robust to imperfect and noisy features. Finally, following standard practice, we z-normalize all features before applying PCA.

In contrast to previous uses, here we apply PCA to dialog data, with featuresets accordingly including features of both speakers' prosodic behavior. As a result, the top factors discoved by PCA tend to be those which explain variance not only in the speaker's behavior but also in the interlocutor's behavior. That is, this predisposes the method to find patterns that have interactional significance, with ones that relate purely to one speaker's behavior destined to rank lower.

Further, again in contrast to previous work, we applied PCA to a heterogeneous and large data set. This is because our aim is to see what we can find, rather than to refine some model or answer some question. In particular, most of our work has been with the Switchboard corpus of American English telephone conversations. In each case we built the models using about 600,000 data points, the maximum our computer's memory could handle for PCA. These were obtained by sampling every 10 milliseconds during 16 dialogs, without regard to any notion of sentence, utterance, or turn. Thus each model is built from about 100 minutes of dialog, across multiple speakers, multiple dialog activities, and multiple topics.

Finally, PCA involves an (optional) interpretation step. While some researchers are happy with fully automatic methods, we think that human judgments are a necessary part of scientific inquiry, and in this respect our method is aligned with classical approaches. We differ in choosing to apply the human interpretation after the data-reduction step, rather than up front.

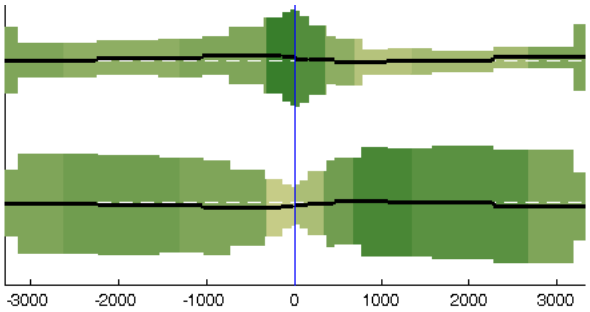


Figure 1: The pattern exemplifying the high side of factor 6 (dimension 6). Time is in milliseconds. First-speaker features are on top and second-speaker features below. Line width shows volume, with the median volume seen at the far left and right of the figure. Height shows pitch, with the median indicated by the dashed line. Darkness shows rate. While what is shown is the strengths of factor loadings — not directly the pitch height, volume or rate — it is not seriously misleading to interpret the figures as indicating typical values for the features across time.

4. Findings

4.1. Interpretable Elements

While PCA is guaranteed to find elements, there is no guarantee that they will be interpretable. While uninterpretable models of prosody can still be useful, interpretable ones are preferable. Luckily, most of the factors that PCA outputs have indeed been interpretable, with each pole corresponding to a simple pattern or 'construction' [26, 27] with an identfiable meaning or function. These generally relate to meanings and functions familiar from the prosody literature. (Although not so far to meanings at the degree of specificity claimed for some sentence-level contours [28, 29, 30].) Space allows just two examples:

4.1.1. The Upgraded-Assessment Pattern

Switchboard dimension 6 is positive to the extent that: the interlocutor was speaking loudly but trails off and this occurs with a low pitch, during which while the speaker was quiet; followed by a loud region by the speaker with a slightly expanded pitch range and increased speaking rate (the upgraded assessment); followed after a short pause by a long and loud continuation by the interlocutor. This is seen in Figure 1.

An example very high on this dimension occurred 309 seconds into dialog sw2402, where A has spoken favorably about warm places:

> A: a lot of people go to Arizona or Florida for the winter and they're able to play all year round but
>
> B: yeah, oh, Arizona's beautiful!

Words frequent in this context include *neat*, *ooh*, *absolutely*, and '*laughter-right*' (laughed tokens of the word *right*).

Thus three kinds of evidence — the feature loadings, impressions of the pragmatics, and statistics of the co-occurring words — provide convergent evidence for a coherent interpretation of the pattern: that it involves one person seeking and the other displaying empathy, in extreme cases in the form of an upgraded assessment. This matches well with a previously-described prosodic construction [31]: a pattern in which a listener expresses agreement with an assessment by producing an

upgraded version, for example when one speaker tentatively observes *it's pretty* and the other follows with *absolutely gorgeous* with increased volume, pitch height and pitch range, and 'tighter' articulation. This is exactly what happens when dimension 6 is high.

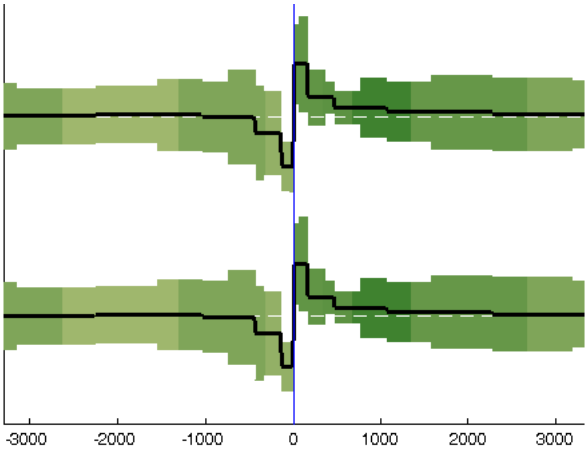


Figure 2: Dimension 26, high side, as above.

4.1.2. The Interestingness-Signalling Pattern

Figure 2 shows dimension 26 of the Switchboard corpus. The loadings on this factor do not distinguish between the tracks, but this is just an artifact of the prevalent cross-track bleeding in this corpus, which affects many of the lower dimensions. Examining places in the corpus where this factor was strongly present, we found that these frequently aligned with backchannels. Quantitatively, of all *uh-huh*s in the data, 79% occurred in contexts where factor 26 was present with a positive weight, significantly more than a 50% baseline. Looking at the behavior in the other track, places where factor 26 was high were often places where one speaker was finishing up the delivery of one piece of information and preparing to continue on with some elaboration or new aspect. In general, this pattern of joint behaviors [32] signals that the recent content was interesting and will be interesting again soon.

Looking at the feature loadings, there is a very salient region of low pitch just before the point of interest, for about 150 milliseconds. This is the pattern that previous work has identified as a cue for an interlocutor backchannel [33]. Despite various work aiming to refine and elaborate this cue [34, 35], not much more had been found. However from the figure we can easily read off more information: that this pattern involves a slightly increased pitch for about a second, followed by a short, somewhat louder region (often a content word), followed by a short low-pitch region with reduced volume, and then, about a second after the backchannel, a short region with faster rate (as the speaker resumes the turn with a fresh start). Here PCA serves to reveal the larger pattern encompassing the salient feature.

Just to complete the story, in the opposite pattern, characterizing points where this factor was strongly negative, the speaker was typically involved in a narrative and speaking with low volume, and appeared to be downplaying the importance of what he's saying, for example when it was just background to a main point to come later.

4.2. Numerous Elements

In the quest to reduce prosody to the minimum number of elements, it would be helpful to have estimates of how many elements are really needed. PCA is useful for questions like this: often it reveals that superficially complex phenomena can be explained by just a handful of underlying factors. However that was not the case here; on the contrary, the top 25 factors in Table 1 account for no more than 86% of the variance, despite this featureset being one with many strong correlations. Moreover, at least 30 of the factors (and thus 60 patterns) have clear and distinct functions, as summarized in Table 1. These observations suggest that the research program of reductively explaining prosody [36, 37] may not work for the prosody of dialog.

4.3. Continuous-Valued Elements

A recurring debate in prosody involves the extent to which prosodic elements are categorical or continuous. For these elements we can address this by examining the distributions of values on each dimension. All looked normally distributed, with only two exceptions (on the first dimension, which is bimodal, and on the second, which is skew), which suggests that they are not categorical. This interpretation is compatible with the functions they bear, all of which seem likely to be experienced in a graded rather than categorical manner.

4.4. General Elements

While the elements of prosody are likely to vary somewhat with domain and speaker and so on, it would be disappointing if those found by one application of PCA were entirely limited to one specific genre. To see whether this was the case, we tried it on a different corpus, Maptask, and using a different set of base features (computed using the same feature extractors, but with different densities at different temporal offsets). Again we found meaningful patterns, and of those analyzed so far, most are similar to those found in the Switchboard data, as seen in Table 2.

5. Prospects

Ultimately the aims of prosody research must surely include the identification of the inventory of prosodic elements, for any given language (despite the difficulties [29]). This paper has presented a way to use PCA to advance us towards that goal. Next steps include: 1. using better and finer-grained features, to infer the exact shapes and timings of the patterns. 2. using features that are phrase-, word-or syllable-aligned, rather than fixed in width and offset, 3. analyzing different types of data, to replicate findings obtained with other methods, and 4. examining individual differences, as it is unlikely that all speakers of a language have identical prosodic elements, even if the functions are shared.

In addition to superposition, a complete model will certainly require other combining mechanisms: most obviously concatenation, but also probably stretching, warping, alignment, synchronization, assimilation, undershoot, and others, especially for prosodic phenomena at finer time scales. It important to elucidate how superposition works together with these other mechanisms [1, 40].

Beyond acoustical compositionality, the compositionality of the meanings of these patterns is an open question [41]. The heterogeneity of the functions (Table 1) suggests that they could be composed without mutual interference, but this needs to be

1	this speaker talking *vs.* other speaker talking	32%
2	neither speaking *vs.* both speaking	9%
3	topic closing *vs.* topic continuation	8%
4	grounding *vs.* grounded	6%
5	turn grab *vs.* turn yield	3%
6	seeking empathy *vs.* upgraded assessment	3%
7	floor conflict *vs.* floor sharing	3%
8	dragging out a turn *vs.* ending confidently and crisply	3%
9	topic exhaustion *vs.* topic interest	2%
10	lexical access or memory retrieval *vs.* disengaging	2%
11	low content and low confidence *vs.* quickness	1%
12	claiming the floor *vs.* releasing the floor	1%
13	starting a contrasting statement *vs.* starting a restatement	1%
14	rambling *vs.* placing emphasis	1%
15	speaking before ready *vs.* presenting held-back information	1%
16	humorous *vs.* regrettable	1%
17	new perspective *vs.* elaborating current feeling	1%
18	seeking sympathy *vs.* expressing sympathy	1%
19	solicitous *vs.* controlling	1%
20	calm emphasis *vs.* provocativeness	1%
21	mitigating a potential face threat *vs.* agreeing, with humor	$< 1\%$
22	personal stories/opinions *vs.* impersonal explanatory talk	$< 1\%$
23	closing out a topic *vs.* starting or renewing a topic	$< 1\%$
24	agreeing and preparing to move on *vs.* jointly focusing	$< 1\%$
25	personal experience *vs.* second-hand opinion	$< 1\%$
26	signaling interestingness *vs.* downplaying the current information	$< 1\%$
29	no emphasis *vs.* lexical stress	$< 1\%$
30	saying something predictable *vs.* pre-starting a new tack	$< 1\%$
37	mid-utterance words *vs.* sing-song adjacency-pair start [38, 39]	$< 1\%$
62	explaining/excusing oneself *vs.* blaming someone/something	$< 1\%$
72	speaking awkwardly *vs.* speaking with a nicely cadenced delivery	$< 1\%$

Table 1: Brief descriptions of the interpretations of some of the top dimensions found in the Switchboard corpus, with the variance explained by each. Visualizations of all dimensions are at http://www.cs.utep.edu/nigel/dimensions/ .

investigated.

As noted in the introduction, some recent applications of prosody use raw features directly, without models, or at least without interpretable models. It would be nice to reverse this, to help reunify the scientific study of prosody and practical uses. PCA-derived elements, being computationally convenient yet also interpretable, may help. We already have found them useful for language modeling for speech recognition [10], for information retrieval [42, 25], for finding important information in dialog [43], and for characterizing the pragmatics of a non-lexical discourse particle [44]. Other potential applications include dialog-act inference, simultaneous interpretation, detecting emotion, detecting social roles, language identification, speaker recognition, realtime behavior prediction, dialog outcomes prediction, computer-assisted language learning, language proficiency evaluation, diagnosis of communication disorders, and speech synthesis.

1	this speaker talking *vs.* other speaker talking	$\sim$s1
2	low activity, low rapport *vs.* highly engaged	new
3	neither speaking *vs.* both speaking	$\sim$s2
4	grounding *vs.* grounded	$\sim$s4
5	turn grab *vs.* turn yield	$\sim$s5
6	topic continuation *vs.* topic change	$\sim$s3
7	slowly describing a difficult configuration *vs.* describing an easy path	new
8	meta-level *vs.* on-task	new
9	comfortable *vs.* awkward	new

Table 2: Interpretations of the top dimensions in the Maptask corpus. The last column notes correspondences to Switchboard-corpus dimensions.

6. Conclusions

Xu and Prom-on envisage models where "a full repertoire of communicative functions can be simultaneously realized in prosody, with all the details of the surface prosody still linked to their proper sources" [6]. PCA-derived prosodic elements can be part of such models, as they meet several important desiderata: 1. a fully explicit composition mechanism for combining elements, here simple addition, 2. groundedness of elements, whose presence at any point in any dataset can unambiguously determined, here by a simple linear combination of easily computable acoustic features, and 3. meaningfulness of elements, here with each bearing a specific communicative meaning or function.

This technique also has other advantages. It works not just for careful, professional speech and the phenomena therein, but for 'messy' unconstrained dialog. It covers elements at multiple 'levels' with a single mechanism. It's single mechanism covers not only pitch but also duration and volume (and potentially also voicing modes, gaze, gestures, etc.). Finally, it has been truly useful for discovery, and in this respect the results it gives, and the visualizations they support, are far clearer than those obtained by previous approaches [45, 46, 47, 35]; in essence this is because PCA is good at stripping out the variation involved in dimensions other than the single one being focused on.

Given the simplicity of the method, these results are surprisingly promising and the potential value seems great.

7. Acknowledgments

I thank Olac Fuentes for suggesting PCA, Alejandro Vega for the initial implementation and much more, Luis F. Ramirez for the initial Maptask analysis, Steven D. Werner for analyzing the distributions of values on each dimension, and the NSF for support, as project IIS-0914868.

8. References

[1] J. P. van Santen, T. Mishra, and E. Klabbers, "Estimating phrase curves in the general superpositional intonation model," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 61–66.

[2] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, pp. 85–115, 2011.

[3] H. Fujisaki, "Information, prosody, and modeling – with emphasis on tonal features of speech," in *Speech Prosody*, 2004.

[4] G. Kochanski and C. Shih, "Prosody modeling with soft templates," *Speech Communication*, vol. 39, pp. 311–352, 2003.

[5] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[6] Y. Xu and S. Prom-on, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[7] E. E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," in *Proceedings of the International Conference on Speech Prosody*, 2004, pp. 575–582.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[9] N. G. Ward and A. Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," in *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.

[10] ——, "Towards empirical dialog-state modeling and its use in language modeling," in *Interspeech*, 2012.

[11] M. Goudbeek and K. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *Journal of the Acoustical Society of America*, vol. 128, pp. 1322–1336, 2010.

[12] M. Charfuelan and M. Schröeder, "Investigating the prosody and voice quality of social signals in scenario meetings," in *Proc. Affective Computing and Intelligent Interaction*, 2011.

[13] H. R. Pfitzinger, "Segmental effects on the prosody of voice quality," in *Acoustics'08*, 2008, pp. 3159–3164.

[14] A. Batliner, J. Buckow, R. Huber, V. Warnke, E. Nöth, and H. Niemann, "Boiling down prosody for the classification of boundaries and accents in German and English," in *Eurospeech*, 2001, pp. 2781–2784.

[15] J.-P. Goldman, "Prosodyn: a graphical representation of macroprosody for phonostylistic ambiance change detection," *Proceedings of Speech Prosody*, 2012.

[16] S. Itahashi and K. Tanaka, "A method of classification among japanese dialects." in *EUROSPEECH*, 1993.

[17] Z.-H. Chen, Y.-F. Liao, and Y.-T. Juang, "Prosody modeling and eigen-prosody analysis for robust speaker recognition," *ICASSP*, 2005.

[18] C. M. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[19] D. Jurafsky, R. Ranganath, and D. McFarland, "Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates," *Computer Speech and Language*, vol. 27, pp. 89–115, 2013.

[20] M. Gubian, F. Cangemi, and L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," in *Speech Prosody*, 2010.

[21] M. Gubian, L. Boves, and F. Cangemi, "Joint analysis of f0 and speech rate with functional data analysis," in *ICASSP*, 2011, pp. 4972–4975.

[22] B. Parrell, S. Lee, and D. Byrd, "Evaluation of prosodic juncture strength using functional data analysis," *Journal of Phonetics*, vol. 41, no. 6, pp. 442–452, 2013.

[23] O. Jokisch, T. Langenberg, and G. Pinter, "Intonation-based classification of language proficiency using FDA," in *Speech Prosody*, 2014.

[24] I. T. Jolliffe, "Principal component analysis for time series and other non-independent data," in *Principal Component Analysis, 2nd ed.*, 2002, pp. 299–337.

[25] S. D. Werner and N. G. Ward, "Evaluating prosody-based similarity models for information retrieval," in *MediaEval Workshop*, 2013.

[26] J.-M. Marandin, "Contours as constructions," 2006, constructions SV1-10/2006.

[27] N. Sadat-Tehrani, "An intonational construction," *Constructions*, vol. 3, 2008.

[28] M. Liberman and I. Sag, "Prosodic form and discourse function," in *Papers from Tenth Regional Meeting, Chicago Linguistic Society*, 1974, pp. 402–427.

[29] A. Cutler, "The context-dependence of "intonational meanings"," in *Papers from the Thirteenth Regional Meeting, Chicago Linguistic Society*, 1977, pp. 104–115.

[30] N. Hedberg, J. M. Sosa, and L. Fadden, "The intonation of contradictions in American English," in *Prosody and Pragmatics Conference*, 2003.

[31] R. Ogden, "Prosodies in conversation," in *Understanding Prosody: The role of context, function, and communication*, O. Niebuhr, Ed. De Gruyter, 2012, pp. 201–217.

[32] H. H. Clark, *Using Language*. Cambridge University Press, 1996.

[33] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of Pragmatics*, vol. 32, pp. 1177–1207, 2000.

[34] L.-P. Morency, I. de Kok, and J. Gratch, "A probabilistic multimodal approach for predicting listener backchannels," *Autonomous Agents and Multi-Agent Systems*, vol. 20, pp. 70–84, 2010.

[35] N. G. Ward and J. L. McCartney, "Visualizations supporting the discovery of prosodic contours related to turn-taking," in *Interdisciplinary Workshop on Feedback Behaviors in Dialog*, 2012.

[36] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contourse in the interpretation of discourse," in *Intentions in Communication*, P. R. Cohen, J. L. Morgan, and M. E. Pollack, Eds. MIT Press, 1990, pp. 271–310.

[37] F. Lie, Y. Xu, S. Prom-on, and A. C. L. Yu, "Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling," *Journal of Speech Sciences*, vol. 3, pp. 85–140, 2013.

[38] D. R. Ladd Jr., "Stylized intonation," *Language*, pp. 517–540, 1978.

[39] J. Day-O'Connell, "Speech, song, and the minor third: An acoustic study of the stylized interjection," *Music Perception*, vol. 30, no. 5, pp. 441–462, 2013.

[40] S. Tilsen, "A dynamical model of hierarchical selection and coordination in speech planning," *PLOS ONE*, vol. 8, no. 4, 2013.

[41] C. Portes and C. Beyssade, "Is intonational meaning compositional?" *Verbum*, vol. XXXIV, 2014.

[42] N. G. Ward and S. D. Werner, "Using dialog-activity similarity for spoken information retrieval," in *Interspeech*, 2013.

[43] N. G. Ward and K. A. Richart-Ruiz, "Patterns of importance variation in spoken dialog," in *14th SigDial*, 2013.

[44] N. G. Ward, D. G. Novick, and A. Vega, "Where in dialog space does uh-huh occur?" in *Interdisciplinary Workshop on Feedback Behaviors in Dialog, at Interspeech 2012*, 2012.

[45] J. Edlund, M. Heldner, and A. Pelcé, "Prosodic features of very short utterances in dialogue," in *Nordic Prosody – Proceedings of the Xth Conference*, 2009, pp. 56–68.

[46] A. Rosenberg, "Classification of prosodic events using quantized contour modeling," in *HLT-NAACL 2010*, 2010, pp. 721–724.

[47] D. Neiberg, "Visualizing prosodic densities and contours: Forming one from many," *TMH-QPSR (KTH)*, vol. 51, pp. 57–60, 2011.

P-centre Position in Natural Two-Syllable Czech Words

Jan Volín, Eliška Churaňová, Pavel Šturm

Institute of Phonetics, Faculty of Arts, Charles University in Prague

jan.volin@ff.cuni.cz

Abstract

The ability to lock motor activity oscillator with external acoustic events is typical of various forms of human behaviour. Previous research showed that the beginning of an action is not necessarily the beginning of the rhythmic phase and led to the concept of p-centres. We present an experiment with 18 natural two-syllable Czech words spoken in synchrony with metronome beats by 18 subjects. Complexity of the consonantal onset and the type of coda together with distinctive phonological vowel length were carefully controlled to reveal a complex but comprehensible relationship between the word structure and phase locking.

Index Terms: Czech, p-centres, speech rhythm, syllable, synchronization

1. Introduction

Speech rhythm is a fascinating phenomenon attracting humans from their pre-linguistic babbling age through teenage years until their old age. Natural rhythmic flow apparently makes cerebral speech processing easier than uncommon or unpredictable patterns of prominence contrasts. Reaction times in monitoring experiments were by more than 150 milliseconds shorter in sentences with modal rhythm than in the same sentences with altered temporal structure [1], but cf. [2]. Certain types of rhythm in speech may contribute to speaker attractiveness or credibility of the speaker's propositions [3].

There is still considerable uncertainty, however, as to the descriptive principles that should be applied to speech rhythm [4]. One of the key missing pieces of the puzzle is the position of the "beat" – the moment of the perceptual emergence of an acoustic event in the mind of the listener. For syllables this moment was named a *p-centre* [5]. Several experiments of the early 1980s confirmed the intuitions of previous researchers about the importance of the vowel onset. It is not the beginning of the acoustic features pertinent to a syllable but the vicinity of its vocalic nucleus that starts the rhythmic phase.

The attempts to predict the position of the p-centre more accurately exploit mainly timing parameters [6,7], sometimes in combination with the information about energy of the signal [8,9,10]. Moreover, [11] suggested a link between the p-centre position and articulatory movements. Both acoustic descriptors and microbeam data explained substantial amount of variance also in [12], while [13] showed relevance of their kinematic data obtained with an articulometer.

However, Barbosa et al. [14] pointed out that mainly the Germanic languages had been investigated from this perspective. Cross-language comparison is therefore highly desirable. On the one hand, the fundamental principles of rhythmic structuring of speech are universal, since all languages are spoken in syllables, and, for instance, all healthy human ears detect the energy difference between /p/ and /a/ in the syllable /pa/. On the other hand, the rhythm type of an individual language predestines its users to specific strategies in speech perception, speech acquisition, and, naturally, speech production (e.g. a review in [15]). Not only the rhythm type (stress-based, syllable-based, mora-based), but also the phonotactic properties of individual phonemes in the language inventory might influence the exploitation of prominence alternations in speech. Our study is, therefore, aimed at mapping the behaviour of speakers of Czech – a west Slavonic language of central Europe.

Another issue addressed in our approach is the number of subjects. As, e.g., [10] noted, the rhythmic behaviour of human subjects is extremely disparate, and [4] rightly complains that the rhythmic proficiency of individual speakers who serve as subjects or respondents in experiments is often ignored. The pioneering study of [6] was based on four subjects, [7,9,10] on three speakers, [13] on six speakers (only three of whom provided kinematic data), while [14] studied data from one speaker only, and [8] hypothesized without empirical testing. Thus, one of our aims was to increase the number of subjects considerably and provide an estimate of the distribution of the rhythm aptitude in the population.

The third goal of our study concerned the speech material. Many of the previous experiments were carried out with meaningless syllables. This is advantageous in making the experiment neat and highly controlled, and desirable at initial stages of research. However, as [4] stresses, speech is not an object – it is a communicative behaviour. Also, de Jong noticed deviant reactions in his experiments to stimuli that departed from naturally sounding speech tokens [12]. On that account, we abandoned some of the comfort provided by nonsense syllables and used naturally occurring words.

In summary, our search for the position of the p-centre in a speech synchronization task was carried out with the following questions in mind:

- Will the results from a typologically underrepresented language in rhythm research – Czech – be consistent with the previous research? Specifically, do the structural and/or durational features of the stimuli affect the position of the p-centre? If yes, how?
- Given that p-centre position is dependent mainly on syllable onsets and less on codas [e.g., 9], will the rest of a two-syllable word matter if it is used instead of a monosyllable typical of p-centre research? (The rest of a word is actually beyond the stressed syllable coda.)
- Three pairs of our words do not differ structurally but in terms of phonetic identity of the onset consonants. A similar problem was studied by [9]. Do their results hold for a different language and testing conditions?
- Since [14] found dissimilar responses for different tempos, while other researchers seem to implicitly presume proportional changes in p-centre position, will two temporal modes natural to Czech speakers (4 and 6 syll/sec.) produce mutually coherent results?

We intend to answer these questions whilst expanding the number of participants beyond the numbers usual in the field.

2. Method

2.1. Target words

The design of the experiment required a reasonably restricted segmental structure of the targets, yet existing Czech words were favoured over nonsense words because of the reasons outlined above. The disyllabic targets contained the vowel pairs /e aː/ and /eː aː/. Other Czech vowels were not included due to their relative infrequency, occurrence only in foreign words or a salient qualitative difference. Moreover, the demand on using real words necessitated the selection of certain verb forms differing in their suffix, and many vowel combinations were therefore precluded. Each stem was used with the third-person singular zero suffix (open syllables), the first-person singular suffix *-m* (/m/), and the second-person singular suffix *-š* (/ʃ/). The intervocalic consonant was always a voiceless plosive (/t/ or /k/). The onset of the word contained either a single consonant (/l/ and /c/), a CC cluster (/sl/, /st/ and /ʃc/), or a CCC cluster (/spl/). Such a design yielded 18 combinations of word-initial syllable onsets and word-final syllable codas (6 different onsets × 3 different codas). The target words are recapitulated in Table 1.

onset	0-suffix	*m*-suffix	*š*-suffix
C	leːtaː	leːtaːm	leːtaːʃ
CC	sleːtaː	sleːtaːm	sleːtaːʃ
CCC	spleːtaː	spleːtaːm	spleːtaːʃ
CC	steːkaː	steːkaːm	steːkaːʃ
C	cekaː	cekaːm	cekaːʃ
CC	ʃcekaː	ʃcekaːm	ʃcekaːʃ

Table 1. *Summary of the 18 target words. Six lexical items (lines) and three grammatical forms (columns).*

2.2. Speakers and task procedure

18 native speakers of Czech participated in the experiment (14 female and 4 male, aged from 20 to 32). They were mostly students at Charles University in Prague, and reported no speech or hearing impediments. They received financial compensation for their involvement.

The speakers listened to a series of metronome beats (over headphones) and pronounced the target word presented on the computer screen several times. They were instructed to synchronize their repeated articulations with the isochronous sequence of beats so that each beat was aligned with the first syllable of the word spoken in isolation. A clear and "natural" pronunciation was demanded, as natural as it was possible in the experimental conditions (they were explicitly asked not to chant or recite the words). Each item contained 12 pulses followed by soft music and speakers began to articulate on the fifth – the initial four pulses served as a lead-in for steadying listeners' attention. The first and last realizations were omitted from analyses since they were considered prone to effects of uncertainty/accommodation at the beginning of the sequence or anticipation of the stimulus end.

As it is hypothesized that listeners use different processing modes depending on the pace [16], two tempos were tested: in the "normal" tempo the metronome pulse appeared at the rate of 70 bpm (i.e. every 857 ms), while in the "fast" tempo the pulse appeared at the rate of 90 bpm (i.e. every 667 ms). The two tempos were chosen to induce production rates of about 4 to 6 syllables per second (the pulse was associated with the disyllabic word, not with individual syllables).

The items were presented in two blocks separated with a one-minute break; the subjects chatted with the experimenter, could stretch their bodies and refresh themselves with a drink. The first block proceeded in the normal tempo and the second in the fast tempo, which resulted in two occurrences of the target items. In addition, filler items (mono- or trisyllabic words with no restriction on the segmental content) were used to provide variation to the task and prevent subjects from lapsing into monotonous behaviour. Each block was also preceded by several training items in which the subjects accustomed themselves to the tempo and task. Items within each block were pseudo-randomized for individual speakers. The duration of the session, comprising 55 items in total, was approximately 15 minutes.

2.3. Extraction of data

Given the purpose of this study two channels were used simultaneously to record the session, one for the metronome beat and one for the spoken production. The stereo recordings were subsequently processed in the programme *Praat* [17], and TextGrid objects were created for annotation. Individual items were segmented into words and phones using the *Prague Labeller* algorithm [18]. Segment boundaries were marked automatically and then manually checked and corrected where necessary. A script based on the detection of an intensity threshold in the metronome channel was used to determine the position of the recorded beats, which allowed us to identify the location of the metrical pulse for each produced word. The duration of each segment was measured, along with the distance of its boundaries from the metrical beat (*zero* = a segment boundary coincides with the beat; *positive numbers* = boundary located after the metronome beat; *negative numbers* = boundary located before the beat). In the subsequent analyses the term synchronization interval (SI) is used for the distance of the first (i.e., stressed) vowel onset and the beat.

3. Results

3.1. Phonological weight of the word

Since all the research in p-centres we have come across relates the position of the 'mental beat' to the onset of the first vowel, we will do the same for the sake of comparison. It should be also pointed out that Czech word stress is fixed to the first syllable – there are no words stressed on the second syllable.

Assuming that each consonant and short vowel weighs 1 phonological unit and long vowels weigh 2 units we could order the target words from the lightest (5 units) to the heaviest (8 units). However, knowing that individual syllable constituents affect the p-centre position with different power and having tied the weight of the intervocalic consonant and the second vowel, we ordered the overall results by the phonological weight of the first syllable and the type of coda of the second syllable. That seems to arrange the results in an internally congruent manner (see Table 2).

Apparently, the synchronization of the first vowel onset with the metronome beat changes with both the complexity of the first syllable and the absence or nature of the word coda. More complex word onsets pull the synchronization point more ahead of the vowel onset (into themselves, so to say.)

first syllable	word coda	word	mean SI (ms)	SI %
CV	∅	ceka:	-3.7	0.7
CV	Son		5.7	-12.7
CV	Obs		18.1	-34.2
CVV	∅	le:ta:	23.2	-14.0
CVV	Son		25.9	-18.4
CVV	Obs		44.4	-29.4
CCV	∅	ʃceka:	36.4	-57.9
CCV	Son		38.5	-60.0
CCV	Obs		72.8	-120.3
CCVV	∅	sle:ta:	65.3	-41.7
CCVV	Son	ste:ka:	64.7	-43.0
CCVV	Obs		57.5	-39.2
CCCVV	∅	sple:ta:	82.1	-57.1
CCCVV	Son		100.2	-71.7
CCCVV	Obs		94.9	-72.5

Table 2. *Mean synchronization intervals (SI) in milliseconds and percentages of vowel durations for individual phonotactic types (n = 36 apart from CCVV, where n = 2 × 36, i.e.,* sle:ta:/ste:ka: *).*

Long vowels seem to have an effect similar to heavier onsets: they prolong the distance between the first (i.e., stressed) vowel onset and the metronome beat.

Although there can be just one or no consonant word finally, this one coda consonant can still be either an obstruent or a sonorant. This distinction apparently matters – adding a sonorant coda (bilabial /m/ throughout our set) influences the synchronization less than adding an obstruent coda (postalevolar /ʃ/ throughout the set), at least in the case of simpler (phonologically lighter) words. Zero codas leave the vowel onset nearer the beat. Figure 1 displays the mean synchronization intervals (SI) as a function of the first syllable structure, Figure 2 as a function of the word coda.

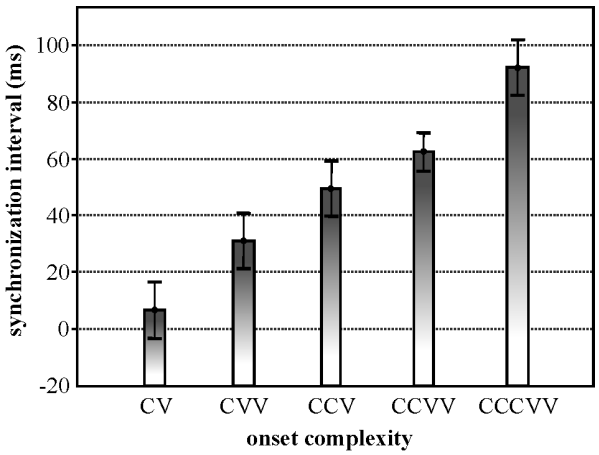


Figure 1. *Mean synchronization intervals for words with various phonological structure of their first syllable. The whiskers delimit the 95% conf. interval.*

Statistical significance of the differences was tested through two-way ANOVA for independent measures with synchronization interval as the dependent variable and ONSET and CODA as independent variables or factors. The main effects of both factors were found significant. For ONSET $F_{(4, 632)} = 43.6$, $p < 0.001$ and for CODA $F_{(2, 632)} = 5.4$, $p < 0.01$. Post-hoc Tukey HSD tests revealed significant differences among all onset complexities with the exception of the CCV which did not differ from CVV and CCVV. Zero codas differed from obstruent codas. No significant interaction was found between the two factors.

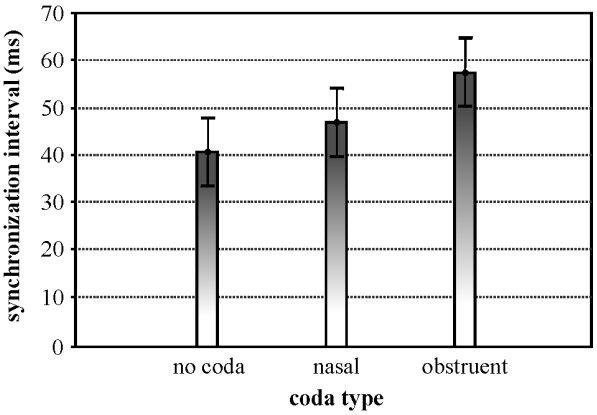


Figure 2. *Mean synchronization intervals for words with various final consonants (zero, nasal, obstruent). The whiskers delimit the 95% conf. interval.*

The situation described above disregards the tempo so it has to be ascertained if both slow and fast modes of synchronization display analogical trends. Inspection of the data showed that in the faster mode the speakers tended to prolong the synchronization interval, while in the slower one the metronome beat tended to be closer to the vowel onset. However, a closer look supported by a two-way ANOVA and post-hoc Tukey tests revealed that the speakers' behaviour was almost identical for simple words. For CV, CVV and CCV the difference between fast and slow tempos produced $p > 0.99$. For CCCVV it was still insignificant with $p \approx 0.52$, and only CCVV produced significant difference ($p < 0.05$). It has to be remembered, though, that the CCVV groups is twice as large as the other groups since it collapses words /ste:ka:/ and /sle:ta:/. (For comparison of these two words see below.) The variance is then lower and the effect is more visible. Clearly, if the effect is sensitive to number of cases in this way, it is not a strong effect. The influence of coda types seemed to be the same for both tempi.

3.2. Pairwise comparisons

Some of the items of our test were designed for direct comparison with others to answer less general questions. The first such comparison concerns the two CCVV words. It was actually carried out prior to the analyses above to make sure that the words /ste:ka:/ and /sle:ta:/ can be collapsed. Although the former has two obstruents in the onset, of which the second is a voiceless plosive, while the latter has got a sonorant sound in the same position, no difference was found in the synchronization interval: $F_{(5, 210)} = 0.44$, $p = 0.817$ (All six forms of these words were entered into the test.) Likewise, no difference was found in their overall duration or of the duration of their consonantal onsets. Internal durational ratios within the onset were different: /k/ was systematically longer than /l/ at the expense of /s/. In summary, although the energy distribution in the onset and the phonetic identities of its constituents differs, the synchronization interval does not.

The pair /ceka:/ × /ʃceka:/ can be observed in Table 2 and it is represented in Figure 1 by the first and the third column. The effect of the additional post-alveolar /ʃ/ is quite substantial. The synchronization moment moves by about 42.5 ms. The durational difference between the /c/ and /ʃc/ onset is about 80 ms, but this value is difficult to measure rigorously since /c/ is a voiceless plosive so the closure phase has to be estimated. Nevertheless, the mean overall duration of the word changes by 35 ms only.

Just as the two previous comparisons addressed the issue of consonantal onset, the following will focus on the vowel length in the first syllable. The relationship of the words /ʃceka:/ × /ste:ka:/ can be estimated from the third and fourth column in Figure 1. Their synchronization differs by about 10 ms which is not a statistically significant result. It is still informative to explore the durational differences in these two items. First of all, the mean duration of /e/ is by 93 ms shorter than that of /e:/, but the durations of the words only differ by 54 ms. It could be inferred that the missing 39 ms are due to the difference between the /st/ and /ʃc/ onsets. That, however is not the case: the difference between the alveolar and palato-alveolar cluster is only 12 ms. The missing time has to be accounted for by durational reorganization of the otherwise identical rest of the word: velar intervocalic consonant, long open vowel and codas. T-tests showed that it is the second vowel that is significantly different ($p < 0.01$). It seems to compensate for what is happening in the first syllable. Interestingly, this complex situation occurs in a pair of words that differ phonologically only in the length of the first vowel. The 10-ms difference in synchronization, which was not found significant actually means a 128-ms distance from the acoustic onset of the word /ʃceka:/ and a 106-ms distance from the acoustic onset of the word /ste:ka:/ suggesting that the first vowel onset is, indeed, a better synchronization point than the very beginning of the word.

Multiple regression analysis

In order to shed more light on the results reported in Table 2, a series of multiple regression analyses were performed with mean synchronization interval (first vowel boundary position) as the dependent variable and several temporal and phonotactic parameters as the predictors. When only TEMPO (slow × fast) and WORD DURATION were taken into account, the explanation power was negligible ($R^2_{adj} = 0.10$), partly because the two factors express similar facts. However, slightly better results were obtained for TEMPO and ONSET DURATION, assuming to reflect the phonotactic complexity of the onset, with $R^2_{adj} = 0.17$. Next, we extended the model with the variable of V1 DURATION, which further increased the amount of variance explained ($F_{(3,643} = 55.3$, $R^2_{adj} = 0.20$, $p < 0.001$). Interestingly, when a phonological category was used instead of a phonetic variable (vowel length instead of vowel duration), the power of the model did not decrease but, on the contrary, slightly increased: $F_{(3,643} = 60.2$, $R^2_{adj} = 0.22$, $p < 0.001$.

Adding information about the coda, whether in the form of CODA DURATION or CODA TYPE, did not make much difference. In both cases the model gained one more percent of explanatory power: $F_{(4,642} = 49.3$, $R^2_{adj} = 0.23$, $p < 0.001$.

If we wanted to use the regression analysis to create an analogy to Marcus' classic formula: $P = 0.65x + 0.25y + k$ (in [6]), where P is the synchronization interval, x is the onset duration and y is the rhyme duration (in the case of our first syllable only vowel), we would get $P = 0.37x + 0.20y - 27.2$. This analogy is not out of proportion, but it has to be remembered that it only explains 16 % of variance and that the methodology leading to it is not comparable with the Marcus' procedure. Be that as it may, the model with the information about coda (z in the following formula) and the whole word duration (w in the following formula) would have:

$$P = 0.61x + 0.50y + 0.27z - 0.29w + 73.6$$

4. Discussion

In a task of synchronizing words with an isochronous auditory sequence, the moment of the acoustic event can be considered the surface manifestation of the p-centre [14]. Our respondents clearly modified their behaviour according to the structure of the word and aligned the words differently for different consonant-vowel (CV) strings and different arrays of features attached to these segments. To answer the first question from the introduction then, we can state that our results based on Czech, typically classified as syllable-timed (but cf. [19]), resonate with previous research on stress-timed English and provide descriptors that can be further tested.

The current analyses are based on durational or structural characteristics only. They explain some of the variance in the data and show the trend for longer and more complex structures to push the p-centre farther ahead of the onset of the first vowel. Word-final segments exert smaller influence on the outcome. They display greater variation and little effect in regression analyses. Little, however, does not mean zero. The triplets of words in our set that differed solely in the presence or absence of the word-final consonant or its manner of production did not behave uniformly. Therefore, the answer to the second question from the introduction is positive as well.

The title of [9] stated that p-centre positions were unaffected by phonetic categorization. We supported that claim using different methodology by showing that our speakers' behaviour was identical for two (triplets of) words that differed in their two-consonant onset. In one of them it consisted of a sibilant + lateral approximant, in the other it was the same sibilant + voiceless plosive. The energy distribution was not equal, but the duration of the onset was. This seems to be the essential property of the item to condition the outcome. Our corroboration concerns the syllable onset only, though. It cannot be generalized for codas where we saw the difference between sonorant /m/ and post-alveolar /ʃ/. The issue will be further investigated in the nearest future as we have already prepared more stimuli for forthcoming testing.

The final question in the introduction reacted to findings in [14], where the respondent produced different patterns of behaviour for different tempi. Our results revealed no interactions between slow and fast mode of testing. That does not provide any evidence against [14] since our methods were incomparable, but rather suggests that for the pace of approximately 4 syllables per second and 6 syllables per second we can expect rather proportionate trends. Our linear regression found word duration as an expression of rate useful.

5. Acknowledgements

This work was supported by the Charles University Grant Agency (GAUK) under Grant 834213, and by the Charles University in Prague programme for science development P10-Linguistics.

6. References

[1] Buxton, H., "Temporal predictability in the perception of English speech", in A. Cutler and D. R. Ladd [Eds], Prosody: Models and Measurements, 111–121, Berlin: Springer-Verlag, 1983.

[2] Dilley, L. C. and Pitt, M. A., "Altering context speech rate can cause words to appear or disappear", *Psychological Science*, 21: 1664–1670, 2010.

[3] Cross, I., "Rhythms of persuasion: The perception of periodicity in oratory", in *Proceedings of Perspectives on Rhythm and Timing*, 27, Glasgow, 2012.

[4] Kohler, K. J., "Rhythm in speech and language: a new research paradigm", *Phonetica*, 66: 29–46, 2009.

[5] Morton, J., Marcus, S., & Frankish, C., "Perceptual centres (P-centres)". *Psychological Review*, 83: 405-408, 1976

[6] Marcus, S., "Acoustic determinants of perceptual center (p-center) location", *Perception & Psychophysics*, 30(3): 247–256, 1981.

[7] Fox, R., and Lehiste, I., "The effect of vowel quality variations on stress-beat location", *Journal of Phonetics*, 15: 1-13, 1987.

[8] Howell, P., "An acoustic determinant of perceived and produced anisochrony", in *Proceedings of the 10th ICPhS*, 429–433, Dordrecht, 1984.

[9] Cooper, A.M., Whalen, D.H., Fowler, C., "P-Centers are unaffected by phonetic categorization", *Perception & Psychophysics*, 39, 187-196, 1986.

[10] Pompino-Marschall, B., "On the psychoacoustic nature of the P-centre phenomenon", *Journal of Phonetics*, 17: 175–192, 1989.

[11] Fowler, C. A., Whalen, D. H. and Cooper, A. M., "Perceived timing is produced timing: A reply to Howell", *Perception & Psychophysics*, 43: 94–98, 1988.

[12] de Jong, K., "Acoustic and articulatory correlates of p-centre perception", *UCLA Working Papers in Phonetics*, 81: 66-75, 1992

[13] Patel, A., Löfquist, A. and Naito, W., "The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-Centre problem", in *Proceedings of 14th ICPhS*, 1: 405–408, San Francisco, 1999.

[14] Barbosa, P. A., Arantes, P., Meireles, A. R., Vieira, J. M., "Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors", in *Proc. of 9th European Conf. on Speech Communication and Technology (Interspeech 2005)*, 1441–1444, Lisbon, 2005.

[15] Cutler, A. and Otake T., "Rhythmic categories in spoken-word recognition", *Journal of Memory and Language*, 46: 296–322, 2002.

[16] Kohno, M., "Two different systems for rhythm processing and their hierarchical relation", in *Proceedings of the 13th ICPhS*, 1: 94-97, Stockholm, 1995.

[17] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer (Version 5.2.37)", 2011, accessed on September 3, 2011 from http://www.praat.org/.

[18] Pollák, P., Volín, J. and Skarnitzl, R.: "HMM-Based Phonetic Segmentation in Praat Environment", in *Proceedings of the XIIth International Conference Speech and Computer – SPECOM 2007*, 537–541, Moscow, 2007.

[19] Dankovičová, J. and Dellwo, V.: "Czech speech rhythm and the rhythm class hypothesis", in *Proceedings of the 16th ICPhS*: 1241-1244, Saarbrücken, 2007.

Correlations between prosody and epistemic bias in negative polar interrogatives in Japanese

Hyun Kyung Hwang[1], *Satoshi Ito*[2]

[1] Department of Linguistic Theory and Structure, National Institute for Japanese Language and Linguistics, Tokyo, Japan
[2] Department of Linguistics, Cornell University, USA
hwang@ninjal.ac.jp, si57@cornell.edu

Abstract

This study investigates correlations between prosodic patterns and speaker's bias observed in Japanese negative polar interrogatives, with special attention given to the perceptual and functional aspects of the correlation. The result of a naturalness rating test and a comprehension test demonstrate that listeners perceive the matching context-prosody pairs to be more natural, compared to the conflicting pairs. The results indicate that the prosodic patterns successfully guide listeners to identify the speaker's bias in negative polar interrogatives.

Index Terms: epistemic bias, negative polar interrogatives, comprehension test, naturalness rating test

1. Introduction

1.1. Speaker's bias in negative polar interrogatives

It has been widely recognized that English negative polar interrogatives often, though not always, convey the speaker's bias toward either a positive or negative answer [1, 2, 3, 4, 5].

(1) (negative interrogative with a positive epistemic bias)
A: John is such a linguist.
B: Yeah, doesn't he even speak Japanese?

(2) (negative interrogative with a negative epistemic bias)
A: There is nothing John can help with here.
B: Doesn't he even know how to keep accounts?

The speaker B in (1) expects confirmation for the proposition "John speaks Japanese", whereas the speaker B in (2) requests confirmation for "John does NOT know how to keep accounts". Compare these with (3), which is neutral in terms of epistemic bias.

(3) (negative interrogative without an epistemic bias)
Context: A and B are making a list of teetotalers for a party.
A: Jane and Mary do not drink.
B: OK. What about John? Does he not drink (either)?
(Examples from [4])

A similar distinction in speaker's bias has also been reported in Japanese [6, 7]. In particular, Ito and Oshima [6] points out that Japanese negative polar interrogatives differentiate their intonation patterns depending on the speaker's bias.

1.2. Prosodic patterns of negative polar interrogatives in Japanese

Negative polar interrogatives in Japanese contain the form of ... *X-nái?*, where *–nái* is the negative morpheme. Two prosodic patterns are pervasively used in this construction; (i) both X and the negative morpheme retain their lexical accents (AA), (ii) only X retains its accent, and the negative morpheme is deaccented (AD). While both the adjective (*nagái* 'long') and the negative morpheme exhibit an F0 fall in the top contour of Figure 1, the negation in the bottom contour lacks an F0 fall.

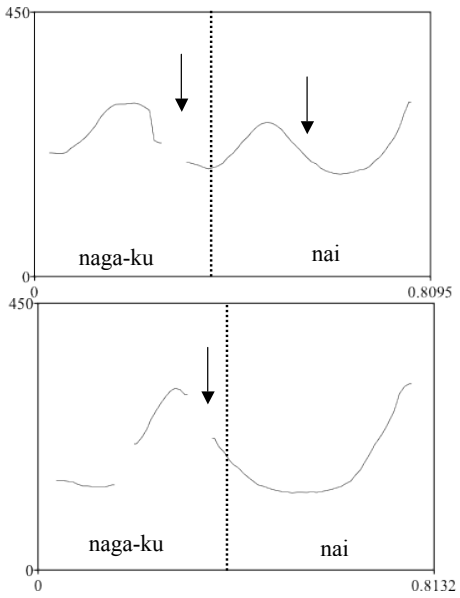


Figure 1: *AA (top) and AD (bottom) pattern of negative polar interrogatives in Japanese; arrows indicate the F0 fall.*

It has been claimed that the AA pattern is associated with a neutral/negative epistemic bias, and the AD pattern with a positive bias [6]. However as yet no experimental/quantitative data was provided to support their claim.

Beside the two patterns above, one more prosodic pattern with adjectival predicates is observed among younger speakers: (iii) both X and the negative morpheme are deaccented (DD). The pitch contour of the DD pattern is demonstrated in Figure 2. It should be noted that no F0 fall is observed either on the adjective or the negative morpheme, and the entire phrase displays a single gradually rising pitch contour.

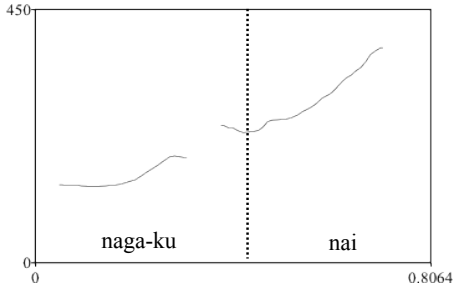


Figure 2: *DD pattern of negative polar interrogatives in Japanese.*

Hara and Kawahara [8] compare the AD and DD patterns in terms of evidentiality. Based on the result of a naturalness rating test, they argue that the DD pattern is felicitous only when the participants of conversation have public evidence which is stronger than hearsay or circumstantial evidence, for the positive answer. However the AA pattern is not discussed in their study.

Thus, experimental investigation is needed to uncover the correlation between the epistemic bias and all three prosodic patterns. In the current experimental study, we aim to test the perceptual and functional aspects of the correlation, involving all three prosodic patterns. Also, we reconsider the difference between the AD and the DD patterns argued in [8]. The results of a naturalness rating test and a comprehension test confirm the correlation between the prosody and the speaker's bias. The experiment also highlights the fact that prosodic patterns play a crucial role in identifying the epistemic bias of negative polar interrogatives in Japanese.

2. Methods

2.1. Material

The tested phrase was *nagáku-nái*? ('isn't it long?'). Only an adjective as a predicate was chosen, as adjectives allow us to test all the prosodic patterns. A major reason for this is that the DD pattern is regularly realized with adjectival predicates, but unavailable or highly marked with non-adjectival predicates. In order to provide a contextual prompt for the speaker's bias, the target phrase is embedded in three different bias conditions: neutral, negative and positive bias. The positive bias condition is further divided into two subgroups depending on the presence or absence of public evidence.

Concerning the prosodic pattern, both neutral and negative biases are expected to exhibit the AA pattern. On the other hand, the positive bias condition can be realized in either the AD or the DD pattern. The bias conditions together with the expected prosodic pattern are summarized in Table 1.

Table 1. *Tested bias conditions and expected prosodic patterns.*

Bias	**Neutral**	**Negative**	**Positive**	
Public Evidence			No	Yes
Prosodic pattern	AA		AD	DD

The contexts provided to the subjects for each bias condition are given below.

(4) Neutral
(Situation) One morning, Taro, a high school student, was talking with another high school student named Hanako (H), who had just transferred from another school that day. Taro (T) and Hanako were speaking just before morning assembly. Hanako needed to go to the bathroom but since it was her first day, she did not know how long morning assembly would last. Therefore, she asked Taro:

H: **Morning assembly isn't that long**, is it?
T: No, it's not that long. Why do you ask?
H: If it's very long, I think I may need to go to the restroom first.

(5) Negative
(Situation) Taro wanted to use his PC but his power cord was too short, so he asked Hanako if she had an extension cord. Hanako handed Taro her extension cord (but the cord was not long enough).

T: Hmmm... Actually, I think this might not be long enough…
H: Oh really? **It isn't long** (enough)?
T: Yeah, it seems like it's not quite long enough…

(6) Positive bias without public evidence
(Situation) Taro and Hanako were in the school yard listening to the principal's speech. Hanako felt like the speech was a little long that day, but she didn't have a watch, so she asked Taro (who was standing in front of her).

H: **Isn't today's morning assembly (a little) long**?
T: Yeah, it's a bit longer (than usual) today.

(7) Positive bias with public evidence
(Situation) When Taro and Hanako were in the school yard listening to the principal's speech, they heard the bell for first period ring. When she heard the bell ring, Hanako asked to Taro (who was standing in front of her):

H: **Isn't today's morning assembly (a little) long**?
T: Yeah. That was the bell for first period, right? But that means first period will be shorter, so that's nice.

Note that the situation in (7) explicitly describes that the bell for the first period ring, indicating the assembly has run overtime.

2.2. Recording

One female and one male speakers of Standard Japanese were recorded. Both speakers were in their late twenties at the time of the recording. The recording was made in a sound attenuated booth at National Institute for Japanese Language and Linguistics. They were instructed to read the situations written in Japanese orthography carefully. In addition, pictures which visualized the situations were provided to ascertain their understanding of the situations. After self-reporting that they fully understood the contexts, speakers were asked to exchange the conversations as naturally as possible. The recording was repeated five times. The female speaker uttered the target phrase, and it is worth noting that this particular speaker exhibited alternation between the AD and the DD pattern regardless of public evidence in the positive condition.

2.3. Stimuli

Stimuli for a comprehension test and a naturalness rating test were created using the utterances obtained in the recording session. For the comprehension test, one rendition was chosen from the five repetitions, which was judged the most appropriate in terms of pronunciation, speed, and intensity. As the speaker showed alternation between the two prosodic patterns for the positive bias condition, we included both cases, which resulted in a total of six stimuli: Neutral-AA, Negative-AA, Positive with/without public evidence-AD/DD. For all the stimuli, the male speaker's answer for the target phrase was deleted, as it was to be identified by participants in the experimental task.

For a naturalness rating test, the target phrase *i.e.* negative polar interrogative was cross-spliced into the conversations of the four different bias conditions: Neutral, Negative, Positive with Public Evidence, and Positive without Public Evidence. As three possible prosodic patterns were observed, a total of twelve combinations (4 bias condition X 3 prosodic patterns) were created.

2.4. Participants and procedure

A total of thirty native speakers of Japanese ranged 18-35 years old took part in the tests. They were all born and grew up in or around Tokyo area. Both tests were conducted in a quiet office, and the stimuli were presented over a headphone with each situation and response choices on a computer screen. All conditions were randomly interspersed.

Specifically, for the comprehension test, participants were asked to read the situations given on a computer screen carefully. Then, they were informed that they would hear a short conversation between a male and a female speakers. The task was to choose the most appropriate answer for the female speaker's question in the situation given. They were asked to click on one of four boxes containing the four choices below.

(6) Yes, it is long. Yes, it isn't long.
No, it is long. No, it isn't long.

For a naturalness rating test, participants were instructed to read the situations carefully, and to rate the naturalness of the prosodic pattern of the female speaker's question on a 1-5 scale, taking into account the situation. For the naturalness judgments, the numbers were labeled as following: 1 "highly unnatural", 2 " somewhat unnatural", 3 "neither unnatural nor natural", 4 "somewhat natural", 5 "highly natural".

3. Results and discussion

3.1. Comprehension of speaker's epistemic bias

The results of the comprehension test reveal that the speaker's epistemic bias and prosodic patterns are highly correlated. (In Table 2, P.E. stands for Public Evidence.)

Table 2. *Percentages of correct responses depending on the epistemic bias and prosodic patterns.*

Bias	Neutral	Negative	Positive			
P.E.			No		Yes	
Prosody	AA		AD	DD	AD	DD
%Correct	94	100	97	100	97	97

Participants yielded extremely high accuracy in interpreting the speaker's bias, confirming the intuitive claim in [6]. The relatively low percentages-97% or 94% indicates only one or two unexpected responses, which could be attributed to a mistake at the performance level.

Surprisingly, there was no substantial difference between the two Positive conditions with/without Public Evidence. It should be reiterated that, according to [8], the Public Evidence condition with the DD pattern or No Public Evidence condition with the AD pattern is expected to exhibit greater accuracy compared the other two conditions. Besides the prosodic alternation observed in the female speaker's production, this result also suggests that evidentiality is not a determining factor for the correlation between speaker's bias and prosody in negative polar interrogatives.

3.2. Naturalness judgments

There are appreciable differences in perceived naturalness between the prosodic patterns, as illustrated in Figure 3.

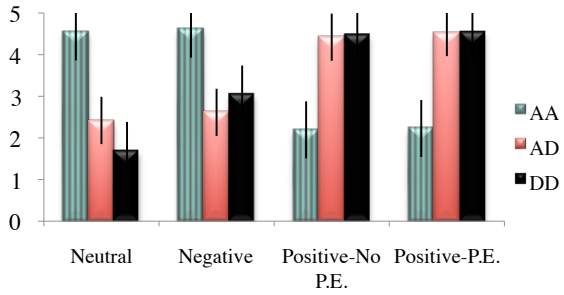


Figure 3: *Mean naturalness ratings depending on the bias and Public Evidence (P.E.) conditions.*

Overall, the AA pattern realized in the Neutral/Negative conditions (striped bars) were judged as highly natural. Likewise, similarly high naturalness ratings were assigned to the AD (bright colored bars) and DD patterns (dark colored bars) in the Positive condition. This result again supports the correlation claimed in [6], and corroborates the finding of the comprehension test.

Considering evidentiality in the Positive condition, both the AD and the DD patterns received equally high ratings. On the other hand, the AA pattern is judged as considerably less natural, regardless of the presence or absence of Public Evidence.

In order to test the statistical significance of the differences, one-way ANOVAs using a generalized linear model were performed using JMP 9. All reported effects were significant at the $p<0.05$ level. The dependent variable considered was the naturalness ratings. The independent factor was prosodic patterns in each bias condition. One-way ANOVAs show that there is a significant effect of the prosodic pattern on the rating across the bias conditions, as summarized in Table 3. For the Neutral condition, post-hoc comparisons using Tukey-Kramer HSD reveal that the AA pattern is perceived as significantly more natural than the other patterns. Interestingly, marginally significant differences are yielded between AD and DD (AD > DD, $p=0.03$). In relation to this, [6] points out that the AD pattern can be used in some other ways besides its "positive epistemic bias" use. The differences between AD and DD might stem from one of these usages,

which is called the "information gap" situation in [6], where the AD pattern has a distinct use where it does not convey an epistemic bias but indicates that the speaker considers the core proposition **P** possible based on some information that may not be available to the hearer. Thus, for example, for the negative polar interrogative with positive epistemic bias such as (1B), **P** will be roughly: λw.[speak⟨John, Japanese, w⟩]. The effect of using AD pattern in such a situation is similar to adding a phrase like: "You may be surprised by my asking this, but (is P the case?)". In this context it is predictable that the AA pattern is acceptable for most participants, as [6] anticipates. In addition, some participants noticed the "information gap" usage of the AD pattern in the neutral context in (4) and judged the pattern is also natural in context (4). That is, they felt that it is also natural to ask "You may be surprised by my asking, but is that morning assembly is long here the case?" by using the AD pattern in context (4). Interpreted this way, Table 3 suggests an interesting consequence: the DD pattern cannot be used for "information gap" situation while the AD pattern can.

Table 3. *Results of statistical analyses.*

df (2, 92)	Ratings	Tukey-Kramer HSD
Neutral	F=27.76 P<.0001*	**AA** > AD > DD
Negative	F=57.65 P<.0001*	**AA** > DD, AD
Positive-No P.E.	F=44.71 P<.0001*	**DD, AD** > AA
Positive-P.E.	F=47.88 P<.0001*	**DD, AD** > AA

Turning to the Negative bias condition, the AA pattern is rated significantly more natural than the AD or the DD pattern, confirming the correlation.

Finally, in the Positive bias condition, the subgroups depending on Public Evidence pattern together with respect to the naturalness rating; significantly higher ratings are assigned to the AD and DD patterns, compared to the AA pattern.

This result is in accordance with the finding in the comprehension test, but contrary to the results obtained in [8]. It is not clear how to account for the discrepancies between the result in [8] and the current finding. It is conceivable that differences in the stimuli played a role. Unlike the current study, only a positive condition was tested in [8], which could influence their participants to focus on the existence of public evidence, and to try to differentiate the two cases.

4. Conclusions

This paper discussed the correlation between prosodic patterns with adjectival predicates and the speaker's bias observed in Japanese negative polar interrogatives through two types of experiment. The result obtained from the comprehension test demonstrates that the distinct prosodic patterns are exploited to comprehend the epistemic bias and to appropriately respond to negative polar interrogatives. The result of the naturalness test reveals that Japanese speakers are sensitive to the prosodic patterns observed in negative polar interrogatives, and the prosodic differences play an important role in naturalness judgments. Taken together, the results reported in the current study confirm that prosodic patterns are highly correlated with speaker's bias in negative polar interrogatives in Japanese. Further, it was shown that the presence or absence of public evidence does not differentiate the AD and the DD patterns.

Future research concerning other parts of speech, in addition to adjectives, needs to be explored. With respect to the difference between the AD and the DD patterns, the current research suggests the possibility that the AD pattern might be able to be used for "information gap" situation while the DD pattern might not. More tests will be necessary to clarify this point. Also, further investigation of the fine-grained prosodic correlates of speaker's bias, and consideration of relevant constructions with greater complexity are necessary to better understand the nature of the interface between prosody and speaker's bias.

5. Acknowledgements

We are grateful to John Whitman for helpful comments and discussions. We would also like to thank all the participants of our experiments.

6. References

[1] Ladd, D. R., "A first look at the semantics and pragmatics of negative questions and tag questions", Chicago Linguistics Society 17, 164-171, 1981.

[2] Büring, D. and Gunlogson, C., "Aren't positive and negative questions the same?", manuscript, University of California Santa Cruz, 2000.

[3] Huddleston, Rodney and Pullum, G. K., "The Cambridge Grammar of the English Language", Cambridge University Press, 2002.

[4] Romero, M. and Han, C-H., "On negative *yes/no* questions", Linguistics and Philosophy 27(5):609-658, 2004.

[5] Asher, N. and Reese, B., "Intonation and discourse: Biased questions", in S. Ishihara, S. Jannedy, and A. Schwar [eds], Interdisciplinary Studies on Information Structure, vol. 8, 1-38, The University of Potsdam, 2007.

[6] Ito, S. and Oshima, D. Y., "On two varieties of negative polar interrogatives in Japanese", in M. Kenstowicz, T. Levin and R. Masuda [eds], Japanese/Korean Linguistics 23, CSLI Publications, to appear.

[7] Sudo, Y., "Biased polar questions in English and Japanese", in D. Gutzmann and H-M. Gaertner [eds], Beyond Expressives: Explorations in Use-Conditional Meaning, Brill, 2013.

[8] Hara, Y. and Kawahara S., "The prosody of public evidence in Japanese: A rating study", in J. Choi, E. A. Hogue, J. Punske, D. Tat, J. Schertz, and A. Trueman [eds], Proceedings of the 29th West Coast Conference on Formal Linguistics, 353-361, Cascadilla Proceedings Project, 2012.

L2 production of Estonian quantity degrees

Einar Meister and Lya Meister

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics at Tallinn University of Technology, Estonia
einar@ioc.ee, lya@phon.ioc.ee

Abstract

The Estonian quantity system involves three contrastive patterns referred to as short (Q1), long (Q2) and overlong (Q3) quantity degrees. Our previous studies have shown that for L2 learners the distinction between Q2 and Q3 is a difficult task in both production and perception. While Q1 and Q2 structures are always distinguished in the orthography, this is not the case in most Q2 and Q3 words excluding the words with plosives between first and second syllable vowels (see examples later in the text). Thus, the orthography might be the reason for the use of the same L2 production pattern for both Q2 and Q3.

The current paper studies the role of L2 orthographic input on the L2 production of Estonian quantity degrees by two groups of subjects with different language backgrounds: Finnish and Russian. The material used in the study involves word structures with and without orthographic manifestation of quantity contrasts.

The results confirm the role of Estonian orthography on the L2 pronunciation, however, the two L2 subject groups show different prosodic patterns.

Index Terms: L2 speech, Estonian, Finnish, Russian, quantity opposition

1. Introduction

The domain of the Estonian quantity degrees is a disyllabic foot – a sequence consisting of the stressed syllable and the following unstressed syllable, e.g. [1], [2], [3], [4]. In the vowel-peaked structures (Q1: CV.CV; Q2: CVV.CV; Q3: CVV:.CV) the quantity contrast manifests mainly in vowels and diphthongs of the stressed syllable, e.g.: Q1: *sada* /sa.ta/ 'hundered', nom.sg.; Q2: *saada* /saa.ta/ 'to send', sg.imperat.; Q3: *saada* /saaː.ta/ 'to get'; Q2: *koera* /koe.ra/ 'dog', gen.sg.; Q3: *koera* /koeː.ra/ 'dog', part.sg.. In the consonant-peaked structures (Q1: CV.CV, Q2: CVC.CV, Q3: CVC:.CV) quantity oppositions occur in consonants and consonant clusters between the first and second syllable vowels, e.g.: Q1: *kala* /ka.la/ 'fish', nom.sg.; Q2: *kalla* /kal.la/ 'arum', nom.sg.; Q3: *kalla* /kalː.la/ 'pour', 2.sg. imperat.; Q2: *lehma* /leh.ma/ 'cow', gen.sg.; Q3: *lehma* /lehː.ma/ 'cow', part.sg..

Phonetically, the duration of a stressed syllable vowel (V1) in the vowel-peaked structures is shortest in Q1 and longest in Q3 – V1 duration in Q2 is ca 1.9 and in Q3 ca 2.5 times longer than in Q1 (as pooled from several studies [5]); the duration of intervocalic consonant (C2) in the consonant-peaked structures increases in a similar amount – ca 1.8 and ca 2.5 times in Q2 and Q3, receptively [6]. The duration of an unstressed syllable vowel (V2) varies inversely to the duration of V1 (or C2) being in Q2 ca 0.8 and in Q3 ca 0.6 times shorter than V2 in Q1 [5], [6]. Despite large variations of V2 duration, no quantity contrast exists in unstressed syllable and V2 is defined phonologically "short". Also, a consonant quantity contrast in word-initial position is not possible.

In addition to the inversely proportional durational relations within a foot, F0 contour is a complementary cue distinguishing the quantity oppositions – in Q1 and Q2 the F0-peak is located close to the end of V1, in Q3 it is located within the first half of V1 eg. [7], [8].

Lehiste [7] has introduced the syllable duration ratio as a characteristic feature distinguishing the three quantity degrees. The typical duration ratio for Q1 is 2:3, for Q2 3:2, and for Q3 2:1; similar ratios have been reported in numerous subsequent studies for both read [2], [8], [9], [10], [11] and spontaneous speech [12], [13].

As our previous studies [6], [14], [15], [16], [17] have shown, for L2 learners with Finnish- and Russian-language backgrounds the distinction between vowel-peaked Q2 and Q3 contrasts is a difficult task in both production and perception. It can be partly explained by the fact that in the orthography, vowel-peaked Q2 and Q3 structures, representing different grammatical words, are not distinguished (see examples above).

In our recent paper [5], we have addressed the role of L2 orthography on the production of Estonian quantities by focusing on consonant-peaked target words manifesting the Q1 – Q2 – Q3 contrast orthographically, i.e. the words with plosives between first and second syllable vowels, e.g.: Q1: *kade* /ka.te/ 'envious', nom.sg.; Q2: *kate* /kat.te/ 'cover', nom.sg.; Q3: *katte* /katː.te/ 'cover', gen.sg.; Q1: *lugu* /lu.ku/ 'story', nom.sg.; Q2: *luku* /luk.ku/ 'lock', gen.sg.; Q3: *lukku* /lukː.ku/ 'lock', part.sg.; Q1: *leba* /le.pa/ 'lay', 2.pers.sg.imperat.; Q2: *lepa* /lep.pa/ 'alder', gen.sg.; Q3: *leppa* /lepː.pa/ 'alder', part.sg.

Notice that, in Estonian orthography, the letters <bdg> and <ptk> denote short and long voiceless plosives, respectively.

The findings confirmed the effect of L2 orthography – L2 subjects with Finnish-language background produced different patterns for Q2 and Q3 structures in the case of target words with plosives, but not in the case of words with non-plosives.

In the current paper we study the role of L2 orthographic input on the L2 production of Estonian quantity oppositions by two groups of subjects with different language backgrounds: Finnish and Russian. The speech material analyzed in the study involves vowel-peaked and consonant-peaked words with and without orthographic manifestation of quantity contrasts.

2. Method

2.1. Subjects

The L2-FI group involves subjects (age 21-49, median 36) from the Helsinki, Turku and Oulu areas born in monolingual Finnish speaking families. They started to study Estonian at the age 18-35 at university and have studied it for 1-5 years; six subjects

use Estonian frequently (daily or weekly), other six rarely.

The L2-RU subjects (age 21-33, median 24.5) were born in monolingual Russian speaking families living in the north-east of Estonia or in the capital area. Most of the L2 subjects started to learn Estonian in school at the age of 6-13, one subject at the age of 16 and one at the age of 20. All (except one) L2-RU subjects communicate at home in Russian, but outside they use Estonian almost every day.

The L1-EE group of native speakers (age 21-54, median 26.5) came from monolingual Estonian-speaking families and represent the pronunciation of standard Estonian.

All subject groups have equal number of subjects balanced by sex (6 male, 6 female); none of the subjects reported any language impairment.

2.2. Speech material

A subset of the Estonian Foreign Accent Corpus [18], [19] involving disyllabic target words representing quantity oppositions in sentence context was used in the study. L2-FI subjects from Helsinki and Oulu were recorded in the recording studios of Helsinki and Oulu universities, the subjects from Turku in a quiet lecture room at Turku University. L2-RU and L1-EE subjects were recorded in the recording studio of the Laboratory of Phonetics and Speech Technology, Tallinn University of Technology. For all recordings the same microphones and high quality recording equipment (sampling frequency 44.1 kHz, resolution 16 bit) were used.

From each subject 48 words were analyzed, including nine triplets of vowel-peaked structures (CV.CV; CVV.CV; CVV:.CV) (referred later as Vowel set) and seven triplets of consonant-peaked structures (CV.CV, CVC.CV, CVC:.CV). Among the latter, three triplets (referred to as Plosive set) involved plosives /k/, /p/ and /t/ between first and second syllable vowels, in four triplets (referred to as Non-plosive set) the consonants /m/, /n/, /s/, and /l/ were present. All target words have been manually segmented and labeled on word and phone levels using Praat [20].

2.3. Measurements

The durations of all constituent segments (C1, V1, C2, V2) in each target word were measured using a Praat-script and the syllable duration ratio was calculated as the duration of the first (stressed) syllable rhyme divided by the duration of the second (unstressed) syllable nucleus [21].

In the case of vowel-peaked quantity contrast (i.e. target words CV.CV, CVV.CV and CVV:.CV) the characteristic duration ratio is calculated as the duration of the stressed syllable vowel (V1) divided by the duration of the unstressed syllable vowel (V2). In the case of consonant-peaked target words (CVC.CV and CVC:.CV) the calculation of duration ratio is more complicated since it involves splitting of the word-medial geminate consonant into two segments: the coda of the first syllable and the onset of the second syllable. For splitting a simple approach from [21] has been adopted. In Q2 (CVC.CV) the intervocalic geminate is divided into two parts of equal duration, in Q3 (CVC:.CV) the second syllable onset is taken equal to one-third of the duration of the intervocalic geminate, and two-thirds of geminate's duration is attributed to the first syllable coda.

Notice that syllable-initial consonants do not participate in forming quantity contrasts, thus they are left out of further analysis. However, durations of C1 and C2 are given in Table 1 and 2 for the reader's interest.

3. Results

3.1. Vowel-peaked structures

Table 1 provides mean segment durations and syllable duration ratios of the Vowel set (36 subjects x 27 words = 972 words) representing the vowel-peaked word structures in Q1, Q2, and Q3 produced by subjects in three groups. ANOVA with factors Subject group and Quantity, and TukeyHSD post-hoc test were applied for statistical analysis. Box plots (Figure 1) and scatter plots (Figure 2) demonstrate variations in V1/V2 ratio, and V1 and V2 duration produced by the three subject groups.

As expected, the segment durations and syllable duration ratios of the L1-EE subject group are in line with those reported in many earlier studies (e.g. [2], [4], [5], [13], and others). Quantity effects durations of V1 [$F(2, 321) = 350.5$; $p < 0.001$] and V2 [$F(2, 321) = 131.6$; $p < 0.001$], and consequently also V1/V2 ratios [$F(2, 321) = 503$; $p < 0.001$], among different quantity degrees.

In the L2-FI group, Quantity has a strong effect on V1 [$F(2, 321) = 210$; $p < 0.001$] and V2 [$F(2, 321) = 92$; $p < 0.001$] duration, and V1/V2 ratio [$F(2, 321) = 248.5$; $p < 0.001$]. However, the post-hoc test confirmed differences in these parameters among Q1 and Q2 ($p < 0.001$), but not between Q2 and Q3. Comparing L2-FI vocalic segments to those of the L1-EE group, there are no significant differences in the case of Q1 and Q3, but large differences exist between the two groups in the case of Q2 in V1, V2, and in V1/V2 ratio ($p < 0.001$). L2-FI subjects do not distinguish the Q2 and Q3 temporal patterns and produce them both similarly to the Q3 pattern of the native group. Notice the almost perfect match of Q2 and Q3 in L2-FI group in Figure 1 and 2 (mid).

Table 1: *Mean duration (in ms) and standard deviation (in parenthesis) of C1, V1, C2, V2, and V1/V2 duration ratio in the three word structures representing the quantity contrasts Q1, Q2 and Q3 read by L1-EE, L2-FI and L2-RU subjects.*

Group	Qs	C1	V1	C2	V2	V1/V2
L1-EE	Q1	**81**	**81**	**61**	**111**	**0.8**
		(20.1)	(15)	(15.7)	(26.9)	(0.2)
	Q2	**85**	**143**	**58**	**85**	**1.8**
		(19.4)	(28.2)	(10.8)	(23.8)	(0.4)
	Q3	**78**	**171**	**59**	**63**	**2.8**
		(21.5)	(30.4)	(11.9)	(12.9)	(0.7)
L2-FI	Q1	**94**	**74**	**71**	**114**	**0.7**
		(25.2)	(22)	(24.9)	(35.9)	(0.2)
	Q2	**106**	**170**	**60**	**66**	**2.8**
		(25.9)	(48.2)	(17.5)	(25.7)	(1)
	Q3	**96**	**173**	**60**	**66**	**2.8**
		(27.4)	(44.8)	(15.7)	(27.7)	(0.9)
L2-RU	Q1	**99**	**98**	**63**	**100**	**1.1**
		(32.4)	(30)	(19.2)	(38.7)	(0.5)
	Q2	**108**	**153**	**61**	**90**	**1.8**
		(28.8)	(35.7)	(16.1)	(33.3)	(0.6)
	Q3	**106**	**155**	**62**	**91**	**1.8**
		(40.2)	(40.6)	(16.6)	(30.8)	(0.6)

In the L2-RU group, in vocalic segments, Quantity has a strong effect on V1 [$F(2, 321) = 88.8$; $p < 0.001$], and on V1/V2 ratio [$F(2, 321) = 61.8$; $p < 0.001$], but a rather weak effect on V2 [$F(2, 321) = 2.8$; $p = 0.06$]. The post-hoc test revealed differences in V1 and in V1/V2 ratio ($p < 0.001$) in Q1 and Q2, but not in Q2 and Q3; Quantity had no effect on V2 duration.

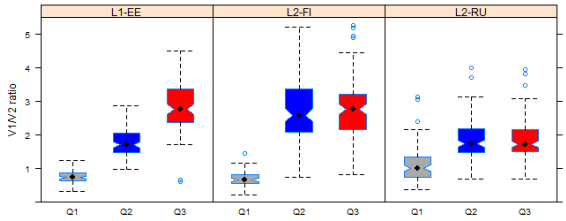


Figure 1: *Box plots of V1/V2 duration ratio in Q1 (gray), Q2 (blue) and Q3 (red) produced by L1-EE (left), L2-FI (mid) and L2-RU (right) subject groups.*

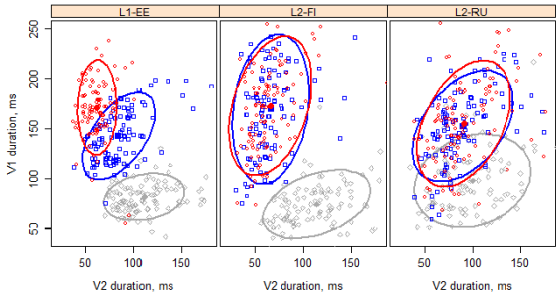


Figure 2: *Scatter plots of V1 and V2 durations in Q1 (gray), Q2 (blue) and Q3 (red) produced by L1-EE (left), L2-FI (mid) and L2-RU (right) subject groups.*

Compared with the L1-EE group, L2-RU vocalic segments deviate in Q1 ($p < 0.001$ in the case of V1 and V1/V2 ratio, $p < 0.05$ in the case of V2) and in Q3 ($p < 0.001$ in the case of V2 and V1/V2 ratio, $p < 0.01$ in the case of V1); no difference between the two groups was found in Q2. Similarly with the L2-FI group, the L2-RU group produces a single temporal pattern for both Q2 and Q3, but it coincides with Q2-pattern of the L1-EE group (see Figure 1 and 2, right).

3.2. Consonant-peaked structures

Table 2 provides mean segment durations of the consonant-peaked word structures in Q1, Q2, and Q3 produced by subjects in the three groups. Part A includes duration data of the Non-plosive set (36 subjects x 12 words = 432 words), and part B of the Plosive set (36 subjects x 9 words = 324 words). Variations in Syll rhyme/V2 ratio and Syll rhyme and V2 durations are shown in box plots (Figure 3) and scatter-plots (Figure 4).

In the L1-EE group, Quantity strongly effects the duration of Syll rhyme in both the Non-plosive [$F(2, 141) = 212$; $p < 0.001$] and in the Plosive [$F(2, 105) = 382.7$; $p < 0.001$] sets. Also V2 duration differs significantly among quantities in the Non-plosive [$F(2, 141) = 26.6$; $p < 0.001$] and the Plosive [$F(2, 105) = 43.2$; $p < 0.001$] sets. As a result, the characteristic duration ratio Syll rhyme/V2 reliably separates the three quantity degrees in both sets (Non-plosive set: [$F(2, 141) = 160.5$; $p < 0.001$], Plosive set [$F(2, 105) = 326.2$; $p < 0.001$]). Comparing L1-EE segment durations in the Non-plosive set to those of the Plosive set, differences in most segments emerge. However, these differences are natural due to the variable phonemic identity of the segments. It is important to notice that in both sets segment durations manifest quantity related patterns similar to the vowel-peaked structures: Syll rhyme duration is proportional to the quantity degree (shortest in Q1 and longest in Q3) and V2 duration is inversely proportional (longest in Q1 and shortest in Q3).

Table 2: *Mean duration (in ms) and standard deviation (in parenthesis) of C1, Syll rhyme (S1-rh), C2-s2, V2, and Syll rhyme/V2 duration ratio (S1-rh/V2) in the three word structures representing the quantity contrasts Q1, Q2 and Q3 in words in the Non-plosive (A) and the Plosive (B) sets read by L1-EE, L2-FI and L2-RU subjects.*

A: Non-plosive set						
Group	Qs	C1	S1-rh	C2-s2	V2	S1-rh/V2
L1-EE	Q1	**87** (17.7)	**85** (13.4)	**63** (20.4)	**93** (29.8)	**1.0** (0.28)
	Q2	**78** (18.7)	**152** (19.1)	**55** (10.1)	**78** (17.4)	**2.0** (0.49)
	Q3	**77** (18.9)	**191** (37.3)	**55** (10.1)	**61** (14.5)	**3.3** (0.93)
L2-FI	Q1	**106** (20.5)	**73** (15.4)	**69** (17.7)	**97** (23.9)	**0.8** (0.22)
	Q2	**98** (17.1)	**154** (36.3)	**67** (19.7)	**71** (26.4)	**2.4** (0.83)
	Q3	**96** (19.3)	**159** (37.8)	**67** (19.7)	**65** (18.1)	**2.6** (0.91)
L2-RU	Q1	**111** (26)	**95** (18.7)	**77** (28.2)	**79** (32)	**1.4** (0.63)
	Q2	**108** (24.3)	**157** (22.3)	**61** (16.3)	**81** (30.6)	**2.1** (0.6)
	Q3	**107** (18.8)	**155** (37)	**61** (16.3)	**63** (20.9)	**2.7** (1.12)
B: Plosive set						
L1-EE	Q1	**58** (14.6)	**74** (16.9)	**74** (12.3)	**100** (25.2)	**0.8** (0.18)
	Q2	**57** (19.2)	**132** (12.1)	**67** (6.7)	**77** (18.4)	**1.8** (0.52)
	Q3	**49** (16.3)	**199** (25.7)	**67** (6.6)	**55** (13.5)	**3.8** (0.68)
L2-FI	Q1	**60** (20.4)	**65** (13.9)	**92** (23.8)	**107** (41.6)	**0.7** (0.28)
	Q2	**58** (16.7)	**134** (34.6)	**75** (29.5)	**74** (36)	**2.4** (1.36)
	Q3	**56** (15.7)	**187** (43.9)	**75** (29.5)	**64** (22.4)	**3.2** (1.3)
L2-RU	Q1	**71** (24.5)	**91** (29.5)	**69** (14.3)	**95** (41.1)	**1.1** (0.53)
	Q2	**74** (20)	**138** (21.3)	**67** (16.5)	**65** (16.2)	**2.2** (0.49)
	Q3	**63** (16.3)	**180** (40.6)	**67** (16.5)	**58** (18.3)	**3.3** (1.15)

Quantity effects Syll rhyme and V2 durations, and also their ratios in both L2 groups, however, differences from the native group exist. In the L2-FI group, in the Non-plosive set, Quantity has strong effect on Syll rhyme [$F(2, 141) = 111.8$; $p < 0.001$]); however, Syll rhyme differs only among Q1 and Q2 (Q1: 73 ms, Q2: 154 ms; $p < 0.001$), but not in Q2 and Q3 (Q2: 154 ms, Q3: 159 ms; $p = 0.7$). In the Plosive set, L2-FI subjects produce different Syll rhyme durations in the three quantity degrees [$F(2, 105) = 121.8$; $p < 0.001$]), the mean durations for Q1 are: 65 ms, for Q2: 134 ms, and for Q3: 187 ms. V2 duration

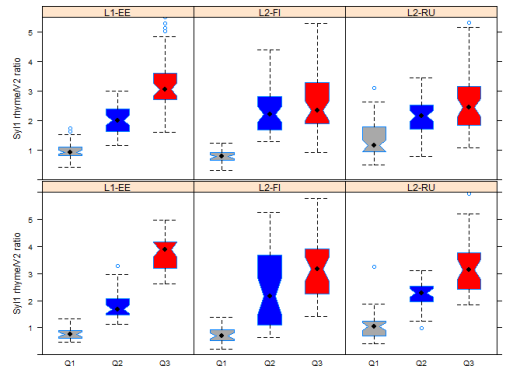


Figure 3: *Box plots of Syl1 rhyme/V2 duration ratio in Q1 (gray), Q2 (blue) and Q3 (red) of Non-plosive (top) and Plosive (bottom)sets produced by L1-EE (left), L2-FI (mid) and L2-RU (right) subject groups.*

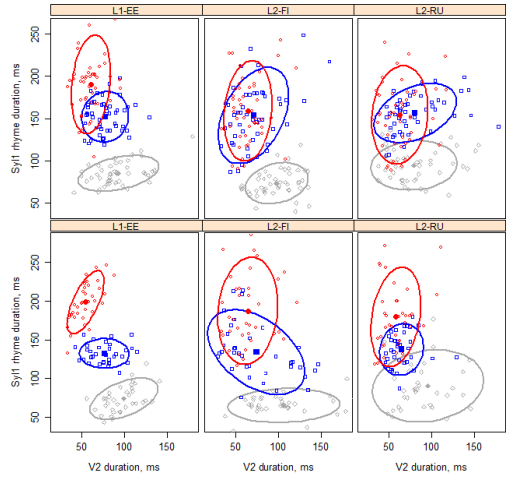


Figure 4: *Scatter-plots of Syl1 rhyme and V2 durations in Q1 (gray), Q2 (blue) and Q3 (red) of Non-plosive (top) and Plosive (bottom) sets produced by L1-EE (left), L2-FI (mid) and L2-RU (right) subject groups.*

has similar patterns in both word sets: V2 duration differs only between Q1 and Q2 ($p < 0.001$), but not between Q2 and Q3. As a result, the Syl1 rhyme/V2 ratio in the Non-plosive set is 0.8 for Q1, 2.4 for Q2, and 2.6 for Q3; these ratios differ between Q1 and Q2 ($p < 0.001$), but not between Q2 and Q3 ($p = 0.2$). In the Plosive set, the duration ratios are 0.7, 2.4, and 3.2, respectively, and provide a reliable difference between all quantity degrees ($p < 0.001$).

For the L2-RU group, Quantity has a similar effect as in the case of L2-FI group for Syl1 rhyme which differs in the Non-plosive set between Q1 and Q2 only (Q1: 95 ms, Q2: 157 ms; $p < 0.001$), and not between Q2 and Q3 (Q2: 157 ms, Q3: 155 ms; $p = 0.9$), and in the Plosive set, it differs among all quantities [$F(2, 105) = 72.2$; $p < 0.001$]), the mean durations are 91 ms, 138 ms, and 180 ms for Q1, Q2 and Q3, respectively. V2 duration shows different patterns: in the Non-plosive set it is almost the same in Q1 and Q2 (Q1: 79 ms, Q2: 81 ms; $p = 0.96$), but it differs between Q2 and Q3 (Q2: 81 ms, Q3: 63 ms; $p < 0.01$); in the Plosive set V2 duration differs between Q1 (95 ms) and Q2 (65 ms) ($p < 0.01$), but not between Q2 and Q3 (58 ms) ($p = 0.6$). Consequently, Syl1 rhyme/V2 duration ratio is distinctive among all quantity degrees in both word sets: in the Non-plosive set Q1: 1.4 vs. Q2: 2.1 ($p < 0.001$) and Q2: 2.1 vs. Q3: 2.7 ($p < 0.01$); in the Plosive set Q1: 1.1 vs. Q2: 2.2 vs. Q3: 3.3 ($p < 0.001$).

4. Discussion

The languages involved in the study differ in the way the duration cue is exploited in phonological contrasts. Estonian and Finnish are quantity languages, both exploiting the duration cue contrastively, while Russian is a non-quantity language lacking duration-based phonological oppositions. There are two theoretical models addressing the role of duration in L2 speech: (1) the *Feature Hypothesis* states that "L2 features not used to signal phonological contrasts in L1 will be difficult to perceive for the L2 learner and this difficulty will be reflected in the learner's production" [22], (2) the *Desensitization Hypothesis* states that duration cues are easy to access whether or not listeners have had specific linguistic experience with them [23].

The first model predicted an advantage for the L2-FI group over the L2-RU group since L2-FI subjects can exploit the binary contrasts available in Finnish. The second model suggested that also L2-RU subjects should be able to distinguish Estonian quantity contrasts even though there is no corresponding prosodic pattern in Russian to rely on. The Q1 vs. Q2 results support the second hypothesis since both L2 groups were successful in producing the Q1 vs. Q2 opposition. With respect to the Q2 vs. Q3 results, neither hypothesis can be favored (more space would be needed to elaborate on theoretical implications).

The most surprising result was that the L2-RU group outperformed the L2-FI group in producing distinct Q2 and Q3 patterns in the Non-plosive set. This may have to do with the fact that our Finnish speakers had learned Estonian as adults while the Russian speakers mostly in childhood, in addition, the exposure to and use of Estonian is much more frequent in the case of the L2-RU group. Also, the methods how the Estonian quantity contrasts have been taught might effect the results.

5. Conclusions

In the paper we studied the effect of L2 orthographic input on L2 pronunciation of vowel-peaked and consonant-peaked quantity contrasts. Only in the Plosive set the quantity contrasts are explicitly expressed in the orthography, unlike the Vowel set and the Non-plosive set. The results confirmed the role of L2 orthographic input in the case of the L2-FI group – in the Plosive set the L2-FI subjects produced different patterns for Q2 and Q3 structures, but not in the other sets. However, the role of orthography can not be conclusive in the case of the L2-RU subjects since they produced different patterns for Q2 and Q3 structures also in the Non-plosive set.

6. Acknowledgements

We thank our volunteer speakers in Finland and Estonia, and Martti Vainio, Kari Suomi and Stina Ojala for their help in arranging the recordings at Helsinki, Oulu and Turku University.

This work has been partially supported by the Estonian Ministry of Education and Research target-financed research theme No. 0140007s12 and by the National Program for Estonian Language Technology.

7. References

[1] Lehiste, I., "Search for Phonetic Correlates in Estonian Prosody", in I. Lehiste and J. Ross [Eds], Estonian Prosody: Papers from a Symposium, Tallinn: Institute of Estonian Language, 11-35, 1997.

[2] Eek, A. and Meister, E., "Simple perception experiments on Estonian word prosody: foot structure vs. segmental quantity", in I. Lehiste and J. Ross [Eds], Estonian Prosody: Papers from a Symposium, Tallinn: Institute of Estonian Language, 71-99, 1997.

[3] Krull, D. and Traunmller, H., "Perception of quantity in Estonian", Proceedings, Fonetik 2000, 85-88, 2000.

[4] Eek, A. and Meister, E., "Foneetilisi katseid ja arutlusi kvantiteedi alalt (I). Häälikukestusi muutvad kontekstid ja välde", Keel ja Kirjandus, 11-12, 815-837, 904-918, 2003.

[5] Meister, L., "Eesti vokaali- ja kestuskategooriad vene emakeelega keelejuhtide tajus ja hääldus es", Ph.D. Dissertation, Tartu: Tartu likooli Kirjastus, 2011.

[6] Meister, E. and Meister, L.,"Production of Estonian quantity contrasts by native speakers of Finnish", Proceedings of Interspeech 2013, Lyon, 330-334, 2013.

[7] Lehiste, I., "Segmental and syllabic quantity in Estonian", American Studies in Uralic Linguistics 1, Bloomington: Indiana University Press, 21-82, 1960.

[8] Liiv, G., "Eesti keele kolme vältusastme vokaalide kestus ja meloodiatüübid", Keel ja Kirjandus, 412-424, 480-490, 1961.

[9] Eek, A., "Observations on the duration of some word structures: I", Estonian Papers in Phonetics, 18-31, 1974.

[10] Krull, D., "Stability in some Estonian duration relations", Institute of Linguistics, University of Stockholm, PERILUS 13, 57-60, 1991.

[11] Krull, D., "Temporal and tonal correlates to quantity in Estonian", Institute of Linguistics, University of Stockholm, PERILUS 15, 17-36, 1992.

[12] Krull, D., "Word-prosodic features in Estonian conversational speech: Some preliminary results", Institute of Linguistics, University of Stockholm, PERILUS 17, 45-54, 1993.

[13] Asu, E. L., Lippus, P., Teras, P., and Tuisk, T., "The realization of Estonian quantity characteristics in spontaneous speech", in M. Vainio, R. Aulanko, and O. Aaltonen [Eds], Nordic Prosody, Proceedings of the Xth conference, Frankfurt, Berlin, New York: Peter Lang, 49-56, 2009.

[14] Meister, L. and Meister, E., "Perception of the short vs. long phonological category in Estonian by native and non-native listeners", Journal of Phonetics, 39(2): 212-224, 2011.

[15] Meister, L. and Meister, E., "The production and perception of Estonian quantity degrees by native and non-native speakers", Proceedings of Interspeech 2013, Portland, 886-889, 2012.

[16] Meister, E. and Meister, L., "Native and non-native production of Estonian quantity degrees: comparison of Estonian, Finnish and Russian subjects", in E. L. Asu & P. Lippus [Eds], Nordic Prosody. Proceedings of the XIth conference, Frankfurt am Main: Peter Lang Verlag, 235-243, 2013.

[17] Meister, E. and Meister, L., "Production and perception of Estonian quantity contrasts by L2 subjects with different language backgrounds", PPLC13: Phonetics, phonology, languages in contact, Paris, Book of Abstracts, 41-43, 2013.

[18] Meister, L. and Meister, E., "Aktsendikorpus ja võõrkeele aktsendi uurimine", Keel ja Kirjandus, 55(8-9): 696-714, 2012.

[19] Meister, L. and Meister, E.," The Estonian Foreign Accent Corpus", PPLC13: Phonetics, phonology, languages in contact, Paris, Book of Abstracts, 141-143, 2013.

[20] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program], Version 5.3.60, retrieved 8 December 2013 from http://www.praat.org/.

[21] Eek, A. and Meister, E., "Foneetilisi katseid ja arutlusi kvantiteedi alalt (II): Takt, silp ja välde", Keel ja Kirjandus, 47(4): 251-271, 2004.

[22] McAllister, R., Flege, J. E., and Piske, T., "The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian", Journal of Phonetics, 30, 229-258, 2002.

[23] Bohn, O-S., "Cross-language speech perception in adults: First language transfer doesnt tell it all", Speech Perception and Linguistic Experience. Issues in Cross-Language Research, [Ed] W. Strange. Baltimore: York Press, 279-304,1995.

Variation in list intonation in American Jewish English

Rachel Steindel Burdin[1]

[1]Department of Linguistics, The Ohio State University, USA

burdin@ling.osu.edu

Abstract

Yiddish-influenced intonation has been previously noted as a potential defining characteristic of American Jewish English and, specifically, list intonation identified as a possible area of differentiation. However, apart from remarks in general descriptions of Standard American English (SAE) prosody, a systematic study of list intonation has not been conducted in SAE. In this study, lists were defined, and extracted from sociolinguistic interviews with Jewish women with varying degrees of exposure to Yiddish. Speakers from different language backgrounds differed significantly in their use of contours, boundary tones and pitch accents on list items, with speakers with less exposure to Yiddish using more of the standard English contour (H* H-L%) than speakers with more exposure to Yiddish. Yiddish bilinguals were more likely to use a rise fall contour (L+H* L-L%), fewer H-L% boundary tones and H* pitch accents, and more rising pitch accents (L+H* and L*+H) than non-bilinguals. In addition, speakers of all language backgrounds used a variety of list intonations, showing the need for more systematic study into the uses and meanings, social and otherwise, of list intonations in English.

Index Terms: list intonation, variation, Jewish English

1. Introduction

"List intonation" is, in some sense, a misnomer, as there are in fact several list intonations that have been described for American English. The distinction between the types has been said to be one of pragmatics: Ladd [1] notes two list intonations, one involving a rise, and the other, a high plateau, with the rising one being used for exhaustive lists (those which have listed all possible items) and the plateau being used for non-exhaustive lists (in which the items are taken to be merely representative of a larger set). The high plateau, H* H-L%, can be seen in figure 1: the pitch, starting at the stressed syllable and through the end of the intonational phrase, is generally high, and flat. However, more than two intonations can be used in lists in English, although their use has not been studied systematically: Schubiger [2], for example, describes five different patterns that can occur on "enumerations".

Intonation has been noted as a potential characteristic feature of American Jewish English speech: in studies of Orthodox Jewish communities, both Fader's [3] and Benor's [4] subjects claimed that they could tell if someone was Jewish by their intonation. It is unclear what, exactly, this distinctive intonation is: A "rise-fall" contour has been proposed as a feature of Jewish English as far back as 1956 [5]. However, this contour does not seem to be phonologically distinct, being ToBI transcribed as L+H* !H-L% [6], a contour which does exist and is used in Standard American English, usually as a vocative or some other stylized utterance [1]. It is more likely, then, that the differences in Jewish English intonation lie in either differences in phonetic implementation of contours, in differences in the use and meaning of contours, or both.

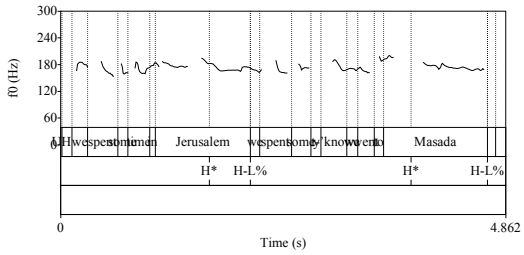


Figure 1: A standard H* H-L% on list items-"We spent some time in Jerusalem, we went to Masada"

One potential area of difference is in lists: some Jewish speakers produced non-exhaustive lists that contained falls on non-final items (described as being rarer by Schudiger [2]), as in figure 2.The speaker is describing how Brooklyn has changed since she lived there: it has become cool, and that there's always something going on, "with artists, with music, with food, with breweries". Each phrase has a H* L-L% contour, with high pitch on the stressed syllable, which then falls to the bottom of the speaker's pitch range. These types of lists, with falling (or rise-falling), rather than flat, contours, sounded distinctly "Jewish" to some listeners.[1]

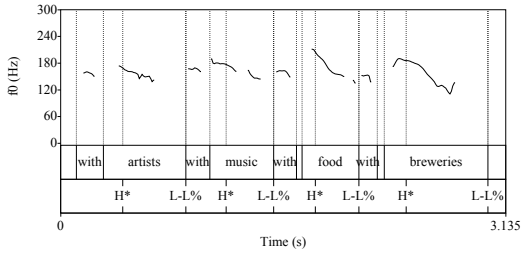


Figure 2: H* L-L% on list items

Data were analyzed from a series of sociolinguistic interviews to see if there was variation in list intonation between 10 Jewish women from the metropolitan NYC area. In addition, this study provides, as far as this author knows, the first systematic study of list intonation in any variety of American English.

[1]I am grateful for participants at the Linguistic Society of America 2013 meeting for this observation.

2. Methodology

2.1. The interviews

Ethnographic-type interviews with 10 Jewish women were conducted in the metropolitan New York area. The sample is non-random, with most of the subjects recruited through personal contacts. The women varied in their degree of exposure to Yiddish and other varieties with significant Yiddish influence. Degree of exposure to Yiddish has been previously found to influence the use of Jewish English features. Jewish speakers with more exposure to Yiddish use forms that show more Yiddish-like traits (e.g., using a loanword like *schmooze* with its original Yiddish meaning of "chat"), and speakers with less use forms more like non-Jewish speakers (e.g., using *schmooze* to mean "network, chat up") [7]. As such, it was expected that the speakers with more exposure to Yiddish would show some influence of Yiddish in their intonation, while those with less exposure would behave more like Standard American English speakers.

Three of the women, the bilingual group, grew up bilingual in Yiddish and English, in neighborhoods in Newark and New York City that had large populations of Yiddish-speaking immigrants. However, it should be noted that these women were, at the time of the interview, English-dominant, with no noticeable Yiddish accent, beyond possibly their prosody. Three of the women, the mixed exposure group, grew up speaking only English, but had significant exposure to Yiddish-influenced varieties: one had Yiddish speaking parents and lived in a neighborhood in New York City with a large immigrant population; the other two spent significant amounts of time in Israel and reported high proficiency in Israeli Hebrew, whose prosody is almost certainly heavily influenced by Yiddish [8]. Finally, four of the women, the limited exposure[2] group, grew up monolingual in primarily English-speaking neighborhoods consisting mostly of second or third generation Americans.

As the target of investigation was potentially a socially meaningful one, it was thought that a more naturalistic setting, with a known interviewer, would prompt the subjects to produce more speech that was markedly Jewish, and indeed, most of the subjects, particularly when discussing Jewish topics, did, e.g., use Jewish English lexical items, mostly loan words from Hebrew and Yiddish. The interview as a whole centered around the interviewee's life, and contained several prompts that were amenable to list-type responses: e.g., What did you do for fun when you were younger? What sort of activities would you do in New York City? The interviews lasted around 45 minutes.

Subjects were recorded using a head mounted microphone with a digital recorder. Transcripts of the interviews were made using ELAN, and analyzed using PRAAT [9].

2.2. Extracting lists

Lists have been described in various ways, with some claiming that only two items are necessary to make a list [10]; and others, three, with three items tending to be preferred [11]. I follow Selting [12] in allowing several characteristics to define a list, with one exception: Selting cites prosody as factor in marking a sequence of items as a list; however, using prosody as a feature to define a list would be obviously problematic for this study. As such, lists were selected solely on the basis of the transcripts of the interviews.

[2] As opposed to no exposure group; these women would have had experience, for example, speaking to women in the first or second group, and had command over various Yiddish lexical items and other linguistic features that are a part of the American Jewish English repertoire

A list was defined as a group of utterances of at least two items that were (1) non-temporally ordered and (2) displayed syntactic parallelism. Parallel NPs ("I had to buy everything-pillows, blankets, pots, pans...") and PPs (as in figure 2) were automatically included; parallel VPs and sentences needed to match in some other way besides simply being VPs and sentences, either (for sentences) by having some repeated part (e.g., "You could walk to Chinatown, you could walk over the bridge.."), or by having a similar argument structure (e.g., "We watched movies, we ate junk food..." but not "We had a lot of fun, we lived on a cul de sac, we used to play softball right in the street with our neighbors"). In addition, sequences of items concluded with a specific marker of list-hood, e.g., including a phrase like "and things like that", or, in one case, explicitly stating "we have a list", were included, regardless of whether or not the list items met the previous criteria.

Although previous studies have described lists as full, definable units, with the same prosody on each item apart from possibly the last (for example, Selting's study [12]), for this study, this was often not the case. While there seemed to be a preference for consistent prosody across list items, speakers did switch in the middle of a list. It was also difficult to determine what should count as a "list", singular. Take the following example, where a speaker is describing a part of Brooklyn:

(1)
1. You could walk to Chinatown
2. You could walk over, over the bridge
3. You can walk to work if you work downtown
4. You could jog over the bridge coming home
5. You had great restaurants
6. You had, um, beautiful, beautiful brownstones
7. And apartment buildings
8. And as I said, tree-lined-
9. Beautiful tree-lined streets.

The first four items are a list due to the syntactic parallelism, all being sentences starting with *you* + a modal; one could also perhaps include lines 5 and 6 due their starting with "you had", but here, the speaker has switched from listing attributes that were good about living near the Brooklyn Bridge to attributes that were good about the neighborhood in general. However, line 6 also seems to go with lines 7, 8, and 9, which are slightly different again: here, she's listing things that were attractive in the neighborhood, and has switched to listing NPs, as opposed to full sentences. Deciding whether this is one, two or three lists is non-trivial.

Similar problems can be seen in example 2, from a speaker describing her trip to Israel:

(2)
1. We spent some time in Jerusalem
2. We spent some time in-
3. We went to Masada
4. We went to the Dead Sea...

The false start in line 2, with the same frame as line 1, indicates that item 1 is meant to be in a list; however, the speaker reconsiders, perhaps because it's not possible to "spend some time" in Masada[3], and sticks with this new frame for the rest of the list. It was decided, in this case, to include item 1, as some syntactic parallelism exists with lines 3 and 4 (simple past tense verbs, with a location), and it fit into the theme of the list – places this speaker had gone in Israel – quite clearly.

[3] As it is a archeological site on a mountain, rather than a city.

Due to the problems of demarcating the lists as a whole, as well as the fact that often times, the lists did not have consistent prosody across all items, the results reported here are based on the list items, rather than by looking at the lists as whole units. This means, for example, that the entirety of examples (1) and (2) were included in the analysis, apart from those lines containing disfluencies at the end of the IP (example (1), line 8, and example (2), line 2).

The list items were extracted, and ToBI transcribed by the author based on the ToBI guidelines [13]. Due to the unscripted nature of the data, phonetic differences were not looked at this time; however, it is likely that they exist as, e.g., wider pitch range has been cited as a feature of Jewish English [14]. Uncertain transcriptions were discarded. Only list items ending with a full IP break were analyzed. Most of the time, there was a clean mapping of list item to IP; however, there were, at times, multiple IPs within a list item. These "extra" IPs were tagged as comments, and were excluded from the analysis. 504 list items were initially extracted; 467 were included in the analysis.

Lists were annotated for being exhaustive or not, as, as mentioned above, it has been claimed that exhaustive and non-exhaustive lists might have different prosody. An exhaustive list is one that a speaker intends to be a full account of possible items; a non-exhaustive list has items that are meant to be representative. Again, the distinction between exhaustive and non-exhaustive was based on the text, not on the prosody of the utterance, and fairly strict definition was used for "non-exhaustive": either the speaker had to name all items that could be in the list, or the speaker had to have stated that a specific number of items were in the list (e.g. "My three sons...").

List items were also annotated based on their position in the lists, first, medial, or last. For most of the lists, this was fairly straightforward; however, for lists like examples (1) and (2), the entire sequence of utterances was taken as one "list", and first, last, and medial were assigned accordingly.

3. Results

Table 1 gives an overview of the nuclear pitch accents and boundary tones used on the list items, sorted by language status; in this and the following tables, only the most common contours are shown for reasons of space. The limited exposure group preferred the contour previously noted for standard English, H* H-L%, shown in figure 1 above, with the H* !H-L% contour, consisting of a high tone on the stressed syllable and either a drop to a mid-range plateau, or a slight fall, shown in figure 3, as a slightly distant second.

Table 1: Nuclear pitch accents and boundary tones on list items

Bilinguals (n= 121)	Mixed (n=93)	Limited (n=253)
H* !H-L% (20%)	H* H-L% (26%)	H* H-L% (45%)
L+H* L-L% (15%)	H* L-L%(18%)	H* !H-L% (16%)
H* H-L% (15%)	H* !H-L% (12%)	H* L-L% (10%)
H* L-L% (14%)	L+H* !H-L% (11%)	L* H-L% (7 %)
L* L-H% (9 %)	H* L-H% (9%)	!H* H-L% (7 %)

The mixed and bilingual groups had more variety in their choice of pitch accents and boundary tones: For the mixed group, the H* H-L% contour, although a plurality of the contour types, was not close to being the majority, and, for the bilingual

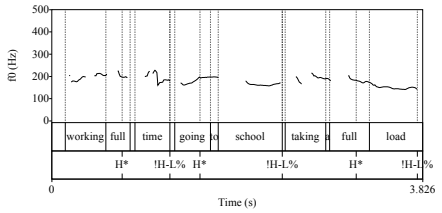


Figure 3: List items with H* !H-L%

group, the standard English contour tied with a rise-fall contour, L+H* L-L% (consisting of a sharp rise to the stressed syllable, then fall to the bottom of the speaker's pitch range, shown in figure 4) in second, with H* !H-L% ahead of it; preference was fairly evenly divided among the top 4 contours; again, cf. the limited group, which showed a clear preference for H* H-L%.

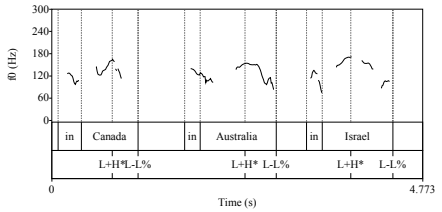


Figure 4: List items with L+H* L-L%

A logistic regression model was built using the contour type (H* H-L% or not[4]) as the dependent variable, as this is, as noted above, the contour previously noted for standard English, and also the contour that showed the most variance in use between the three groups. Language status, exhaustiveness of the list, and list position were included as fixed effects, and speaker as a random effect. Sum contrasts were used for all fixed effects. Log likelihood testing comparison revealed no significant differences between models including exhaustiveness and position and those without, so reports from the simplest model, with language status as a fixed effect, and speaker as a random effect, are included here. It was found that the limited exposure speakers and the mixed speakers were more likely to use the standard English H* H-L% contour ($p<0.001$, and $p<0.01$) than the average; the bilingual speakers were less likely to use the contour ($p<0.001$) than the average.

Another model was built using the L+H* L-L% contour or not as the comparison, as this was the second most used contour by the bilinguals, and showed, again, variance between the three groups. Position and exhaustiveness were found to be non-significant, and were removed from the model. The bilinguals used this contour significantly more ($p<0.05$), and the mixed group, less ($p<0.05$). The limited group also trended in this direction, but not significantly so ($p = 0.0508$). For the H* L-L% contour, no significant differences emerged for language status.

Table 2 shows the percentage of different boundary tones used by the speakers on list items. Again, a logistic regression model was built, using the same effects as above; however, neither position nor exhaustivetivity were significant. A log like-

[4]Converting the data to binary variables was necessary; multinomial logistic regression models were attempted, but failed to converge, probably due to the amount of data.

lihood comparison between models with and without these effects was non-significant, so the results from the model with only speaker and language status will be reported. The boundary tones were converted to binary factors: use of H-L% or not (the standard English contour), and use of L-L% (the contour preferred by the bilinguals) or not. The mixed and limited group were more likely to use H-L% boundary tones ($p<0.001$ and $p<0.05$) than the average; the bilingual group was less likely ($p<0.001$). No significant difference was found for the use of the L-L% boundary tones.

Table 2: Boundary tones

Tone	B (n= 121)	M (n=93)	L (n=253)
H-L%	20 %	36 %	48 %
L-L%	36 %	28 %	26 %
!H-L%	26%	22 %	17%
L-H%	16 %	11 %	7 %
H-H%	2 %	2 %	1%

Table 3 shows the pitch accents used on the nuclear accent of the IP. The pitch accents were converted to binary factors: H* pitch accent (the standard, preferred by all groups) or not, and rising pitch accents (L+H* and L*+H) or not (as these seemed to be preferred by the bilinguals), and logistic regression models were run with the same fixed and random effects as above; in this case, position was found to be significant, but exhaustiveness of the list was not, and so position, but not exhaustiveness, was included in the model. For the first model, the H* pitch accent was more likely to be used for first items. Bilinguals were less likely to use H* overall ($p< 0.01$); the limited exposure group and the mixed group were more likely to use H* pitch accents overall ($p<0.01$ for both). For the second model, looking at the rising pitch accents (L+H* and L*+H), again, position was significant, with first and last items being less likely ($p<0.01$ and $p<0.05$) to receive rising accents than middle items. The bilingual group was more likely to use the rising pitch accents ($p<0.01$); the mixed group trended in this direction, but not significantly so ($p = 0.059$). The limited group was less likely to use these accents ($p<0.01$).

Table 3: Pitch accents

Pitch accent	B (n= 121)	M (n=93)	L (n=253)
H*	55 %	68 %	74 %
L+H*	20 %	15 %	7%
L*+H	10 %	5 %	0 %
!H*	10 %	3 %	7 %
L*	15%	7 %	11%

4. Discussion

Table 4 provides a summary of the findings above. The differences between the groups showed up as significant in the use of the H* H-L% contour, L+H* L-L% contour, H-L% boundary tone, and H* and rising pitch accents.

The mixed group overall, looked more like the limited exposure group than the bilingual group, with the exception of use of the rising pitch accents, where they were not significantly different from the average, as well as in their use of the L+H* L-L% contour, where the limited group was not significantly different from the average. The mixed group, appropriate to their name, might occupy a space in between the bilingual group and the limited group in a very real sense, using mostly standard English prosody, but being able to adopt various features otherwise associated with the bilingual group when needed or wanted. The bilingual group, on the other hand, was markedly different from both groups. It is very likely that this difference comes from their Yiddish background. There has not yet been a systematic study of Yiddish prosody, so this remains speculative; however, the use of falling contours on lists has been noted in German [12], a language very closely related to Yiddish.

Table 4: Summary of findings: Use of contours, boundary tones, and pitch accents, compared to average. *($p<0.05$), **($p<0.01$), ***($p<0.001$)

	Bilingual	Mixed	Limited
H* H-L%	less***	more **	more ***
L+H* L-L%	more*	less*	n.s.
H* L-L%	n.s.	n.s.	n.s.
H*	less**	more**	more**
L+H*/L*+H	more *	n.s.	less**
H-L%	less***	more*	more**
L-L%	n.s.	n.s.	n.s.

In addition to this variability across groups, there was considerable variation within the groups, as even the speakers with limited access to Yiddish showed some variety in the type of contours used: although the H* H-L% was preferred by this group, other contours were used as well. It is likely that some of these contours have a difference in meaning or use from the H* H-L% contour above: although exhaustivity did not come out as significant in any of the models, this may be due to the fact that in some cases, the list's status as either exhaustive or not might be marked primarily by the prosody, rather than being able to be extracted from the text string, and that the prosody was the only clue to exhaustiveness vs. non-exhaustiveness. However, this still leaves the question of what, for example, separates the H* H-L% lists from the H* !H-L% lists.

These findings also open up the possibility that, in addition to more pragmatic differences between the use of the contours, like exhaustive and non-exhaustive, there might be social meaning differences as well, with the use or non-use of certain contours perhaps marking a speaker as older, or Jewish. The differences might also have to do with attitudes towards the items being listed: for lists with falling contours, the speakers occasionally sounded dismissive, or annoyed. However, perception studies will be needed to more accurately study what these contours mean, both in terms of social meaning and other meanings.

5. Acknowledgements

The author would like to thank Cynthia Clopper, Kodi Weatherholz, Rory Turnbull, and Joseph Tyler, for helpful discussion and comments, as well as the OSU Department of Linguistics for providing travel funding and equipment.

6. References

[1] Ladd, R. D. 1980. *The structure of intonational meaning*. Bloomington: Indiana University Press.

[2] Schubiger, Maria. 1958. *English intonation: Its form and function*. Tübingen: Max Niemeyer Verlag.

[3] Fader, A. 2009. *Mitzvah Girls: Bringing up the Next Generation of Hasidic Jews in Brooklyn*. Princeton: Princeton University Press.

[4] Benor, S. B. 2004 Second style acquisition: The linguistic socialization of newly Orthodox Jews. PhD Thesis: Stanford.

[5] Weinreich, U. 1956. Notes on the rise-fall contour. In *For Roman Jakobson*, M. Halle ed. The Hague: Mouton.

[6] Burdin, R. S. 2012. Variation in Jewish English intonation. Presented at NWAV 41, Bloomington, IN.

[7] Benor, S. B. 2011. *Mensch, bentsch* and *bagalen*: Variation in the American Jewish Linguistic Repertoire. *Language and Communication* **3 1/2**, special issue "Jewish Languages in the Age of the Internet". 141-154.

[8] Amir N., Vered, S.V-., and Izre'el, S. 2004. Characteristics of intonation unit boundaries in spontaneous spoken Hebrew: Perception and acoustic correlates. *Proceedings of the 2nd International Conference of Speech Prosody*. Nara, Japan. 677-680.

[9] Boersma, P. 2001. Praat, a system for doing phonetics by computer. Glot international. 341-345.

[10] Lerner, G. H. 1994. Responsive list construction: a conversational resource for accomplishing multifaceted social action. *Journal of Language and Social Psychology* **12**, 20-33.

[11] Jefferson, G. 1990. List construction as a task and interacitonal resource. In *Interactioanl Competence*, Psathas, G. (ed.). New York: University Press of America. 63-92.

[12] Selting, A. 2007. Lists as embedded structures and the prosody of list construction as an interactional resource. *Journal of Pragmatics* **39**(3): 483:526

[13] Beckman, M.E., and Ayers, G.M. 1994. Guidelines for ToBI labeling guide, ver. 2.0. Manuscript, Ohio State University, Columbus, OH.

[14] Tannen, D. 2005. "New York Jewish conversational style." *Intercultural Discourse and Communication: The Essential Readings.* 135-149.

Incorporating Prosodic Boundaries in Unsupervised Term Discovery

Bogdan Ludusan[1], *Guillaume Gravier*[2], *Emmanuel Dupoux*[1]

[1]LSCP - EHESS/ENS/CNRS, Paris
[2]IRISA - CNRS, Rennes

bogdan.ludusan@ens.fr, guillaume.gravier@irisa.fr, emmanuel.dupoux@gmail.com

Abstract

We present a preliminary investigation on the usefulness of prosodic boundaries for unsupervised term discovery (UTD). Studies in language acquisition show that infants use prosodic boundaries to segment continuous speech into word-like units. We evaluate whether such a strategy could also help UTD algorithms. Running a previously published UTD algorithm (MODIS) on a corpus of prosodically annotated English broadcast news revealed that many discovered terms straddle prosodic boundaries. We then implemented two variants of this algorithm: one that discards straddling items and one that truncates them to the nearest boundary (either prosodic or pause marker). Both algorithms showed a better term matching F-score compared to the baseline and higher level prosodic boundaries were found to be better than lower level boundaries or pause markers. In addition, we observed that the truncation algorithm, but not the discard algorithm, increased word boundary F-score over the baseline.

Index Terms: term discovery, prosody, prosodic boundary

1. Introduction

During their first year of life, human infants extract word-like units from continuous speech without supervision [1]. In parallel, unsupervised term discovery (UTD) algorithms are increasingly used within speech technology [2, 3, 4, 5]. In both cases, the task consists in finding repetitive patterns, while using as input only the speech signal. An examination of how infants are solving this task may reveal useful strategies that could be implemented into UTD algorithms.

Many researchers have pointed out that prosody is an important cue that helps infants to segment continuous speech. Newborns are sensitive to the acoustic cues correlated with the presence or absence of phonological phrase boundaries in otherwise identical stretches of speech (e.g. /mati/ in "mathématicien" (mathematician) versus "panorama typique" (typical panorama) [6, 7]. Nine-month-olds use these cues to posit breaks within sentences [8, 9]. Ten- and thirteen-month old infants use these boundaries to constrain word recognition, i.e. they fail to recognize a string that straddles a phonological phrase boundary [10]. Similarly, adults use these boundaries to constrain online lexical cognition, i.e. they do not produce false alarms on word forms that straddle a phonological phrase boundary[11].

Unsupervised term discovery, so far, does not use prosodic information. The existing algorithms are based on computing a similarity score between stretches of speech signal, usually done by means of dynamic time warping (DTW). The proposed systems return a list of matched pairs [2, 3] and/or a library of clusters of discovered terms [2, 4, 5]. Some systems scan the entire corpus for repetitions [2, 3], while others only scan a small time buffer and match the signal against an incrementally built library of terms [4, 5]. The terms discovered with UTD systems have already been proven useful in a number applications like keyword spotting [12], topic segmentation [13], or document classification [14].

Based on the findings regarding the role of prosodic boundaries in speech processing for both children and adults, we investigated the use of such boundaries in unsupervised term discovery. The current study uses manually annotated prosodic boundaries in order to establish the upper boundary of their impact on the discovery task. The rest of the paper is organized as follows: the UTD system used in the experiments and the evaluation method are presented in section 2, while a short description of the corpus employed in this study is given in section 3. Two experiments are illustrated in section 4, in which we varied the strategy for using prosodic information. The paper concludes with a discussion of the results obtained and some possible paths to follow.

2. Methods

2.1. System Presentation

An open source system for spoken term discovery, MODIS [15], was employed for the experiments. It is based on a generic approach to mining repeating sequences, tolerant to term variability [5], and it uses a limited search buffer, making it more psychologically realistic than systems performing an exhaustive search.

MODIS takes a speech signal as input (represented as either MFCCs or posteriorgrams) and delivers a library of repeated terms as its output. Term discovery is based on the notion of seed fragments. A seed fragment is a stretch of signal segmented from the input stream and searched for in a fixed-length buffer ahead of the seed using a segmental variant of the DTW algorithm. If a match for the seed is found, the seed is extended to find the maximal length matching pattern and, if it exceeds a minimal term length, it is stored in the library of terms. This library is used as a long term memory to search for repeating terms: Each new seed considered is first matched against entries in the library before searching for self-repetitions in the buffer. Potential re-occurrences detected by DTW are validated using self-similarity matrix (SSM) comparison.

The key parameters of the seeded discovery algorithm are the seed and term lengths and a set of similarity thresholds used to validate template comparison (higher thresholds means potentially more variability at the expense of precision).

2.2. Evaluation Method

We evaluated two aspects of the discovered terms: the matching quality and the word boundary quality. For the matching

quality, we used a similar method to [16], which transforms the speech chunks corresponding to the found terms into a symbolic representation, based on the phonetic transcription of the speech signal. Then, the precision and the recall are determined based only on the strings of phonemes corresponding to the obtained terms. For a formal definition of the measures used for the evaluation, see Muscariello et al study [16]. The following steps are performed during the evaluation process:

- All phonemes falling inside the time interval corresponding to the found term are concatenated. A phoneme is considered to belong to the term if at least 50% of its duration falls within the term.
- For each class of terms, a centroid is computed, defined as being the string with the lowest normalized edit distance from all the other strings belonging to the class.
- The precision is calculated as the percentage of class members, out of the total number of tokens in the class, falling within a certain distance from the class centroid. The neighbourhood threshold was set to 0.2.
- Next, the recall is determined as being the percentage of how many strings belonging to the centroid neighbourhood were found, from the total number of occurrences of those strings in the whole corpus.

The second measure we evaluated, the word boundary quality, comes from the field of natural language processing (NLP) and it can be a useful measure when the terms discovered by the UTD systems are used in a downstream application. The boundary quality was computed by comparing the set of discovered term boundaries to the set of gold word boundary of the corpus, as done in [17]. We expect a very low recall on this metric, since UTD systems do not attempt to exhaustively segment a corpus, contrary to NLP systems, that perform term discovery based on text input.

3. Materials

The materials used in this paper are a subset of the Boston University Radio News Corpus (BU corpus) [18], which contains news stories recorded by 7 professional speakers. Out of the whole corpus, around 3.5 hours of data are annotated prosodically for phrase breaks and accent tones. The prosodic annotation is based on the ToBI system [19] for American English, which uses a 5-level scheme for prosodic boundaries of increasing strength, starting with cliticized word boundaries (level 0) and ending with intermediate phrase boundaries (level 3) and intonational phrase boundaries (level 4).

We used only levels 3 and 4 as we are interested in the effect of prosodic boundaries that can, in principle, be detected in an unsupervised fashion. We removed the recordings for which these two levels were missing and those with no phone-level segmentation, because this type of annotation was necessary for the evaluation procedure. Thus, for our experiments, we used about 3 hours of data, including 6 speakers (3 males, 3 females) distributed into 403 files. In these materials there were a total of 6059 intonational phrase boundaries and 2731 intermediate phrase boundaries annotations.

4. Experiments

We propose to investigate the usefulness of prosodic boundary information for term discovery, by integrating this type of information in MODIS in two experiments. In Experiment 1, we examined the idea of using boundary information to prune away

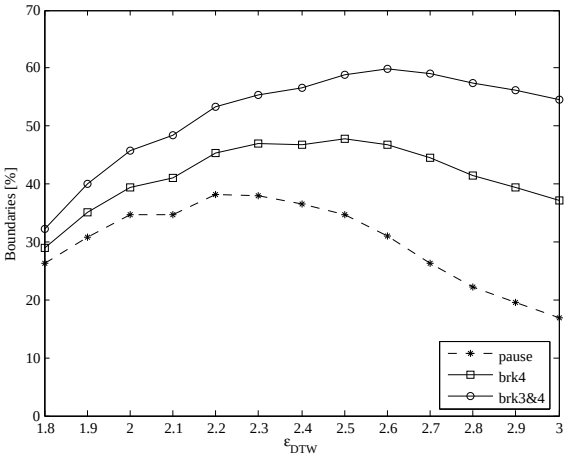


Figure 1: *Percentage of terms found by the baseline system overlapping pause markers, level 4 boundaries and level 3 and 4 boundaries.*

the terms found to be straddling a prosodic boundary ($discard$). In Experiment 2, we tested the option of truncating the terms instead of discarding them ($truncate$). In both experiments, besides intonational phrase boundaries ($brk4$) and intonational plus intermediate phrase boundaries ($brk3\&4$) we also tested another cue which denotes finality - the silent pause ($pause$). Pauses were chosen as they correlate well with prosodic boundaries, but are easier to extract automatically. In order to perform a fair comparison to the case when prosodic boundaries are used, the pauses were extracted from the manual transcription and were defined as a time instant (the beginning of the pause). We considered to be a pause all silent regions of speech having a length of at least 200 ms, resulting in 2723 pause boundaries. The pause markers correspond mostly to level 4 boundaries, but there are some which indicate level 3 boundaries.

The speech signal was represented by standard spectral features: 12 MFCCs plus energy and their first and second order temporal derivatives. The system used a seed length of 250 ms, a 90 second future buffer when searching for terms and it accepted a candidate as a term if it was at least 500 ms long. A found term was represented by its median occurrence, i.e. the token closest to all the other ones in terms of a dissimilarity score and SSM checking was also employed. In the two experiments done we varied the similarity threshold used by the DTW algorithm (ϵ_{DTW}) in the range [1.8, 3.0] and we reported the results for all the values.

4.1. Experiment 1

For the baseline system, spoken term discovery was only constrained by file boundaries (403 markers in total), which were processed by the algorithm in the same manner as the prosodic boundaries (here, discarded). Figure 1 shows the percentage of terms found with the baseline system which straddle a level 4 break, a level 3 or 4 break or a pause marker. One can observe that up to 60% of the terms straddle either a level 3 or a level 4 prosodic boundary. Given that prosodic boundaries match with constituent boundaries, it is likely that the straddling terms will be less meaningful to downstream applications. In addition, it is probable that such terms are purely coincidental, and therefore correspond to low quality clusters of word fragments. We will

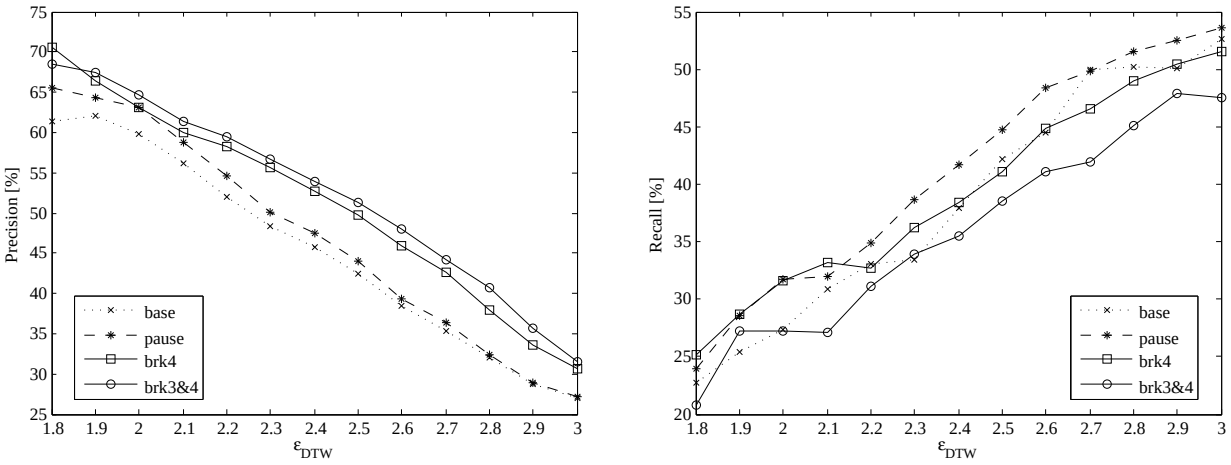


Figure 2: *Term matching precision (left panel) and recall (right panel) obtained for different ϵ_{DTW} values and various types of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

evaluate this premise next.

In order to measure the change in performance after the removal of straddling terms, compared to the baseline, we used the precision and recall, as computed with the method introduced in section 2.2. The results obtained with our baseline, and the same system employing prosodic or pause boundaries to discard straddling terms are illustrated in Figure 2. The left panel shows the precision obtained, while the right panel presents the recall rate. The baseline is represented with a dotted line, the system taking advantage of pause markers with a dashed line, and the results obtained with the prosodic boundary informations by a continuous line. The square represents the results for $brk4$, while the circle illustrates the $brk3\&4$ system.

It can be seen that, by adding extra boundary information into the system, besides file boundaries, the terms found are consistently more accurate. The average increase in precision is 6.0% for $brk4$, 7.3% for $brk3\&4$, and 1.8% for $pause$. In terms of recall, the system using intonational phrase boundaries gives similar performances to the baseline, which, in turn, performs better than the system having knowledge of both types of prosodic phrase boundaries, but worse than when pauses are known.

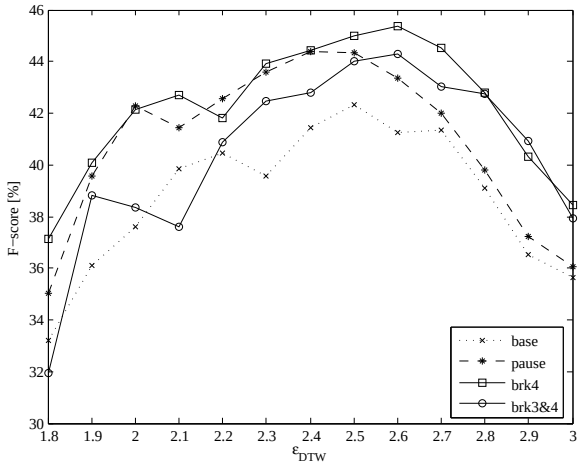


Figure 3: *Term matching F-score obtained for different values of ϵ_{DTW} and various type of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

We also looked at the F-score curve over the different ϵ_{DTW} values tested (Figure 3). When comparing the baseline with the other three systems, one sees a consistent improvement throughout the range of values investigated, resulting in an average F-score increase of 3.4% for $brk4$, 1.6% for $brk3\&4$ and 2.1% for the $pause$ system. We observed similar $brk4$ or worse $brk3\&4$ performance than for pauses at low values of the DTW threshold, but consistently better results for boundaries at high values of ϵ_{DTW}. This demonstrates that, when more heterogeneous terms are found, the boundaries tend to help discriminate better between occurrences of different terms, showing that prosodic boundaries encode more information than the pauses.

For an overall comparison of performance, we have summarized in Table 1 the two measures used for evaluation: the goodness of the obtained terms and the word boundaries discovered by the terms. The number in each cell represents the average precision, recall and F-score computed over all DTW threshold values (ϵ_{DTW}). The term matching measurements illustrated in the table mirror well the results displayed in Figures 2 and 3, showing a better performance than the baseline. Still, in terms of of word boundaries, the systems employing boundary information are penalized by a low recall and have lower F-scores, even if the word boundaries found are much more accurate.

	Term			Word Boundary		
	P	R	F	P	R	F
Baseline	45.3	38.5	38.8	22.8	2.8	4.7
Pause	47.1	40.9	40.9	24.1	2.5	4.2
Break 4	51.3	39.2	42.2	24.1	2.1	3.6
Break 3&4	**52.6**	35.8	40.4	24.4	1.7	3.0

Table 1: *Average precision, recall and F-score for discovered terms and word boundaries respectively, when the discard method is used.*

4.2. Experiment 2

In the previous experiment, we have observed that the method used for incorporating boundary information tends to persistently decrease the recall rate. This was due to the fact that the approach used consisted in discarding all terms found straddling a boundary. In this last experiment we wanted to compare this method to a different one which does not discard the terms spanning over several prosodic units, but shortens them ($truncate$). In this method, a term straddling a boundary would be truncated so it would include only the speech signal belonging to the prosodic unit having the highest overlap in time with the found term. The minimum term length constraint is applied after the truncation procedure.

	Term			Word Boundary		
	P	R	F	P	R	F
Baseline	45.2	42.4	40.3	23.0	2.9	4.9
Pause	46.5	**43.1**	41.6	26.3	**3.2**	**5.2**
Break 4	49.1	41.6	**42.4**	27.5	3.1	**5.2**
Break 3&4	51.3	38.5	41.9	**28.3**	3.0	5.1

Table 2: *Average precision, recall and F-score for discovered terms and word boundaries respectively, when the $truncate$ method is used.*

Table 2 illustrates the results obtained with this approach of incorporating prosodic information for the various types of information added. It contains the same type of information as Table 1. As we expected, the approach which truncates terms straddling a boundary gives a better recall than $discard$, at the expense of a slightly lower precision. In terms of F-score, an overall increase in performance is observed for all conditions.

The results in Table 2 show a small advantage in terms of word boundary F-score when boundary information is used. We detailed these results in Figure 4, by plotting the F-score over the range of ϵ_{DTW} values tested. It seems that for lower values of the DTW threshold, the baseline performs slightly better than the other systems, but, as the system becomes more permissive, the boundary information becomes more useful for discriminating words. Thus, it encourages the use of prosodic breaks information in conjunction with higher DTW thresholds for improved performance of UTD systems.

Interestingly, we found that the system using intonational phrase boundaries generally outperforms the system using smaller breaks (intermediate boundaries), which correspond to phonological phrase boundaries within the prosodic hierarchy [20]. Smaller breaks give better precision but this, in turn, is compensated by a worse recall which yields a slightly lower F-score, also for the $truncate$ case. This stands in contrast to findings in psycholinguistics studies where smaller breaks do seem to play a role, both for processing online speech in adults and for boosting speech segmentation in babies [11, 8, 9, 10]. We speculate that this may be due to the fact that our system imposes a 500 ms limit on the size of the discovered terms, which could affect proportionally more the breaks 3 and 4 compared to breaks 4. In order to prove this, we looked at the length of fragments delimited by boundaries. We discovered that when level 4 breaks are considered, 0.6% of the total speech fragments are shorter than 500 ms, the minimum term length employed in this study. This proportion increases to 3.7% when both level 3 and level 4 boundaries are taken into account, but it is equal to 0% for pauses. It means that for a percentage of the corpus no terms can be found. In this case, it would be interesting to investigate whether a lower limit for minimum term length will boost the recall rate, while not affecting too much the precision.

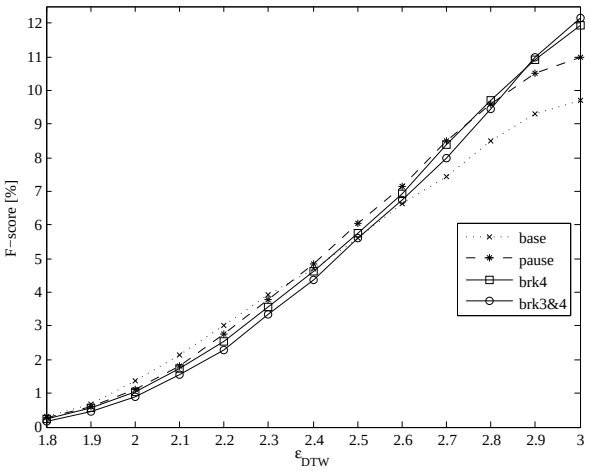


Figure 4: *Word boundary F-score obtained for different values of ϵ_{DTW} and various type of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

5. Conclusions

We have presented in this paper a preliminary study regarding the usefulness of prosodic boundaries for unsupervised term detection. Our findings show that boundary information, either intonational boundaries or intermediate and intonational boundaries, increases the performance of the system. The better results obtained are mainly due to increasingly accurate found terms, reflected in an improved precision. We have also compared prosodic boundaries against pauses, a prosodic cue which is generally easier to detect automatically. The system employing pauses outperformed the one using both level 3 and 4 boundaries, but it behaved worse than the system having knowledge of level 4 boundaries. We have discovered this advantage of pauses over level 3 and 4 boundaries to be due to a much lower recall, caused in part by constraints imposed for the length of the terms found.

The results we obtained encourage us to further continue our investigation by planning to use in a future study automatically extracted prosodic boundaries. In order to achieve this, we are currently focusing on methods of prosodic boundary detection based exclusively on acoustic cues. A second direction to pursue would be extending the study to several other languages. While unsupervised term discovery was applied until now only to less than a handful of languages we would expect prosodic boundary information to bring a consistent improvement in any language.

6. Acknowledgements

The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG) and the Fondation de France. It was also supported by ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC. The authors would like to thank Armando Muscariello for providing them with the evaluation code.

7. References

[1] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-month old infants," *Science*, vol. 274, pp. 1926–1928, 1996.

[2] A. Park and R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proc. of INTERSPEECH 2010*, 2010, pp. 1676–1679.

[4] R. Flamary, X. Anguera, and N. Oliver, "Spoken WordCloud: Clustering recurrent patterns in speech," in *Proc. of Int. Workshop on Content-Based Multimedia Index*, 2011, pp. 133–138.

[5] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2031–2044, 2012.

[6] A. Christophe, E. Dupoux, J. Bertoncini, and J. Mehler, "Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition," *Journal of the Acoustical Society of America*, vol. 95, pp. 1570–1580, 1994.

[7] A. Christophe, J. Mehler, and N. Sebastián-Gallés, "Perception of prosodic boundary correlates by newborn infants," *Infancy*, vol. 2, no. 3, pp. 385–394, 2001.

[8] P. Jusczyk, D. Kemler-Nelson, K. Hirsh-Pasek, L. Kennedy, A. Woodward, and J. Piwoz, "Perception of acoustic correlates of major phrasal units by young infants," *Cognitive Psychology*, vol. 24, no. 2, pp. 252–293, 1992.

[9] L. Gerken, P. Jusczyk, and D. Mandel, "When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases," *Cognition*, vol. 51, no. 3, pp. 237–265, 1994.

[10] A. Gout, A. Christophe, and J. Morgan, "Phonological phrase boundaries constrain lexical access II. Infant data," *Journal of Memory and Language*, vol. 51, no. 4, pp. 548–567, 2004.

[11] A. Christophe, S. Peperkamp, C. Pallier, E. Block, and J. Mehler, "Phonological phrase boundaries constrain lexical access: I. Adult data," *Journal of Memory and Language*, vol. 51, no. 4, pp. 523–547, 2004.

[12] A. Muscariello, G. Gravier, and F. Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *Proc. of INTERSPEECH 2011*, 2011, pp. 921–924.

[13] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proc. of ACL 2007*, 2007, pp. 504–511.

[14] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP 2010*, 2010, pp. 460–470.

[15] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, "MODIS: an audio motif discovery software," in *Proc. of INTERSPEECH 2013*, 2013, Software available online at `https://gforge.inria.fr/projects/motifdiscovery/`.

[16] A. Muscariello, G. Gravier, and F. Bimbot, "Variability tolerant audio motif discovery," in *Proc. of Int. Multimedia Modeling Conf. on Advances in Multimedia Modeling*, 2009.

[17] R. Daland and J. Pierrehumbert, "Learning diphone-based segmentation," *Cognitive Science*, vol. 35, no. 1, pp. 119–155, 2011.

[18] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[19] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labeling English prosody," in *Proc. of ICSLP 1992*, 1992, pp. 867–870.

[20] M. Nespor and I. Vogel, "Prosodic structure above the word," in *Prosody: Models and measurements*, pp. 123–140. Springer, 1983.

GlóRí - the *Gl*ottal *R*esearch *I*nstrument

John Dalton, John Kane, Irena Yanushevskaya, Ailbhe Ní Chasaide, Christer Gobl

Phonetics and Speech Laboratory,
School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin

jrdalton@tcd.ie, kanejo@tcd.ie, yanushei@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

Abstract

This papers presents GlóRí - the glottal research instrument. GlóRí is a speech analysis interface which offers a flexibility and multiplicity of approaches to voice analysis. The system allows for fully automatic processing, for instance for analysis of large corpora. However, for more fine-grained studies, which may require precise voice source measurements, the systems facilitates manual optimisation of parameter settings. The present paper highlights the main features of the GlóRí system and provides illustrations of the usefulness of this approach.

Index Terms: Glottal source, voice source, phonetic features, voice quality

1. Introduction

The research being carried out by the voice processing group at the Phonetics and Speech Laboratory in Trinity College Dublin is concerned with the development of robust voice source processing methods and analysing the function of the voice source in prosody. As part of this endeavour, we have been developing our speech analysis methods so as to be able to handle the inherently different acoustic characteristics of the speech signal and to be adaptive and flexible according to the phonetic and prosodic context. This paper presents GlóRí - the glottal research instrument. Note that [glo:ˌi:] is the Irish (Gaelic) word for *voices*. GlóRí is the voice analysis system within which our ongoing developments in voice source analysis and data visualisation will be integrated.

The ability to derive precise and robust measurements of the voice source is becoming increasingly important for speech technology applications (e.g., speech synthesis [1, 2], emotion classification [3, 4]), as well for linguistic analysis on the prosody of the voice [5, 6] and also for voice pathology and voice function assessment [7]. Ideally, one would wish to be able to derive robust parameterisation of the voice source completely automatically. However, this process typically involves three non-trivial steps.

The first step, in order to allow glottal pulse-synchronous analysis, is to estimate glottal closure instants (GCIs, [8]). After several decades of research state-of-the-art GCI detection is at a sufficiently high level of performance for the analysis of neutral read speech. Despite this, a recent study [9] demonstrated how different phonation types (and in particular creaky voice) deteriorated GCI detection performance. For disordered voices, where there is indeed a significant excitation within each glottal cycle, the deterioration is likely to be significantly more.

The next step, typically involves glottal inverse filtering, the process of compensating for the effect of vocal tract resonance from the speech signal. The most commonly used vocal tract model is an all-pole model, which can be estimate by linear predictive analysis, or similar methods. For nasals consonants and nasalised vowels there are generally zeros present in the vocal tract spectrum, making the all-pole model less suitable and this has negative effects on subsequent voice source parameterisation [10]. For speech with a low first formant frequency (F1), discrimination of F1 from the glottal formant can be problematic. This is particularly true when combined with a high f_0, which has the additional effect of having more widely spaced harmonics making the vocal tract spectral envelope more difficult to estimate effectively.

Finally, once an estimate of the voice source has been derived, for many purposes one typically then requires a parametric description of the signal. The two main approaches to this are to either take measurements from the voice source estimate directly or fit a mathematical model to the individual glottal pulses. Quotients characterising the timing of important events in the glottal cycle are generally thought to be the most salient parameters, regardless of which of the two approaches are used. One critical event is the instant of glottal opening, a reference point required for most time-quotient parameters. Localisation of this time instant is extremely difficult, not least because the glottal opening does not always display a significant discontinuity (e.g., in lax phonation). For the direct measure approach, the more commonly used parameters are generally amplitude-based [11, 12] or frequency domain correlates of time quotients [13].

From the three steps described above there is clearly wide scope for the introduction of significant errors, particularly in the case of natural, expressive speech or indeed even moderately disordered speech. In many instances, an experienced speech science researcher may be able to make adjustments to the automatic process and improve the overall effectiveness of the analysis. With this firmly in mind, the newly developed analysis system, GlóRí, has been designed to allow both fully automatic analysis while also facilitating manual intervention and optimisation at various stages in the analysis. From the outset there are three main characteristics that are fundamental to our system design.

1. **Adaptive** The system should allow a multiplicity of approaches, e.g., for research on large corpora fully automatic analysis can be deployed, but for more fine-grained analysis the researcher should be able to manually optimise the analysis to ensure maximal precision. Furthermore, the analysis should be adaptive to the phonetic and prosodic context, e.g., allowing glottal inverse filtering with an adaptive vocal tract modelling.
2. **Modular** Ongoing development of various voice source and speech analysis algorithms should be easy to incor-

porate into the system. To this end we created the interface in the Matlab programming environment. As our algorithm development (as well as much of the signal processing development in the speech research community) is done using this environment, it facilitates newly developed algorithms being easily incorporated.

3. **Knowledge** We intend for the system to incorporate various sources of *knowledge* in the analysis. This can be considering the given phonetic class being analysed and adapting the analysis accordingly (e.g., using a vocal tract model which included pole-zeros for nasal regions). It can also include incorporating knowledge from speech production theory, e.g., precluding parameterisation which is outside the physical boundaries of human speech.

1.1. Existing voice source analysis systems

There are a small number of interfaces for voice source analysis available in the literature. The APARAT system is one interface which facilitates automatic glottal inverse filtering and voice source parameterisation using a range of existing parameters from the literature [14]. The system is available under an open-source licence[1] and has encourage using voice source feature extraction in a range of speech-related areas.

Another freely available interface for voice analysis is the Voice Sauce program[2] [15]. Voice Sauce enables a wide range of voice-related speech analysis including f_0 and harmonic extraction, formant tracking and the formant compensation proposed by Hanson [13], harmonic and subharmonic to noise ratio, energy and cepstral peak prominence extraction. Voice sauce also includes a facility for analysis of electroglottographic (EGG) waveforms.

Although there is some overlap with these interfaces, GlóRí is a useful complement and there are several major differences compared with existing systems. First, GlóRí allows manual intervention and optimisation. Second, GlóRí includes some very recently developed voice quality related analysis methods. A third major difference is that we are intending for the system to allow incorporation of speech production knowledge and to involve pre-processing steps which could be used to contrain possible analysis settings and also, where necessary adapt the analysis (e.g., the structure of the vocal tract model) to more closely match the acoustic structure of the given speech segment. Furthermore, GlóRí includes resynthesis and data vizualisation components that facilitate construction of stimuli for perception experiments as well as allow to represent the analysed data for visual inspection in a number of ways.

2. System features

This section serves to illustrate the main system features of the GlóRí system. The system was designed to be user-friendly and to allow manual analysis, if it is deemed necessary, or completely automatic voice source feature extraction.

2.1. Manually-optimised analysis

Voice source analysis, including the possiblity for manually-optimised analysis, can be carried out using the analysis window shown in Figure 1. When a speech sample is loaded into

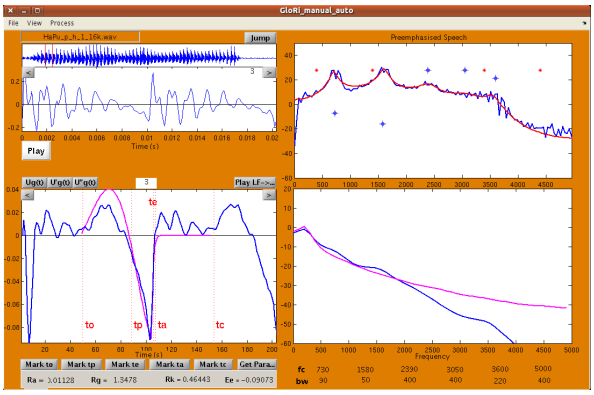


Figure 1: Screenshot of the manual analysis interface of the GlóRí system. Users can choose from a selection of parameterisation approaches, derived from the voice source estimate, using parameters derived from model-fitting and parameters derived from the speech signal.

system, it is resampled to 10 kHz. GCIs are located automatically using our recently developed algorithm (SE-VQ, [9]) and GCIs detected in unvoiced regions are excluded. GCI locations can then be manually edited later if required using the GCI editor. Locations that are judged to be false can be deleted and undetected locations can be added. Although state-of-the-art GCI detection has reached a mature level of performance, this can still degrade when analysing speech involving wide variation in phonation type [9] or the voice is disordered. Allowing a facility for manual intervention here may enable more precise analysis for these types of speech.

For each GCI-centred two pulse length frame, the vocal tract model can be constructed by setting the formants frequencies and bandwidths. The frequency and bandwidth of each formant can be adjusted using the keyboard arrow keys, and a time and frequency domain representation is available to assess the effect of the inverse filter. As each anti-formant nears its optimal location, the oscillations of the corresponding formant will be dampened in the time domain (see bottom left panel of Figure 1), and the formant peak will be largely attenuated in the frequency domain (see bottom right panel). Once the speech signal has been inverse filtered the user can then move to the parameterisation step. The manually optimised system allows parameterisation of the estimated voice source signal by fitting the Liljencrants-Fant (LF) voice source model [16] to the individual glottal pulses. An LF model can be fitted to the inverse-filtered pulse by manually adjusting the time-points of the model (see bottom left panel of 1). Fitting is facilitated with both the time and frequency domains displays in the two adjacent panels, allowing the user to achieve accurate time-point matches while also ensuring close spectral fitting.

2.2. Fully automatic analysis

In contrast to the manually-optimised analysis, a fully automatic analysis approach is included in the GlóRí system (see Figure 2). The analysis relies entirely on the use of automatic algorithms, without any intervention from the user. A folder of speech samples is loaded through the interface, and the desired analysis parameters are selected.

These fall under three categories. The category title "Glottal params" is further subdivided into two different types. Un-

[1] http://sourceforge.net/projects/aparat/
[2] http://www.ee.ucla.edu/~spapl/voicesauce/

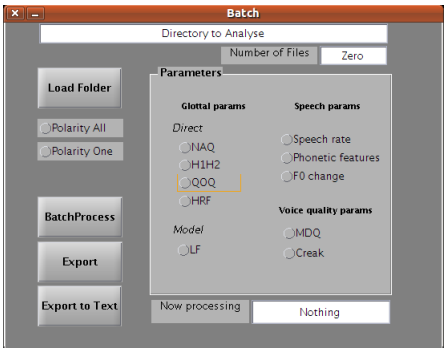


Figure 2: Screenshot of the fully automatic analysis interface of the GlóRí system. Users can choose from a selection of parameterisation approaches, derived from the voice source estimate, using parameters derived from model-fitting and parameters derived from the speech signal.

der "Direct" one can select parameters which are derived using direct measurements of the glottal inverse filtered signal. These include: the normalised amplitude quotient (NAQ; [12]), the difference between the amplitude of the first two harmonics (H1-H2; [13]), the quasi-open quotient (QOQ; [11]), and the harmonic richness factor (HRF; [19]). Under "Model" one can select to have the glottal inverse filtered signal parameterised by fitting LF-model pulses to the individual glottal pulses using our recently developed automatic fitting algorithm [20]. Note that prior to estimation of these parameters, the inputted speech signal is inverse filtered using iterative and adaptive inverse filtering (IAIF, [21]).

Under the category "Voice quality parameters", one can select the maxima dispersion quotient (MDQ; [22]) and Creak [23]. MDQ is a wavelet based algorithm which discriminates breathy and tense voice quality by assessing the dispersion of peaks across a range of frequency bands relative to the GCI. The Creak parameter gives the binary output of a decision tree classifier, using two input features derived from the Linear Prediction (LP) residual signal.

Finally, the "Speech params" category allows selection of parameters related directly to aspects of the speech signal. Phonetic feature extraction selected and this outputs a continous score on the likelihood of the presence of a range of phonetics features {voiced, syllabic, nasal, liquid, fricative, plosive}. This is done using the algorithm recently proposed in [24]. Note that this algorithm provides important information on the underlying manner of articulation is various speech regions which can facilitate analysis strategies which are adaptive to the phonetic context. This algorithm can also be harnessed for deriving a 'speech rate' measurements, in terms of syllables per second.

2.3. Synthesis interface

Once a given speech signal has been analysed, using either manual or automatic methods, one can then load the exported analysis file into a synthesis interface. As shown in Figure 3 a user is provided with parameter contour displays. The user can modify parameter contours by clicking new points on the panel, as has been done for the f_0 parameter (top panel). It is then possible to resynthesise the speech using the modified parametric setting. This is a useful facility for stimuli generation to be used in perception experiments.

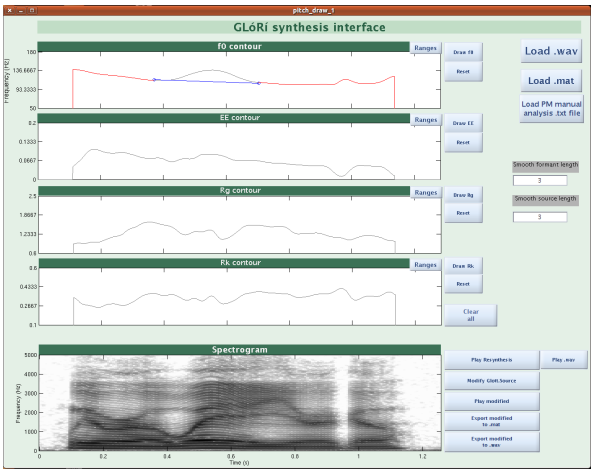


Figure 3: Screenshot of the synthesis interface of the GlóRí system. Users are shown display of parameter settings and can make alterations to these contours and resynthesise the speech.

2.4. Visualisation interface

An interface is also included for easy visualisation of extracted parameter contours. By loading in an analysis file, again either using manual or automatic systems, one can select combinations of parameters to be plotted together with the speech spectrogram. We have also begun to experiment with novel visualisation approaches for showing high dimensional parameter data in a clear single plot, e.g., using spidergrams (illustrated below). These novel developments are incorporated within this interface component of the GlóRí system.

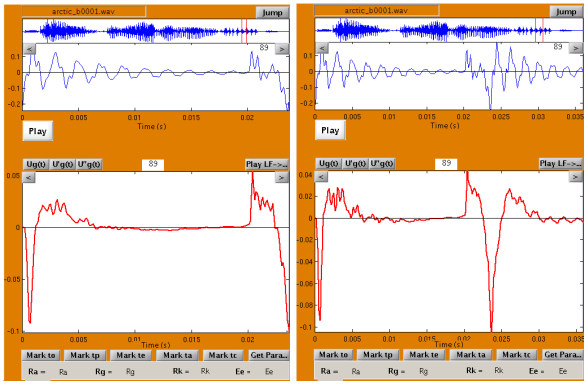
3. Illustrations

This section serves to provide illustrations of how the system features of GlóRí may be beneficial for a range of analysis purposes.

The first illustration highlights the importance of allowing manual intervention to improve the precision of the analysis in certain cases. In the left panel of Figure 4 one can observe the negative impact of a false positive GCI, as often occurs in creaky voice (see [9]), on the overall analysis. The main glottal excitation should be located close to the centre of the analysis panel, and, hence, the false positive observed here will preclude the possiblity of obtaining sensible parameter values if a completely automatic approach was used. However, by exploiting the ability for manual intervention, in this case facilitated by the manual GCI editor, one can easily delete this false positive and proceed to effective voice source modelling as shown in panel (b) of Figure 4.

One crucial intention in the development of the GlóRí system is to facility incorporating knowledge (in its various forms) to help constrain and augment the analysis. In particular, it is desired to facilitate analysis that is sensitive and adaptive to the phonetic environment of the speech signal. As mentioned previous, our phonetic feature extraction algorithm can provide us with initial information on the underlying manner of articulation in a given utterance [24]. Figure 5 shows the output of this feature extraction for a sample utterance.

The information yielded by these phonetic feature extractors may be beneficial for a range of purposes in the analysis



(a) GCI false positive (b) Corrected GCI false positive

Figure 4: Screenshot of analysis of a creaky voice pulse involving a false positive GCI (a) and with the false positive corrected (b).

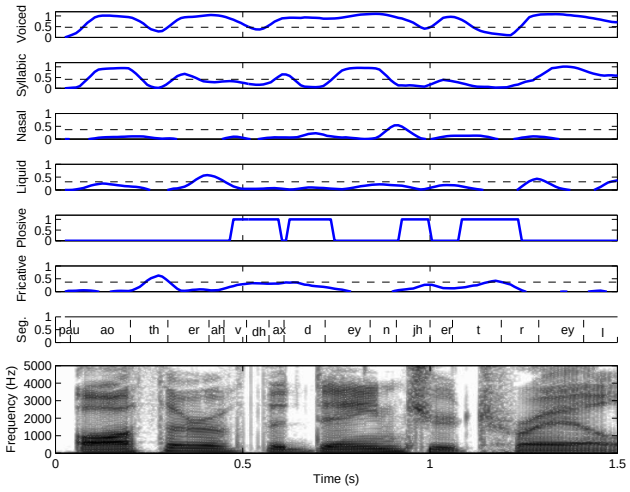


Figure 5: Illustration of phonetic feature extraction for the utterance “Author of The Danger Trail ...”, spoken by an American male.

interface. For voice source feature extraction for the purpose of voice quality classification one may choose to exclude certain phonetic regions known to cause problems for analysis (e.g., areas of frication, or nasality). Another usage could be for adaptive glottal inverse filtering, where the vocal tract model could be adapted to the given region, e.g., in nasalised regions pole-zeros could be incorporated into the vocal tract model. Besides these uses, the phonetic feature extraction may also provide a useful guide for the researcher when carrying out manually-optimised voice source analysis.

Another important component of the GlóRí system is data visualisation. A frequently used approach when analysing voice source parameters is to reduce the data, often to a single shape parameter. However, this approach may at times be premature and may involve losing important information to do with the glottal pulse shape. In order to avoid premature data reduction and to display the voice source parameter data in an accessable form the GlóRí system allows plotting of the data as a “spider-

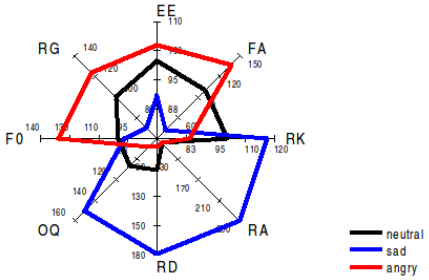


Figure 6: Spidergram plot summarising high dimensional voice source data into a single plot. Note that the axes are the parameter values in percentage relative to neutral.

gram” as shown in Figure 6 [4]. In the spidergram, parameters are arranged in such a way that increased parameter levels above the horizontal axis typically indicate a tenser phonation. Similarly, levels extending below the horizontal axis point to a laxer phonation. The illustration in Figure 6 shows an example spidergram for a sentence spoken with three types of affective colouring: neutral, sad and angry. The blue web for *sad* with it’s increased parameter levels below the horizontal line provides strong evidence of a laxer phonation type, whereas *angry* (with the red web) indicates a tenser phonation.

4. Discussion & conclusion

This paper presented the new voice analysis system, GlóRí. The system is shown to facilitate completely automatic voice source feature extraction, and incorporates a range of state-of-the-art voice source analysis developments as well as existing parameters from the literature. The automatic system may be extremely useful for studies across a range of speech-related disciplines when analysing large corpora, and could in particular be useful for allowing voice source feature extraction for researchers from a non-technical or non-voice related background.

A further benefit of the GlóRí system over existing analysis systems, is that it facilitates manually-optimised analysis. This may be critical for very fine-grained analysis studies which require precise voice source parameter data. Manual intervention here may help reduce the effect of error introduction in the various stages of analysis.

The GlóRí system is intended to be a constantly work-in-progress development. One main direction for ongoing and future research is to bring to bear our knowledge of speech production so we can constrain possible vocal tract model and voice source parameterisation solutions. Our newly developed fully automatic techniques (for instance for deriving information to do with breathy, tense and creaky voice, as well as the underlying phonetic features) can provide prior information that can be used to constrain vocal tract filter and voice source modelling. We intend to make this system publicly available in the near future.

5. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET) and the Irish Department of Arts, Heritage and the Gaeltacht (ABAIR project).

6. References

[1] Cabral, J., Renals, S., Richmond, K., Yamagishi, J., (2011) "HMM-based speech synthesiser using the LF-model of the glottal source", Proceedings of ICASSP, Prague, Czech Republic, 4704-4707.

[2] Degottex, G., Lanchantin, A., Roebel, A., Rodet, X., (2012) "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis", Speech Communication, 55(2), 278-294.

[3] Lugger, M., (2007) "The relevance of voice quality features in speaker independent emotion recognition", Proceedings of ICASSP, Hawaii, USA, 17-20.

[4] Yanushevskaya, I., Gobl, C., Ní Chasaide, A., (2009) "Voice parameter dynamics in portrayed emotions", Proceedings of Maveba, 21-24.

[5] Yanushevskaya, I., Gobl, C., Kane, J., Ní Chasaide, A., (2010) "An exploration of voice source correlates of focus" Proceedings of Interspeech, Makuhari, Japan, 462-465.

[6] Ní Chasaide, A., Yanushevskaya, I., Gobl, C., (2011) "Voice source dynamics in intonation" Proceedings of ICPhS, Hong Kong, 1470-1473.

[7] Gómez-Vilda, P., Fernández-Baillo, R., Rodellar-Biarge, V., Lluis, V. N., Álvarez-Marquina, A., Mazaira-Fernández, L., M., Martínez-Olalla, R., Godino-Llorente, J. I., (2009) "Glottal Source biometrical signature for voice pathology detection", Speech Communication 51(9), 759-781.

[8] Naylor, P., Kounoudes, A., Gudnason, J., Brookes, M., (2007) "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm" IEEE Transactions on Audio Speech and Language processing, 15(1), 34-43.

[9] Kane, J., Gobl, C., (2013) "Evaluation of glottal closure instant detection in a range of voice qualities", Speech Communication 55(2), 295-314.

[10] Gobl, C., Mahshie, J., (2013) "Inverse filtering of nasalized vowels using synthesized speech", Journal of Voice, 27(2), 155-169.

[11] Hacki, T., (1989) "Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie" Folia Phoniatrica, 43-48.

[12] Alku, P., Bäckström, T., Vilkman, E., (2002) "Normalized amplitude quotient for parameterization of the glottal flow" Journal of the Acoustical Society of America, 112(2), 701-710.

[13] Hanson, H. M., (1997) "Glottal characteristics of female speakers: Acoustic correlates" Journal of the Acoustical Society of America, 10(1), 466-481.

[14] Airas, M., (2008) "TKK Aparat: An environment for voice inverse filtering and parameterization" Logopedics Phoniatrics Vocology, 33, 49-64.

[15] Shue, Y-L, Keating, P., Vicenik, C., Yu, K., (2011) "Voice sauce: a program for voice analysis" Proceedings of ICPhS.

[16] Fant, G., Liljencrants, J., Lin, Q., (1985) "A four parameter model of glottal flow" KTH, Speech Transmission Laboratory, Quarterly Report, 4, 1-13.

[17] Gobl, C., (1988) "Voice source dynamics in connected speech", KTH, Speech Transmission Laboratory, Quarterly Report, 29, 123-159.

[18] Walker, J., and Murphy, P., (2007) "A review of glottal waveform analysis" *in* Progress in nonlinear speech processing, 1-21.

[19] Childers, D. G., Lee, C. K., (1991) "Voice quality factors: Analysis, synthesis and perception", Journal of the Acoustical Society of America, 90, 2394-2410.

[20] Kane, J., Gobl, C., (2013) "Automating manual user strategies for precise voice source analysis", Speech Communication 55(3), 397-414.

[21] Alku, P., (1992) "Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering", Speech Communication, 11(2-3), 109-118.

[22] Kane, J., Gobl, (2013) "Wavelet maxima dispersion for breathy to tense voice discrimination", IEEE Transactions on Audio, Speech and Language Processing, 21(6), 1170-1179.

[23] Kane, J., Drugman, T., Gobl, (2013) "Improved automatic detection of creak", Computer Speech and Language, 27(4), 1028-1047.

[24] Kane, J., Aylett, M., Yanushevskaya, I., Gobl, [Under review] "Phonetic feature extraction for context-sensitive glottal source processing", Speech Communication.

Speech segmentation is modulated by peak alignment: Evidence from German 10-month-olds

Bettina Braun, Muna Pohl & Katharina Zahner

Department of Linguistics, University of Konstanz, Germany
{bettina.braun, muna.pohl, katharina.zahner}@uni-konstanz.de

Abstract

In two headturn preference experiments, we tested whether German infants' speech segmentation skills are sensitive to the position of the pitch peak relative to the stressed syllable. Specifically, we compared target words with medial-peak accents (where the pitch peak is aligned with the stressed syllable, i.e. H* accents) and early-peak accents (where the pitch peak is early with respect to the stressed syllable, i.e. H+L* accents). Such differences in accent type signal mostly pragmatic distinctions in German, such as the difference between contextually new and recoverable information. We familiarized infants with target words produced with one of the two intonation conditions; target words were embedded in sentences. We measured looking times to lists of trochaic part-words that were either embedded in the target words or were novel to them. Results showed a novelty effect only in the medial-peak condition, suggesting that German infants at 10 months of age are very sensitive to pitch information for segmenting running speech.

Index Terms: speech segmentation, pitch, metrical prominence, headturn preference procedure, German

1. Introduction

The ability to segment units from continuous speech is a vital skill for infants to build up their mental lexicon. Research has shown that the segmentation skills in the second half of the first year of life correlate positively with later vocabulary size [1, 2]. The focus of our investigation is on infants' use of metrical stress for segmentation. Previous findings suggest that stressed syllables are preferred word onsets for infants whose mother tongue is a stress-timed language (English [3, 4], German [5], Dutch [6]), but not for infants who are exposed to syllable-timed languages such as French [7]. Although the exact timeline has not been determined yet, it seems that infants shift their attention from statistical to prosodic cues and that by 9 months of age stress cues outweigh distributional cues in segmentation [8]. In stress-timed languages, metrical prominence is generally signaled by a wide variety of acoustic cues: Stressed syllables are longer and louder than unstressed syllables [9, 10], they are produced with increased vocal effort [11] [see 12 for an overview] and often have more peripheral vowel qualities [13]. When stressed syllables additionally receive phrase-level prominence (pitch accents), they are additionally produced with audible pitch movement. However, in intonation languages, the actual type of pitch accent (and hence the alignment of the pitch peak with regard to the stressed syllable) can vary. For instance, the pitch peak can be early, medial or late with respect to the stressed syllable, resulting in distinct accent types [14, 15]. These are described as H+L*, H* or L*+H in the framework of autosegmental metrical phonology [16, 17]. Factors that determine the realization of pitch accents are, among others, utterance position (prenuclear accents are realized with later peaks than nuclear accents, cf. [18]) as well as the information status of the accented referent (non-identifiable referents are preferably associated with medial-peak accents and identifiable referents with early-peak accents, cf. [19]) and information structure (new information is associated with medial-peak accents, inferable information with early-peak accents, cf. [14]). Hence, while a prominence-lending pitch movement in general is a salient cue for metrical stress, the exact position of the pitch *peak* is not a reliable indicator for the position of the stressed syllable in a given word. It must be pointed out that the distinction between early, medial and late pitch peaks is not only signaled by the position of the pitch peak but also in the duration and intensity distribution over the stressed and neighboring unstressed syllables [20]. Specifically, in early-peak accents, the contrast in duration and intensity between the pre-accented (high-toned) syllable and the accented (low-toned) syllable was reduced compared to medial-peak accents. [21] further showed that listeners also use duration and intensity for pitch accent type interpretation. These changes in the intensity and duration distribution suggest that it is more difficult to identify the stressed syllable in early-peak accents than in medial-peak accents. In the present study we investigate if peak alignment affects German 10-month-old infants' segmentation of trochaic part-words (which are salient to young infants) as a function of peak position (early vs. medial peak accents).

High pitch may play a crucial role in segmentation as infants are highly sensitive to pitch information. Specifically, they show a strong listening preference for f0-patterns in infant-directed speech over amplitude and duration patterns [22] and the preference for high pitch and expanded melodic contours is present very early in infancy [23-25]. [26] suggest that this early sensitivity to pitch information might be caused by its phonetic salience and by its distribution across languages. Regarding salience, [27] hypothesize that acoustic salience affects speech perception and infants indeed seem to be especially salient to pitch contrasts, even earlier than to other acoustic features, such as duration [28]. Regarding the distribution of pitch contrasts, [29] claims that melodic contours are present in all languages, either in form of lexical tone, lexical pitch accent or intonation. It has been shown that such common and salient acoustic distinctions as falling vs. rising pitch contours are discriminated early in infancy [26].

Previous segmentation studies have not manipulated peak alignment when familiarizing infants with test words. It can be assumed that the infant-directed speech samples used in these studies have a high proportion of medial-peak accents [30], but there is not enough information in the literature to know whether peak alignment modulates infants' ability to segment words or part-words that start with metrically strong syllables. In the present study, we used the headturn preference procedure [3] with a familiarization phase, in which infants hear two out of four possible passages containing the target words. In the test phase, infants listen to lists of isolated words, half of which already occurred in the familiarization

phase and half of which were novel to the infants. The crucial manipulation was the intonational realization of the target words in the familiarization phase: In Experiment 1, the target words were presented with a medial-peak accent, in Experiment 2, they were presented with an early-peak accent. We hypothesize that infants associate high pitch with metrical prominence and hence show different looking times to familiar and unfamiliar test lists only in the medial-peak condition and not in the early-peak condition.

2. Experiment 1

Experiment 1 investigated whether infants are able to segment embedded trochees from continuous speech when the stressed syllable carries the pitch peak. The target words were presented with a medial-peak accent in the familiarization phase. This is the intonational realization that was most likely used in earlier segmentation studies [3] and we expect to see differences in looking times to familiar and unfamiliar part-words. The target words were all trisyllabic with penultimate stress. This word-prosodic structure was chosen as it allowed us to present the pitch accent on the target word in both intonation conditions (in disyllabic trochaic words, the pitch peak would have been placed on different other words in the early-peak condition, as the target words were presented within whole sentences in the familiarization phase). The trochaic part of the target words (i.e., the last two syllables) served as test words, since infants are very good at segmenting trochees from running speech (see Introduction). Thus, compared to other segmentation studies in which infants were tested on their ability to detect trochees in fluent speech (e.g., [2]), the task demands are more challenging here: Since infants have to extract *embedded* trochees, they cannot rely on acoustic cues to word onsets, and statistical co-occurrences are not helpful either in that they would suggest to extract the complete trisyllabic target word. While the lexical activation of embedded words in adult listeners (e.g., *date* from *sedate* or *bone* in *trombone*) is often disputed [31], the task is different for young infants who are not affected by lexical competition (and inhibition) in the same way as adults are.

2.1. Methods

We conducted a headturn preference study, in which infants were familiarized with target words that were embedded in sentences (see [3]). In the test phase, infants heard word lists consisting of isolated disyllabic trochees that were either part-words of the target words in the familiarization phase or not.

2.1.1. Participants

We tested 21 infants between 37 and 41 weeks from monolingual German-speaking homes and who had not been exposed to languages other than German. All infants were born full-term. For the analysis, we could only include those 16 infants (7 male, 9 female) who finished the familiarization phase and all 12 test trials. They had an average age of 38.9 weeks (SD = 1.26 weeks). Parents were reimbursed for public transport fees and received a small present for the child.

2.1.2. Materials

We chose four low-frequency trisyllabic words (less than 0.1 occurrences per million in the CELEX word form dictionary, cf. [32]) with open syllables as target words (*Kanone* [kaˈnoːnə], 'cannon'; *Lagune* [laˈguːnə], 'lagoon'; *Kasino* [kaˈsiːno], 'casino'; *Tirade* [tiˈʁaːdə], 'tirade'). All were stressed on the second syllable. Note that in all these target words, the unstressed initial syllable is not reduced to [ə] but produced with a full vowel. For each target word we constructed six carrier sentences, such that the target word appeared in different lexical contexts and different sentence positions (twice each in sentence-final position, four times early in the sentence following an article or prenominal adjective). We used naturally produced auditory stimuli in our experiments since the perception of stress is influenced by a variety of acoustic cues that are distributed over the stressed and neighboring unstressed syllables [e.g., 33]. A native female speaker of German recorded the 24 target sentences with a medial-peak accent on the target words. To achieve equally salient f0-movements across target words, the average f0-excursion was matched across the four sets of target words (average f0-excursion was 9.34st, SD = 1.51st). The pitch contour of an example sentence is provided in Figure 1.

The four test words consisted of the second and third syllable of the target words: [ˈguːnə] (taken from *Lagune*), [ˈsiːno] (taken from *Kasino*), [ˈnoːnə] (taken from *Kanone*) and [ˈʁaːdə] (taken from *Tirade*). These disyllabic trochees were recorded approximately 30 times. For the experiment we chose 15 items of each disyllable, such that the average f0-excursion of the pitch fall and the average duration did not differ across test words (average f0-excursion = 10.0st, SD = 1.8st, average duration = 791ms, SD = 72ms). Two consecutive tokens in the test lists were separated by 800ms silence.

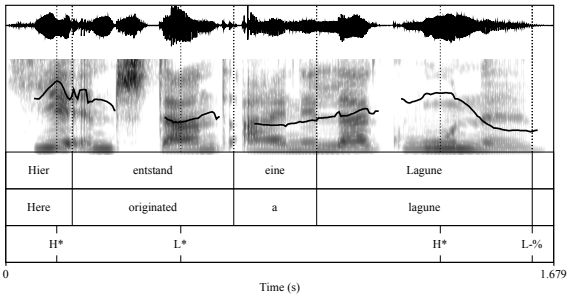


Figure 1: *Example pitch track of a target sentence in the medial-peak condition (f0-range is shown between 100 and 300Hz).*

2.1.3. Procedure

Half of the infants were assigned to the *Kanone* and *Tirade* familiarization trials and the other half to the *Kasino* and *Lagune* familiarization trials. All infants listened to three randomized repetitions of the four test lists (twelve test trials in total). Infants were tested in a three-sided experimental booth at the University of Konstanz. The infants sat on a caregiver's lap. The caregivers wore headphones and did not hear the auditory stimuli the infants were exposed to. Each trial started with a green blinking light at the center of the screen. As soon as the infant oriented towards the center light, a red light to the right or left of the child started blinking. When the infant turned his/her head towards the sidelight, an auditory word list started playing. It played as long as the infant oriented towards this side. If the child looked away for more than 2 seconds, the next trial started. In the familiarization phase, the two passages were presented randomly from the left or the right side until the child had

listened to each of the two paragraphs for at least 45 seconds. In the test phase also, the word lists played randomly from the left or the right side. Looking times were coded online by an experimenter who monitored the child via a video camera and controlled the experiment. The experimenter wore headphones and was not aware of the condition that was played (familiar or novel word list). Prior to testing, parents filled in a questionnaire regarding the infant's language background. The experimental session lasted approximately 5 minutes.

2.2. Results

Looking times in seconds were averaged by familiarity condition for each infant. Log-normalized looking times were analyzed using linear-mixed effects regression models with *familiarity* (familiar vs. novel) as fixed factor and *participants* and *lexicalization* as crossed random factors, allowing for random adjustments of intercepts and slopes for within-group factors [34, 35]. Data points with residuals beyond 2.5 SD of the mean were removed and the model was refitted. p-values were calculated on the basis of model comparisons, using the anova()-function in R [34]. In other words, only when a model with a given main effect improved compared to a simpler model without that effect, the main effect was considered to be significant. Results showed a significant effect of *familiarity* (ß = 1.13, SE = 0.05, t = 2.3, $p < 0.05$), see left-hand bars in Figure 3. Participants fixated on average 1.1 seconds longer to novel test lists (9.3 sec) than to familiar ones (8.2 sec).

2.3. Discussion

Participants looked significantly longer to the novel than to the familiar word lists, suggesting that they were able to segment the trochaic part-words from the trisyllabic words that were used in the familiarization phase. German 10-month-olds can hence segment trochaic syllable sequences from fluent speech even if they are embedded in trisyllabic carrier words, and thus contain neither probabilistic nor acoustic cues to word onsets. Our findings replicate earlier studies on the metrical segmentation strategy in German and English [3-5]. Importantly, we found evidence for segmentation with familiarization and test stimuli that had comparatively small f0-excursions. This suggests that infants are able to segment strings from fluent speech even when exposed to adult-directed speech.

3. Experiment 2

Experiment 2 examined whether a misaligned pitch peak affects infants' segmentation of embedded trochees.

3.1. Methods

3.1.1. Participants

Another group of 19 infants took part in Experiment 2. They fulfilled the same criteria as the infants in Experiment 1. The data of 16 infants could be analyzed (9 male, 7 female, mean age = 38.7 weeks, SD = 1.18). There was no significant difference in age across experiments ($t(30) = 0.3$, $p > 0.7$).

3.1.2. Materials

The four test words were identical to Experiment 1. The same speaker as in in Experiment 1 recorded the 24 sentences again, this time with an early-peak accent on the target word (Fig. 2).

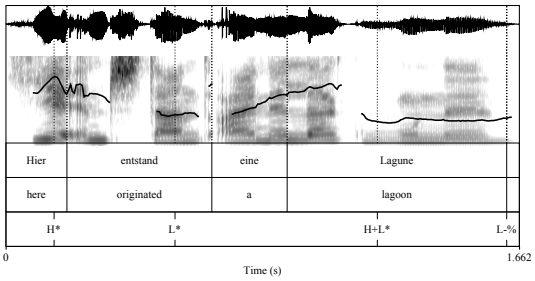


Figure 2: *Example pitch track of a target sentence in the early-peak condition (f0-range is shown between 100 and 300Hz).*

The target sentences were recorded until the productions did not differ from the recordings in the medial-peak condition for Experiment 1 in terms of f0-excursion of the pitch fall, total duration of the target word, the duration of the stressed syllable and the onset consonant of the stressed syllable (mean values are shown in Table 1). Also, the intonation conditions did not differ with respect to the spectral realization of the vowels in the first and second syllable (measured in terms of the Euclidian distance in the F1/F2 space between a particular vowel and the average over all 15 [ə]s in the target words, see (1)).

$$(1)\ \sqrt{(F1_{vowel} - F1_{[ə]})^2 + (F2_{vowel} - F2_{[ə]})^2}\ ).$$

Yet, since the target words were all naturally produced, there are some differences in the materials of the two experiments, mirroring the effects of early- and medial-peak accents described in [20].

3.1.3. Procedure

The procedure was the same as in Experiment 1.

3.2. Results

Looking times were analyzed in the same way as for Experiment 1. Results showed no effect of familiarity ($p > 0.8$), see right-hand graphs in Figure 3.

To statistically corroborate the differences across experiments, we calculated a single linear mixed effects regression model with *intonation condition* (early vs. medial) and *familiarity* as fixed factors and *participants* and *lexicalization* as crossed random factors. The results of this model showed no main effects, but a significant interaction between *intonation condition* and *familiarity* (ß = 0.16, SE = 0.08, t = 2.2, $p < 0.05$).

3.3. Discussion

Looking times did not differ significantly for familiar and unfamiliar disyllables in Experiment 2, in which high pitch was misaligned with the stressed syllable. This suggests that in this intonation condition infants were not able to segment the part-words starting with a stressed syllable. Taken together, the results of the two experiments indicate that trochees could only be successfully extracted from the speech stream when the pitch peak was aligned with the stressed syllable but not when the peak preceded the stressed syllable.

Table 1. *Acoustic realization of target words in the familiarization phase for both intonation conditions.*

Acoustic variable	medial-peak condition	early-peak condition	p-value (paired t-test, df =23)
F0-excursion of the pitch fall	9.3st	9.6st	n.s.
Duration of 1st, unstressed syllable	182ms	193ms	$p < 0.005$
Duration of 2nd, stressed syllable	253ms	256ms	n.s.
Duration of onset consonant of 2nd, stressed syllable	81.3ms	80.6ms	n.s.
Intensity in middle of initial vowel	67dB	66dB	n.s.
Intensity in middle of stressed vowel	71dB	64dB	$p < 0.001$
Euclidean distance of 1st vowel from [ə]	293.3Hz	307.8Hz	n.s.
Euclidean distance of 2nd vowel from [ə]	837.9Hz	737.7Hz	n.s.

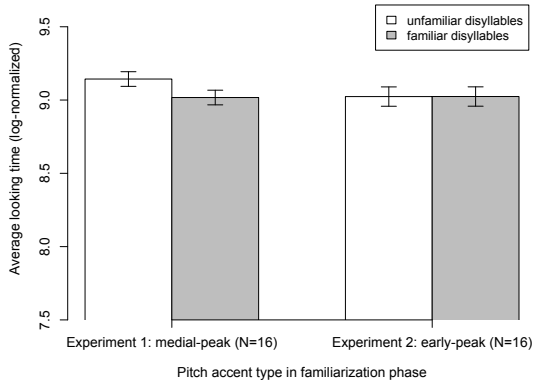


Figure 3: *Average looking time in the medial- and early-peak condition split by familiarity. Whiskers represent standard errors*

4. Discussion and Conclusion

The present study shows that German infants at 10 months of age are able to segment trochaic part-words from trisyllabic target words. However, this is only possible when the familiarization phase provides target words in which the pitch peak is aligned with the stressed syllable (medial-peak condition in Experiment 1) and not when the peak precedes the stressed syllable (early-peak condition in Experiment 2). We see four possible explanations for this finding. First, under the assumption that German 10-month-olds treat stressed syllables as preferred word onsets [36], stress perception appears to be partly linked to high pitch. When the stressed syllable carried a low tone, the segmentation of the trochaic part-words failed. This may be interpreted as an effect of cue strength, in analogy to findings showing that infants are able to distinguish different phrasings only when the prosodic phrase boundary is signaled by phrase-final lengthening and pitch movement, but not, when one of these cues is missing [37]. Second, recent research within the framework of the iambic/trochaic law [38, 39] showed that Italian infants at 7 months group high-low sequences as trochaic patterns while ignoring information on duration [28], suggesting that intonation may be a stronger cue to metrical stress than other stress cues. Third, high-pitched syllables may be likely word-onset markers per se, irrespective of their perceived metrical prominence. Fourth, our effects may not be related to the acoustic salience of pitch at all but may be explained in terms of frequency of occurrence. In infant-directed speech medial peaks are more frequent than early peaks [30]. Thus, infants' segmentation abilities displayed in the medial peak condition might correlate with the relative frequencies in the input.

To investigate the role of pitch on the perception of metrical stress more closely, we test infants' attention to the first two syllables of the target words, produced with a trochaic stress pattern (e.g., ['la:gu] for *Lagune*, ['ka:no] for *Kanone*). If high pitch is the most salient cue to metrical prominence, an effect of familiarity is expected only in the early-peak condition, but not in the medial-peak condition. Furthermore, we plan to use resynthesized stimuli to be able to focus on the role of f0-information in speech segmentation.

5. Acknowledgements

We thank Jana Schlegel and Sophie Egger for recording, data analysis and testing. We acknowledge support from an AFF research grant from the University of Konstanz.

6. References

[1] Singh, L., Reznick, S., and Xuehua, L., "Infant word segmentation and childhood vocabulary development: a longitudinal analysis", Developmental Science, 15(4):482-495, 2012.

[2] Newman, R., Ratner, N.B., Jusczyk, A.M., Jusczyk, P.W., and Dow, K.A., "Infants' Early Ability to Segment the Conversational Speech Signal Predicts Later Language Development: A Retrospective Analysis", Developmental Psychology, 42(4):643-655, 2006.

[3] Jusczyk, P., Houston, D.M., and Newsome, M., "The beginnings of word segmentation in English-learning infants", Cognitive Psychology, 39:159-207, 1999.

[4] Jusczyk, P.W. and Aslin, R.N., "Infants' detection of the sound patterns of words in fluent speech", Cognitive Psychology, 29:1-23, 1995.

[5] Bartels, S., Darcy, I., and Höhle, B., "Schwa syllables facilitate word segmentation for 9-month-old German-learning infants", in J. Chandlee, et al., [Eds], BUCLD 33: Proceedings of the 33rd Annual Boston University Conference on Language Development, 73-84, Cascadilla Press: Somerville M.A., 2009.

[6] Kooijman, V., Hagoort, P., and Cutler, A., "Prosodic structure in early word segmentation: ERP evidence from Dutch ten-month-olds", Infancy, 14:591-612, 2009.

[7] Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., and Alcantara, C., "Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences", Journal of Memory and Language, 54:283-299, 2006.
[8] Thiessen, E.D. and Saffran, J.R., "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month old infants", Developmental Psychology, 39:706-716, 2003.
[9] Jessen, M., Marasek, K., Schneider, K., and Clan, K. "Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German", in Proceedings of the 13th International Congress of the Phonetic Sciences. Stockholm, 1995.
[10] Schneider, K. and Möbius, B. "Word stress correlates in spontaneous child-directed speech in German", in 8th Annual Conference of the International Speech Communication Association. Antwerp, Belgium, 2007.
[11] Sluijter, A.M.C., Van Heuven, V.J., and Pacilly, J.J.A., "Spectral balance as a cue in the perception of linguistic stress", Journal of the Acoustical Society of America, 101:503-513, 1997.
[12] Cutler, A., "Lexical Stress", in D.B. Pisoni and R.E. Remez, [Eds], The Handbook of Speech Perception, Blackwell: Oxford, 2005.
[13] Delattre, P., "An acoustic and articulatory study of vowel reduction in four languages", International Review of Applied Linguistics and Language Teaching (IRAL), 7:294-325, 1969.
[14] Kohler, K., "Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics", Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 25:115-185, 1991.
[15] Kohler, K., "A model of German intonation", Arbeitsberiechte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 25:295-360, 1991.
[16] Pierrehumbert, J.B., "The Phonetics and Phonology of English intonation", 1980, MIT: Bloomington.
[17] Baumann, S., Grice, M., and Benzmüller, R., "GToBI – a phonological system for the transcription of German intonation", in S. Puppel and G. Demenko, [Eds], Prosody 2000: Speech recognition and synthesis, 21-28, Adam Mickiewicz University: Poznan, 2001.
[18] Silverman, K.E. and Pierrehumbert, J.B., "The timing of prenuclear high accents in English", in J. Kingston and M.E. Beckman, [Eds], Papers in Laboratory Phonology I: Between the grammar and physics of speech, 72-106, Cambridge University Press: Cambridge, 1990.
[19] Baumann, S. and Grice, M., "The Intonation of Accessibility", Journal of Pragmatics, 38:1636-1657, 2006.
[20] Niebuhr, O., "Perzeption und kognitive Verarbeitung der Sprechmelodie. Theoretische Grundlagen und empirische Untersuchungen", New York: Mouton de Gruyter, 2007.
[21] Niebuhr, O. and Pfitzinger, H. "On pitch-accent identification - The role of syllable duration and intensity", in 5th International Conference on Speech Prosody. 2010.
[22] Fernald, A., "Four-Month-Old Infants Prefer to Listen to Motherese", Infant Behavior and Development, 8:181-195, 1985.
[23] Nazzi, T., Floccia, C., and Bertoncini, J., "Discrimination of pitch contours by neonates.", Infant Behavior and Development., 21(4):779-784, 1998.
[24] Papoušek, M., Papoušek, H., Bornstein, M.H., Nuzzo, C., and Symmes, D., "Infant responses to prototypical melodic contours in parental speech", Infant Behavior and Development, 13(4):539-545, 1990.
[25] Culp, R.E. and Boyd, E.F., "Visual fixation and the effect of voice quality and content differences in 2-month-old infants", Monographs of the Society for Research in Child Development, 39(5-6):78-91, 1974.
[26] Frota, S., Butler, J., and Vigário, M., "Infants' Perception of Intonation: Is It a Statement or a Question?", Infancy, 19(2):194-213, 2014.
[27] Narayan, C.R., Werker, J.F., and Speeter Beddor, P., "The interaction between acoustic salience and language experience in developmental speech perception: evidence from nasal place discrimination", Developmental Science, 13(3):407-420, 2010.
[28] Bion, R.A.H., Benavides-Varela, S., and Nespor, M., "Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences", Language and speech, 54(Pt 1):123-140, 2011.
[29] Gussenhoven, C., "The Phonology of Tone and Intonation". Research Surveys in LinguisticsCambridge, UK; New York: Cambridge University Press. xxiv, 355 p., 2004.
[30] Fernald, A. and Mazzie, C., "Prosody and focus in speech to infants and adults", Developmental Psychology, 27:209–221, 1991.
[31] Norris, D., Cutler, A., McQueen, J.M., and Butterfield, S., "Phonological and conceptual activation in speech comprehension", Cognitive Psychology, 53:146 - 193, 2006.
[32] Baayen, H.R., Piepenbrock, R., and Gulikers, L., "The CELEX lexical database [CD-ROM]. : Linguistic Data Consortium", 1995, University of Pennsylvania: Philadelphia, PA.
[33] Kohler, K. "Segment duration and vowel quality in German lexical stress perception", in 6th International Conference on Speech Prosody. Shanghai, China, 2012.
[34] Barr, D.J., Levy, R., Scheepers, C., and Tily, H., "Random-effects structure for confirmatory hypothesis testing: Keep it maximal", Journal of Memory and Language, 36:255-278, 2013.
[35] Cunnings, I., "An overview of mixed-effects statistical models for second language researchers", Second Language Research, 28(3):369-382, 2012.
[36] Norris, D. and Cutler, A., "The role of strong syllables in segmentation for lexical access", Journal of experimental psychology. Human perception and performance, 14(1):113-121, 1988.
[37] Wellmann, C., Holzgrefe, J., Truckenbrodt, H., Wartenburger, I., and Höhle, B., "How each prosodic boundary cue matters: Evidence from German infants", Frontiers in Psychology, 2012.
[38] Hayes, B.P., "Metrical stress theory: principles and case studies", Chicago [u.a.]: Univ. of Chicago Press, 1995.
[39] Hay, J.S.F. and Diehl, R.L., "Perception of rhythmic grouping: testing the iambic/trochaic law", Perception & psychophysics, 69(1):113-122, 2007.

The Perception of Korean Boundary Tones by First and Second Language Speakers

Hae-Sung Jeon

School of Language, Literature and International Studies, University of Central Lancashire, UK

HJeon1@uclan.ac.uk

Abstract

This paper reports an experiment which investigated the perception of prosody in Korean or non-word utterances by native Korean speakers and English learners of Korean. Listeners rated the degrees of positivity and excitement of resynthesized utterances with different pitch ranges and durations. The results revealed no significant differences between the two groups of listeners. The variations in pitch range and duration had systematic effects on the ratings. However, the interactions between various factors suggest that the mapping between prosodic shapes and their paralinguistic meaning is not straightforward.

Index Terms: prosody, second language, emotion

1. Introduction

Intonation has linguistic and paralinguistic (e.g. emotional, attitudinal) functions which are important in social interactions [1]. There seems to be much cross-linguistic similarity in the paralinguistic (more specifically 'emotional') meaning of prosody; for example, a large pitch movement or an overall increase in the acoustic parameters is related to emotional arousal, e.g. [2]. However, cross-dialectal or -language differences in linguistic or paralinguistic uses of intonation are reported [e.g. 3, 4, 5]. Second language (L2) learners need to acquire the phonology of their target language intonation. Also, linguistic or paralinguistic meaning of prosody can be misinterpreted by speakers of foreign languages.

This study was motivated by English learners' impressionistic comments that Korean prosody often sounds like 'child-whining'. This is probably due to the frequent use of complex boundary tones and significant phrase-final lengthening. Jun [6] identifies nine Intonational Phrase (IP) boundary tones in Korean, and previous studies reported significant lengthening in the IP- or utterance-final syllable, i.e. the IP- or utterance-final syllable is longer than the medial syllable by a factor of 1.6-1.8 (see Ch. 2., [7] for a review). The previous studies mainly dealt with declarative sentences ending with a L or H tone; final-lengthening is likely to be even greater when there are multiple tones associated with the right edge of the phrase or utterance.

The aims of the study were to investigate: 1) whether emotional interpretation of prosody differs between native Korean speakers and English speakers learning Korean; 2) the potential effect of the lexical cue on the emotional interpretation of prosody; 3) the relationship between utterance-final lengthening or pitch range expansion and the emotional interpretation of prosody; and 4) whether different boundary tones are associated with different emotional interpretations.

2. Boundary tones in Korean

The nine tones in [6] include L%, H%, LH%, HL%, LHL%, HLH%, LHLH%, HLHL% and LHLHL% which are transcribed at the L(ow) and the H(igh) targets in the F0. These boundary tones are associated with the right edge, particularly the final one or two syllables, in the IP or utterance. Although [6] and [8] list the tone types and their use, they acknowledge that it is difficult to pinpoint the basic meaning of each boundary tone, since the same shape of pitch movement could signal different meanings depending on the context (e.g. [9]).

The descriptions of intonation in [6] and [8] (the transcription in [8] follows the British Framework) can be summarised as follows:

- L%: declaratives; neutral, formal, as-a-matter-of-fact, polite-apologetic
- H%: seeking information as in yes/no questions
- LH%: questions, continuation rises, explanatory endings in declaratives; annoyance, irritation, disbelief
- HL%: declaratives, wh-questions; kind, polite, apologetic
- LHL%: intensifying the meaning of HL%; persuasive, insisting, confirmative; annoyance, irritation
- LHLH%: intensifying some of the meaning of LH%, i.e. annoyance, irritation, disbelief
- HLHL%: intensifying the meaning of confirming and insisting on one's opinion; nagging, persuading
- LHLHL%: similar to LHL%, signalling more annoyance

That is, L% and H% are described as emotionally relatively neutral in a statement or a question respectively, and it is only HL% which may signal speakers' positive feelings in declaratives. Complex tones are associated with irritation or annoyance; similarly, [10] suggests that the tones with three or more targets do not have a distinct meaning of their own but they are used to intensify the meaning of the less complex tones, e.g. HLHL% intensifies the meaning of HL%.

In this experiment, eight boundary tones, L%, H%, LH%, HL%, LHL%, LHLH%, HLHL% and LHLHL%, were employed. These were chosen intentionally to balance the number of relatively simple tones (L%, H%, LH% and HL%) and complex tones (LHL%, LHLH%, HLHL% and LHLHL%). HLH% was excluded in the main experiment, since the stimuli with this tone sounded unnatural compared to the others. HLH% was used in the practice session preceding the main experiment, together with HLHLH% (see Section 3. 1. 2).

3. Experiment

3.1. Method

3.1.1. Materials

There were four types of trisyllabic stimuli, two Martian (non-word) utterances and two Korean utterances. For the Korean materials which were created first, the experimenter (a trained phonetician, female native Korean speaker) recorded two utterances, [migugɛ] ('in the USA' which can be either declarative or interrogative) and [hatsima] ('don't do it', imperative or interrogative) in the ten boundary tones (L%, H%, LH%, HL%, LHL%, LHLH%, HLHL% and LHLHL% for the main experiment; HLH% and HLHLH% for the practice sessions). The recording took place in a sound-attenuated booth in the Phonetics Laboratory, University of Cambridge (digital recording with a sampling rate of 44.1 kHz). Then all resynthesis was performed with [11] using the PSOLA method. The F0 contours of the recorded utterances were stylised to create source utterances for further resynthesis; the L and H targets were identified and the F0 between two adjacent targets was interpolated. The experimental materials were created by resynthesising F0 (PITCH, two levels; Natural, Expanded) and DURATION (three levels; Shortened, Neutral, Lengthened). For the Natural condition, no additional manipulation of F0 was done. To create the Expanded condition, the H peaks were increased by 2 ST; the only exception was with L% in which the valley was lowered by 2 ST. For the Lengthened duration condition, the utterance-final vowel was lengthened by 50%; for the Shortened condition, the vowel was shorted by 50%; the original duration was maintained in the Neutral condition.

The Martian stimuli ([sesapʰa], [samuda]) were created by concatenating syllables from Finnish utterances spoken by a female speaker. None of the syllables were adjacent to each other in the original utterances. After the concatenation, their pitch and duration properties were resynthesised to be identical to the Korean counterparts ([samuda] and [migugɛ], [sesapʰa] and [hatsima]). In Korean, phrase-initial [s, h] are often associated with H and [m] with L [6]. However, Korean listeners are insensitive to the violation in the tone-segment type association [12] and all stimuli sounded natural.

The paper-and-pencil method was used to collect listeners' responses. A two dimensional model of emotion with valence (i.e. the degree of positivity) and arousal (i.e. the degree of excitement) was employed, with answer sheets from the Self-Assessment Manikin (SAM) rating system [13]. This method is particularly useful for cross-cultural studies when the interpretation of emotional words can be culture-specific. Although there is a third dimension, dominance, it was not included since its reliability has been questioned and it was necessary to keep the experiment simple. The answer sheets had pictures representing five steps of valence (the continuum between smiling face and sad face) and five steps of arousal (the continuum along the excitement scale). Listeners could put a mark on a picture or between two pictures, and therefore a 9-step scale for each dimension was provided. High scores for valence and arousal indicate a high degree of happiness and excitement, respectively. Following the descriptions in [6], [8], and [10], it is expected that the complex tones will show lower positivity (valence) and higher excitement (arousal) than the simple tones, although HL% is likely to be positive with lower excitement.

3.1.2. Experimental procedure

Thirty eight participants without speech or hearing problems participated in the experiment (19 native Korean speakers at the Korea National University of Education, South Korea, age Mean = 21.68 years, SD = 2.75; 19 native British English speakers at the University of Central Lancashire, UK, age Mean = 20.68, SD = 1.59). English speakers were students in a Korean language class at beginner level who had been learning Korean for six months for 2 hours a week. Although their command of Korean was at a basic level, they could understand the experimental materials presented in Korean.

All participants in the same first language (L1) group were tested together. Participants sat in a quiet classroom and the stimuli were played on loudspeakers connected to a PC. Participants were told that they would hear a Martian speaking and their task was to guess the degree of positiveness/negativeness (valence) and the degree of excitement (arousal) from the speech. They were instructed to make full use of the nine-degree scale. They were informed that there would be Korean utterances only after the completion of the Martian part. The presentation order was: [samuda], [sesapʰa], [migugɛ] 'in the USA' and [hatsima] 'don't do it'. Within each stimulus type, the main experiment was preceded by a short practice session which allowed participants to familiarise themselves with the task. In the practice sessions, stimuli with HLH% and HLHLH% in two pitch conditions (Neutral, Expanded) with natural duration were played. Each stimulus was played three times and the presentation order was randomised. The experiment took approximately 35 minutes. In total, 192 utterances were played in the main experiment (4 STIMULUS TYPE × 8 TONE TYPE × 3 DURATION × 2 PITCH).

3.2. Results

3.2.1. Overall patterns

Figures 1-4 show the ratings for each TONE TYPE. Overall, there is little difference between the two L1 groups (as supported by the ANOVA results reported below), although the valence ratings seem slightly higher (more positive) in the English group. For valence, the ratings were relatively high (more positive) for the simple tones (L%, H%, LH%, HL%) than for the complex tones. The valence is particularly high for H%, which was probably interpreted as a yes-no question with a relatively neutral emotional state. As predicted, HL% has a high rating for valence; however, contrary to the description in Section 2, LH% also shows high valence. For arousal, the complex tones (LHL%, LHLH%, HLHL%, LHLHL%) were rated higher (more excited). Pitch range expansion (Figs 1 and 2) led to a higher degree of valence and arousal than the natural pitch contour, although the arousal rating seems less affected than valence rating. For duration (Figs 3 and 4), lengthening seems to be associated with lower valence and higher arousal.

A series of mixed ANOVAs were conducted for valence and arousal respectively, with a between-subject factor, L1 (Korean, English), and within-subject factors, STIMULUS TYPE (ST: [migugɛ], [hatsima], [samuda], [sesapʰa]), TONE TYPE (TT: L%, H%, LH%, HL%, LHL%, LHLH%, HLHL%, LHLHL%), DURATION (D: Neutral, Short, Long) and PITCH RANGE (P: Natural, Expanded) (Greenhouse-Geisser correction applied, α level adjusted to 0.01 after Bonferroni correction).

The L1 main effect was not significant as for either valence or arousal. For valence, complicated interactions involving ST were statistically significant together with the main effects of TT, P and D; significant effects were revealed with ST, TT, P, D, ST × TT, TT × P, ST × D, ST × T × D, and P × D. The mean rating was highest for [sesapʰa] (Martian, M = 4.76, SE = 0.14) and lowest for [hatsima] 'don't do it' (Korean, M = 4.1, SE = 0.12) ([migugɛ] M = 4.73, SE = 0.83; [samuda] M = 4.42, SE = 0.12). For arousal, again, the interaction effects involving ST were significant (ST: [migugɛ] M = 4.73, SE = 0.16; [hatsima] M = 4.87, SE = 0.14; [samuda] M = 4.78, SE = 0.14; [sesapʰa] M = 4.68, SE = 0.15). The statistically significant effects included TT, P, D, ST × TT, TT × P, ST × T × P, ST × D, TT × D, ST × TT × D, TT × P × D and ST × T × P × D. Due to the interactions involving ST, further analyses were carried out with the data split by ST ([migugɛ], [hatsima], [sesapʰa], [samuda]). The results of the further ANOVAs are summarised in Table 1. The table shows that there are few effects involving L1, whilst the effects of TT, P, D and TT × D tend to be significant across the STs. Figures 5 and 6 allow an examination of the TT × D interaction.

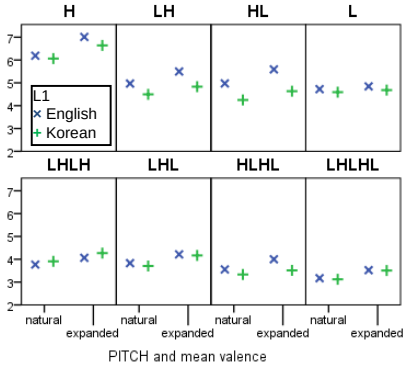


Figure 1: *Valance rating for PITCH. Panels in descending order for mean rating.*

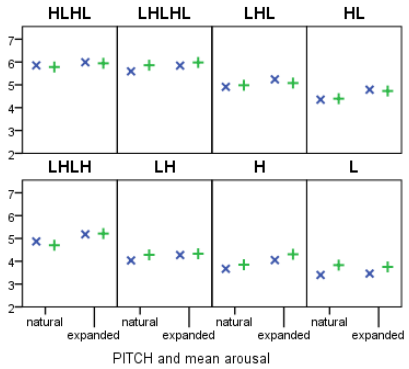


Figure 2: *Arousal rating for PITCH. Panels in descending order for mean rating.*

In Figure 5, [hatsima] ('don't do it') shows little variation across the DURATION conditions with relatively low valence ratings. The other Korean utterance, [migugɛ], shows patterns similar to the Martian utterances in general. Other notable points are: for H%, the valence rating is consistently high for all STs; and for HLHL%, the Korean materials show differences from the Martian materials in that little variation across the DURATION conditions is observed in the Korean materials.

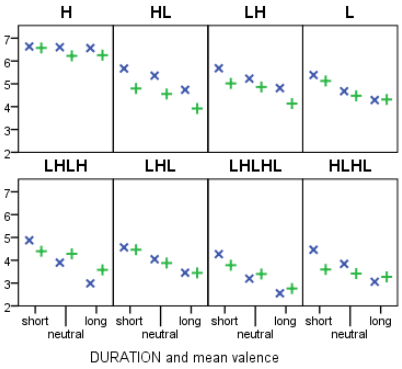


Figure 3: *Valence rating for DURATION. Panels in descending order for mean rating.*

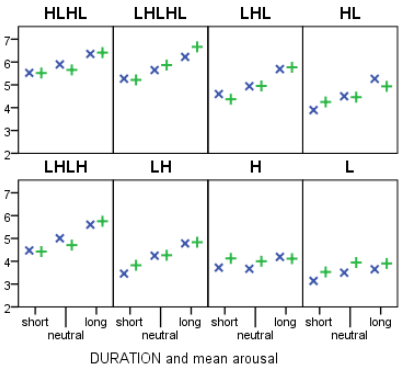


Figure 4: *Arousal rating for DURATION. Panels in descending order for mean rating.*

	valence				arousal			
	K1	K2	M1	M 2	K1	K 2	M 1	M2
TT	**	**	**	**	**	**	**	**
L1	-	-	-	-	-	-	-	-
TT × L1	-	**	-	-	-	-	-	-
P	**	**	**	**	**	-	**	-
P × L1	-	-	-	-	-	-	-	-
D	**	-	**	**	**	-	**	**
D × L1	-	-	-	-	-	-	-	-
TT × P	-	-	-	-	-	-	*	-
TT × P × L1	-	-	-	-	-	-	-	-
TT × D	**	-	*	**	**	*	**	-
TT × D × L1	-	-	-	-	-	-	*	-
P × D	-	-	-	-	-	-	-	-
TT × P × D	-	-	-	-	-	*	**	-
P × D × L1	-	-	-	-	-	*	-	-
T × P × D × L1	-	-	-	-	-	-	-	-

Table 1. *ANOVA results (α level = 0.01), ST: K1 [migugɛ], K2 [hatsima], M1 [samuda], M2 [sesapʰa]. significance codes: ** p < 0.001, * p < 0.01, and - not significant*

In Figure 6, there seems to be a positive correlation between the degree of lengthening and the arousal rating, particularly for Martian, although there are exceptions such as the results for H% and LHL%. For Korean, although the two STs show similar patterns in general, listeners' response patterns differed notably between the two for L%, HL% and HLHL%.

4. Discussion

4.1. Effect of L1

The results show that the emotional interpretation of prosody did not differ between native Korean speakers and native English speakers. English speakers were leaners of Korean at beginner level with a very limited command of Korean and they had been attending the classes for 6 months.

Although the English speakers' valence ratings appeared generally higher than those of the Koreans, the between-L1-group difference did not reach statistical significance. The results seem to support the theory that there is a cross-culturally shared component in the paralinguistic interpretation of prosody related to the expansion of pitch range and durational variations (e.g. [2]). But the generally higher valence ratings by English speakers and the slightly different response patterns between the two groups in some cases (e.g. HLHL% in Fig 3) show that there may also be a language-specific part to be learned which deserves further investigation. Another possibility is that English learners were quick at acquiring the paralinguistic meaning of the Korean boundary tones. Although prosody at the sentence or discourse level is little discussed in the language classroom, students had been exposed to authentic Korean speech through TV shows, films, etc.

4.2. Effect of lexical cue

The fact that listeners' ratings of the Martian utterances did not show a random distribution implies that it is possible to interpret paralinguistic meaning only from prosody without any lexical cues. However, complicated interactions between the STIMULUS TYPE and the TONE TYPE can be seen in Figures 5 and 6, particularly for the Korean utterances, although details are not discussed in the present paper. For example, the Korean sentence whose basic meaning entails annoyance in the declarative, [hatsima] ('don't do it'), received the lowest valence score, and showed less variation in score compared to other types of stimulus. Therefore, the same prosody would signal different paralinguistic meaning depending on the lexical content of a given utterance. A strict and straightforward mapping between semantics and prosody must not be assumed.

4.3. Duration and pitch

For both emotional dimensions, pitch range expansion tended to be associated with intensification, i.e. a greater degree of positivity and excitement. Lengthening was associated with low valence (negativity) but high arousal (excitement).

4.4. Boundary tones

The results generally support the observations that, other things being equal, the complex boundary tones are likely to signal negative emotions and a high degree of arousal, such as irritation or annoyance. However, the effects of lexical cue and variations in duration and pitch span suggest that it would be counterproductive to assume a straightforward mapping between the type of boundary tone and its basic meaning, whether it is linguistic or paralinguistic. In addition, the pitch movement at the boundary and final lengthening would interact with more global characteristics of speech, such as overall pitch register and variations in speaking rate (e.g. [2]).

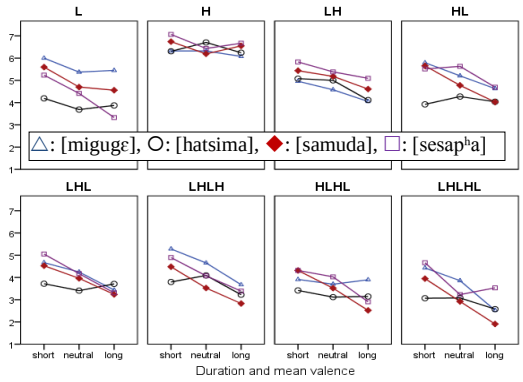


Figure 5: *Valence rating for DURATION. Each line represents a different Stimulus Type*

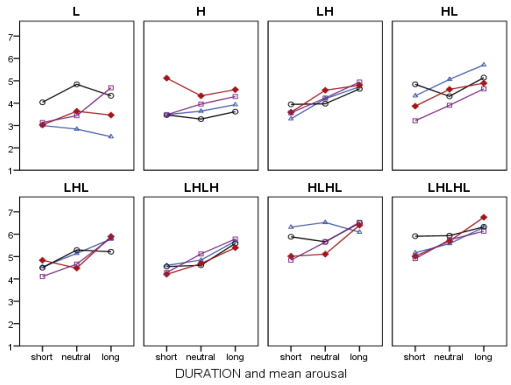


Figure 6: *Arousal rating for DURATION. Each line represents a different Stimulus Type*

5. Conclusions

The emotional interpretation of eight Korean boundary tones did not differ between the two listener groups (native Korean speakers vs. native English speakers learning Korean). The complex boundary tones tended to be perceived as more negative and excited than the simple tones. An increase in pitch span signalled higher positivity and excitement, whilst lengthening signalled higher negativity and excitement. However, the complicated interactions between tone type, utterance-final lengthening, and the lexical cue suggest that the mapping between the type of boundary tone and its paralinguistic meaning is not straightforward.

6. Acknowledgements

I would like to thank Prof. Hyunsong Chung for his help with the experiment in Korea and the members of the Phonetics Laboratory, University of Cambridge, for their help in recording. Thanks are also due to the students in the Beginner's Korean language class 2012-2013 at the University of Central Lancashire who willingly participated in the experiment.

7. References

[1] Ladd, D. R. "Intonational Phonology", 2nd ed. Cambridge University Press, Cambridge, UK, 2008.

[2] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A. & Kotz, S. A. "Factors in the recognition of vocally expressed emotions: A comparison of four languages", Journal of Phonetics, 37: 417-435, 2009.

[3] Grabe, E. & Post, B. "Intonational variation in English", Speech Prosody, 343-346, 2002.

[4] Chen, A., Gussenhoven, C. & Rietveld, A. "Language-specificity in perception of paralinguistic intonational meaning", Language and Speech, 47: 311-350, 2004.

[5] Chen, A. "Perception of paralinguistic intonational meaning in a second language", Language Learning, 59: 367-409, 2009.

[6] Jun, S.-A., "K-ToBI (Korean ToBI) labelling conventions (Ver. 3.1)", UCLA Working Papers in Phonetics, 99: 149-173, 2000.

[7] Jeon, H.-S., "Prosodic Phrasing in Seoul Korean: The Role of Pitch and Timing Cues", Unpublished PhD thesis, University of Cambridge, 2011.

[8] Lee, H.-Y., "Kwukeumsenghak" (Korean Phonetics), Thayhaksa, Seoul [in Korean], 1996.

[9] Kim, H. R. S. "A high boundary tone as a resource for a social action: The Korean sentence-ender –ta", Journal of Pragmatics, 42: 3055-3077, 2010.

[10] Park, M.-j. "The Meaning of Korean Prosodic Boundary Tones", Brill, Leiden, the Netherlands, 2012.

[11] Boersma, P. & Weenink, D. "Praat: Doing Phonetics By Computer" [Computer program]. Version 5.3.23, 2012.

[12] Cho, H. S. "Etude des propriétés acoustiques de la structure prosodique du coréen", unpublished PhD thesis, Université Aix-Marseille I. 2009.

[13] Bradley, M. M. & Lang, P. J. "Measuring emotion: the self-assessment manikin and the semantic differential", Journal of Behavior Therapy and Experimental Psychiatry, 25: 49-59, 1994.

The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps

Irena Yanushevskaya, John Kane, Céline De Looze, Ailbhe Ní Chasaide

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences
Trinity College Dublin, Ireland

yanushei@tcd.ie, kanejo@tcd.ie, deloozec@tcd.ie, anichsid@tcd.ie

Abstract

As part of a broader study of voice prosody in speech communication, this paper looks at intonation in turn-taking. It examines the distribution of pitch patterns and communicative types in the interpausal units (IPUs) preceding pause or gap silences extracted from a corpus of spontaneous speech of Irish English. IPUs preceding speaker change ('Gaps') and IPUs preceding silence where the same speaker continues talking ('Pauses') were selected in the course of automatic extraction of pause/gap silences in dyadic dialogue interactions. A listening test was conducted to establish 'human predictable' pause/gap data sets which were subsequently manually annotated in terms of pitch patterns and communicative types. Overall, the Gaps and Pauses subsets show differentiation in terms of both their communicative types and pitch tunes. Declaratives and Questions are mainly found in Gaps, whereas in Pauses we mainly find Hesitations and Incomplete Declaratives. Gaps are generally characterised by falling or rising pitch patterns, whereas in Pauses a large proportion of speech samples are realised with level pitch. Classification experiments reveal discrimination of pauses and gaps for both prosodic and functional annotation labels. Follow-up work aims to relate intonational characteristics of turn taking with voice quality and temporal dynamics, to provide a holistic view of the processes involved.

Index Terms: dialogue speech, pause, gap, intonation, communicative type

1. Introduction

This paper is part of a broader study of the interaction of intonation and voice source parameters in prosody at the Phonetics and Speech Lab, Trinity College Dublin. The present study complements parallel research exploring voice source [1], f_0 and temporal features (see, for example [2] on the role of prosodic features as well as [3] on f_0 range declination trends in turn-taking organisation). In this paper we describe intonational (pitch patterns) and functional (communicative types) annotation of the interpausal units (IPUs) preceding pause or gap silences extracted from a corpus of spontaneous speech of Irish English.

Robust prediction of turn-taking is crucial for dialogue systems. To date, prediction is largely based on the duration of pause or gap silent intervals, and on the speech interval immediately preceding them. The present detailed analysis of both functional and intonational characteristics of pre-silence chunks for a corpus of Irish English, aims to establish their linkage to turn-taking, and their potential for discriminating speaker changes vs. holds. It further aims to provide the intonational baseline with which we can later correlate voice source and temporal features.

The decision on whether and when we begin to speak in a conversation depends on numerous factors, e.g., lexical and syntactic [4], prosodic [5], vocal effort and audible respiratory cues [6], as well as gestural signals (e.g., head movement, gaze [7] etc.). The importance of lexical-syntactic features for turn-taking management has been emphasised since the early scientific work in this field [4, 8] as well as in more recently reported work [9]. In fact, the perceptual experiment carried out in [9] showed that artificially flattening intonation contours had less of an impact on the predictability of 'end-of-turn' compared to artificially removing the intelligibility of the utterance (by low-pass filtering).

Nonetheless, many researchers focus entirely on prosodic features, an approach which is somewhat justified given that previous studies (e.g., [2, 10, 11]) have found significant discriminative power of prosody-related features, and that this has direct relevance for prosody-only turn prediction (e.g., [12]) in dialogue systems. The role of prosodic patterns in turn-taking has been discussed in [5, 9, 13], see also references therein. For a number of languages (English, German, Dutch, Japanese and Mandarin Chinese), it has been reported that level pitch accents or flat contours at the end of an utterance are indicative of a pause (silent interval within the speech of the same speaker) while any other terminal contours such as rises and falls are indicative of a gap (silent interval between the speech of different speakers) [14-19].

However, the picture emerging is not always as clear cut. In some studies, similar intonation contours have been found for both turn-taking and turn-holding. In [20], 51% of the rising intonation patterns co-occurred with speaker changes while 49% of rises were associated with speaker holds. Furthermore, most studies report a high level of inter-speaker variability.

Other researchers have suggested that turn-taking is likely to be positively affected by the number of prosodic cues present [21, 22]. In addition to pitch contours, prosodic features reported as contributing to turn-management include voice quality (e.g., creaky voice [23]), speech rate and final lengthening [24]. At the level of prosody, we feel it is the dynamic patterning of the voice as a whole (the combination of intonation, voice quality and temporal aspects) that effectively cues speaker changes and holds. While the focus of this paper is on the formal and functional aspects of pitch contours as relating to turn-taking, the bigger picture is a longer term objective.

2. Speech data

2.1. Recordings

The speech data for annotation is taken from the Dublin Institute of Technology Emotional Speech Corpus [26] which consists of seven 10-minute dyadic (male-male and female-female, Irish English) interactions. Six dyadic interactions involving six male and six female speakers were selected from

the original corpus to ensure gender balance. The interactions were elicited in a shipwreck scenario game where participants were presented with 15 items and were given 10 minutes to jointly rank them in order of usefulness for their survival. Recordings were made with participants in separate booths using a professional Neumann microphone connected to an Apple Mac-based Digidesign Pro-Tools Mbox2 recording system. The audio signal was recorded using Pro-Tools software as two separate audio streams and digitised at 96 kHz/24 Bit. Audio was then downsampled to 16 kHz/8 Bit.

2.2. Extraction of the IPUs preceding pause and gap silent intervals

Automatic identification of pauses and gaps was carried out on the speech data using an approach similar to that described in [27]. Binary voice activity detection (VAD) using the VAD algorithm proposed in [28] was carried out on both speaker channels for each dyadic interaction. The threshold for silence interval duration was set to 200 ms to avoid false detection of pauses for speech events like plosives. Silent intervals below this threshold were bridged. Fig. 1 illustrates schematically the output of the VAD process. Overall, 460 gaps and 410 pauses were identified automatically.

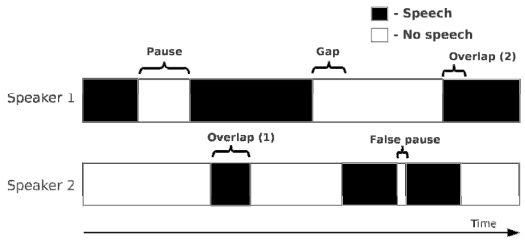


Figure 1. *Schematic representation of a dialogue interaction illustrating pauses, gaps and overlaps. The 'false pause' indicates a silence which is below the threshold (here set to 200 ms).*

As one of our goals is to establish whether and to what extent prosodic characteristics of the speech-chunks immediately preceding pause or gap silent intervals allow automatic prediction of turn-taking in human-machine interaction, we (in a previous study [2]) selected a subset of data where pauses and gaps were clearly predictable by human listeners. A listening test was conducted in which the IPUs (n=870) preceding automatically identified pause and gap silent intervals were presented individually to three raters in random order. Each rater was to indicate, on a 5-point scale, whether in their opinion a pause (same speaker continues) or a gap (speaker change) follows. The rating scale was defined as follows: (1) Very certain the CURRENT speaker continues, (2) Quite certain the CURRENT speaker continues, (3) Don't know! (4) Quite certain the OTHER speaker begins, (5) Very certain the OTHER speaker begins. The raters had an option to indicate that there was an error in the automatic extraction of stimuli, e.g., due to premature truncation of utterances. In total, 6% of the stimuli were marked as an error by the raters. The inter-rater agreement was measured using Krippendorff's α [29]. Analysis revealed fairly high inter-rater agreement (α = 0.74). Only the samples which all three raters identified as being followed by a pause or a gap were retained to form the ultimate 'human predictable' dataset. In total, 302 IPUs preceding gaps and 288 IPUs preceding pauses were retained, which amounts to 70% of the original dataset. This 'human predictable' data set was subsequently manually annotated to explore the distribution of pitch patterns and communicative types of the utterance in the speech-chunks immediately preceding pause and gap silent intervals. For simplicity, we will refer to the IPUs immediately followed by gaps (speaker change) as 'Gaps' and the IPUs immediately followed by pauses (same speaker continues talking) as 'Pauses'. The terms 'Pauses' and 'Gaps' are therefore used to refer to speech-chunks immediately before silent intervals rather than the silent intervals themselves. The data samples in the 'human predictable' data set represent male and female speakers in fairly equal proportion, however the amount of data selected from individual speakers is not necessarily balanced.

3. Annotation of the selected data

The selected Gaps and Pauses data were annotated separately. The aim of this preliminary annotation was to explore the patterns in the distribution of communicative types and pitch tunes in these two data subsets. The manual annotations involved auditory analyses of the extracted data and were initially done by one annotator. Pitch patterns were independently analysed by a second annotator, and the two annotators agreed on 71% of the data. The labels used for the annotation are described in the sections below.

COMMUNICATIVE TYPES: A variety of approaches exist in dialogue speech annotations, e.g., [30-32] and annotation schemes usually include both communicative types of the utterance and functional analyses of dialogue acts. Here we report only the results of communicative type annotation. Since the IPUs were obtained by automatic extraction using a predefined minimum pause threshold, they may contain more than one sentence. An example of such IPU would be *I still like the life jacket. You could drown, like.* In such cases, only the sentence closest to the pause/gap silence (in this example, *You could drown, like*) was analysed. The following communicative type labels were used:

Declarative - grammatically complete/well-formed declaratives, e.g., *Yeah, but I mean we've already lost so many points.*

Incomplete Declarative - grammatically incomplete fragmented declaratives, e.g., *…starting from the most important. Seven…er… a knife.*

Yes/No-Q - grammatically complete/well-formed Yes/No questions, e.g., *Does that mean we picked right every other one?*

WH-Q - grammatically complete/well-formed WH-questions, e.g., *What's the knife gonna do?*

Incomplete Q - grammatically incomplete questions, e.g., *I'd say radio next? …the survival guide and the knife, does it?*

Alt Q - alternative questions, e.g., *Did you say compass or map again ?*

Dec Q - declarative question, e.g., *Yep, flare sounds good to me, so flare's for four?*

Tag Q - tag questions, e.g., *We did do, didn't we?*

Exclamation - [largely based on the intonation with which these utterances were produced], e.g., *Binoculars! Binoculars! Oh god! We did so well!*

Imperative - e.g., *So, do you wanna rank them? Mac, please!*

Hesitation - filled pauses, e.g., *erm, ahm, er,* repetitions and self-corrections, or markedly prolonged words, e.g., *And th[e:]n….*

Backchannel - short acknowledgements such as *sure, yeah, uhum, yep.*

Other labels included '**?**' for uncertain cases and '**n/a**' for the IPUs with no propositional content (e.g., only laughter).

PITCH PATTERNS: In the annotation of pitch patterns in Pauses and Gaps subsets we described the nuclear tunes (each containing the intonation-phrase final pitch accent and its associated boundary tone) using the IViE system [33]. The following tune labels were used:

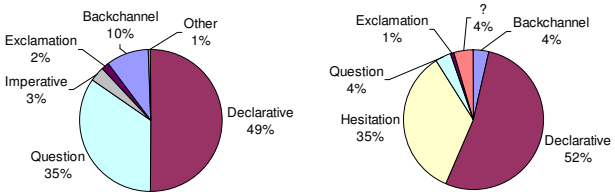
H*+L 0%	fall
H*+L H%	fall-rise
L*+H 0%	low rise
H* H%	high rise
L*+H L%	rise-fall
!H* 0%	downstep (e.g., in item lists)
H* 0%	level (no change/movement of pitch)**.**

Analysis of the pitch patterns reported below was conducted on the final intonational phrase in the sentence closest to the pause/gap silent intervals rather than on the whole IPU.

4. Results and discussion

4.1. The distribution of communicative types

The distribution of communicative types in the Gaps and Pauses subsets is shown in Figure 2. Due to space limitations, the different types of questions are pooled into one category and the same is done for well-formed and incomplete declaratives.



(a) Gaps (b) Pauses

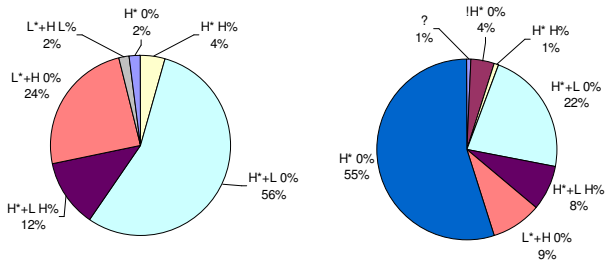
Figure 2. *The distribution of communicative types in the Gaps (a) and Pauses (b) subsets.*

As is clear from Figure 2, almost half of the IPUs in the Gaps subset are Declaratives (of which about a quarter are incomplete). Questions comprise 35% of the Gaps subset, with Incomplete Questions and Yes/No Questions making up the majority of the question types. Among the remaining communicative types, Backchannels are the most frequent (10%), with only 5% of IPUs classified as either Imperatives or Exclamations.

Similar to Gaps, a large part of the Pauses subset is made up by Declaratives (55%). However, the majority of the declaratives in Pauses are incomplete (74% of all declaratives). A fairly large proportion of the IPUs from the Pauses subset are classified as hesitations, a communicative type that does not appear in the Gaps set. The proportion of questions in the Pauses subset is 4% which is substantially lower than in the Gaps set. The proportion of Backchannels is also lower in Pauses, only 4%. The least common communicative type occurring in Pauses is Exclamation (1%). In 4% of the cases the communicative types of the extracted IPUs in the Pauses set were ambiguous and were not annotated.

4.2. The distribution of pitch patterns

The overall distribution of pitch patterns in the Gaps and Pauses data subsets is given in Figure 3 (the data is pooled across all communicative types). More than half of the intonation phrases (IPs) preceding gaps (56%) are realised with a falling pitch, H*+L 0%. Rises L*+H 0% (24%) and fall-rises H*+L H% (12%) are the next most frequent tune types in Gaps. The most frequently occurring pitch pattern in the IPs preceding pauses is level tone H* 0% (55%). The second most common tune type here is fall H*+L 0%, although its proportion in the Pauses subset is substantially lower (22%) than in the Gaps subset (56%). Generally speaking, Gaps are characterised overwhelmingly by pitch movement, whereas Pauses have level tone in the majority of cases. These findings corroborate what has been described in the literature, e.g., [5, 17]. A more detailed analysis of pitch patterns that characterise communicative types most frequently found in Gaps and Pauses is given in the sections below (although not all are illustrated due to space limitations).



(a) Gaps (b) Pauses

Figure 3. *The distribution of pitch patterns in the Gaps (a) and Pauses (b) subsets (across all communicative types).*

A closer look at the pitch tunes in different communicative types found in the Gaps set reveals similar pattern for Declaratives, Incomplete Declaratives, WH Questions and Backchannels: in about 60-70% of the cases, the H*+L 0% tune (fall) is used, followed by L*+H 0% (low rise) as the next most frequent tune type. The proportion of rises is the lowest in Declaratives (10%), it is higher in Incomplete Declaratives and WH questions (17%), and is the highest in Backchannels (24%). In Incomplete Questions, the pitch is predominantly rising (only 12% of samples here have falling pitch). In Yes/No Questions, both falling and rising pitch pattern is used, with a slight preference for rises (about 54% in total).

The distribution of pitch patterns in the Pauses subset is examined mainly for Declaratives and Hesitations which comprise 98% of this data set. Overall, Hesitations are realised predominantly with level pitch (H* 0%), with only a small proportion (14%) having a falling pitch. Incomplete Declaratives are realised with either level or rising pitch (in 80% of the cases), with falls occurring in only 20% of the cases. The proportion of falling pitch is higher in [well-formed] Declaratives (41%), however, in the majority of cases the pitch is either rising or stays level.

4.2.1. *The distribution of tunes: a case of Declaratives*

We compare here in some detail the distribution of tunes in declaratives which make up about 50% of communicative types in each of these two data sets (see Figure 2). The tunes found in Declaratives and Incomplete Declaratives are shown in Figure 4 separately for Gaps (left panel, a) and Pauses (right panel, b). Note that the proportion of incomplete and complete declaratives is reversed in the Pauses set compared to the Gaps set: well formed/complete declaratives constitute 76% of all declaratives in Gaps, but their proportion is reduced to only 26% in Pauses.

There is relatively little difference in the distribution of tunes between complete and incomplete declaratives in the Gaps set. A striking feature of incomplete declaratives in Pauses is a high number H* 0% (level) tunes.

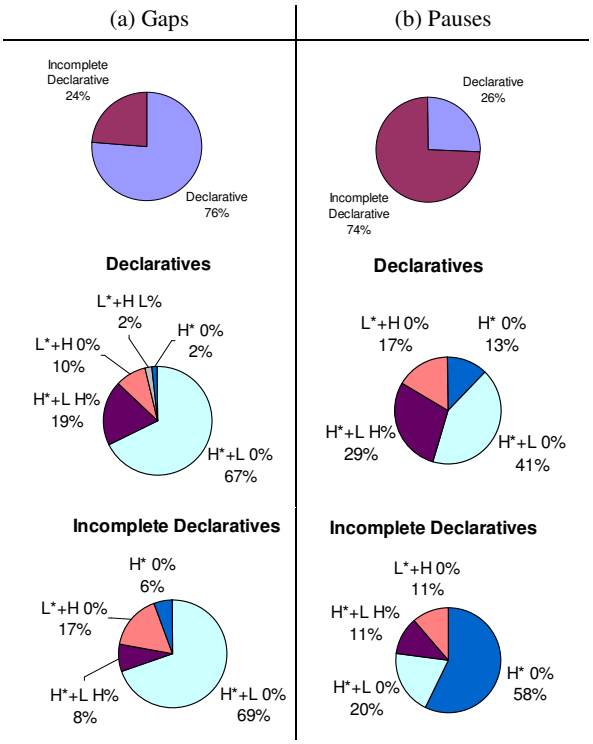


Figure 4. *The distribution of pitch patterns in Declaratives and Incomplete declaratives in the Gaps (a) and Pauses (b) subsets.*

4.3. Intonation and communicative type annotation in classification experiments

In order to investigate the combined discriminative power of functional and intonation labels (derived from speech-chunks immediately preceding pause and gap silent intervals) for differentiating pauses and gaps we carry out a speaker independent classification experiment. For this we utilise a support vector machine (SVM) based classifier with a radial basis function kernel. As input features we use the manually obtained annotation labels, separated into multiple binary features (e.g., the feature for H*+L 0% would have all samples with this annotation label assigned the value of 1 and all others 0). Classification is carried out using a leave-one-speaker-out procedure where the data of a single speaker is held out solely for testing, with the remainder of the data used for training the SVM classifier. This procedure is repeated for all 12 speakers with classification error (%) retained each time.

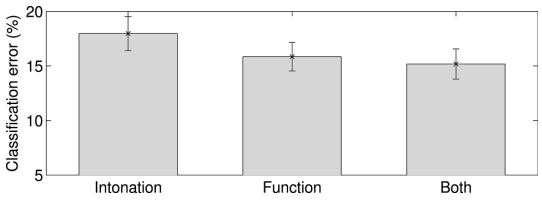


Figure 5. *The results of the support vector machine-based classification experiment. Shown are mean and standard error values.*

The results of this analysis are summarised in Figure 5 and are shown for intonation and functional label features separately and combined. The overall classification error is strikingly low for what is an extremely difficult discrimination problem. Using intonation labels alone one achieves a mean classification error of around 18%. The result provides a strong motivation to produce robust automatic characterisation of such intonation patterns.

Functional labels provide an even lower mean classification error (~15%). This also suggests that detection of functional labels would be beneficial for the prediction of pauses and gaps. However, deriving such information automatically would require the combination of an automatic speech recognition component as well as a subsequent text analytics procedure both of which are liable to introduce significant errors. The combination of the intonation and functional labels brings only a minor reduction in mean classification error and suggests a high level of redundancy between the two classes of labels.

5. Conclusions

In this paper we examined the distribution of communicative functions and pitch tunes in the 'human predictable' Pauses and Gaps subsets selected from the Dublin Institute of Technology Emotional Speech corpus. Overall, Gaps and Pauses subsets show differentiation both in terms of their communicative types and pitch patterns. Declaratives and Questions are commonly found in Gaps, whereas in Pauses it is mainly Hesitations and Incomplete declaratives. Gaps are mainly characterised by falling or rising pitch patterns (pitch movement), whereas in Pauses a large proportion of speech samples are realised with level pitch. Results suggest that including information on pitch patterns in the speech-chunks immediately preceding pause and gap silent intervals appears to enhance automatic discrimination of pauses and gaps. Our future work will exploit the findings of this study to examine other prosodic dimensions, voice quality and temporal characteristics, and their interaction with intonational features.

6. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET). We would like to thank Dr. Brian Vaughan (Dublin Institute of Technology) for providing the DIT Emotional Speech Corpus.

7. References

[1] J. Dalton, J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "GlóRí - the glotal research instrument," Speech prosody 2014, Dublin, Ireland, [accepted].

[2] J. Kane, I. Yanushevskaya, C. De Looze, B. Vaughan, and A. Ní Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," Interspeech 2014, Singapore, [submitted].

[3] C. De Looze, I. Yanushevskaya, J. Kane, and A. Ní Chasaide, "Pitch range declination and reset in turn-taking organisation," Speech Prosody 2014, Dublin, Ireland, [accepted].

[4] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language,* vol. 50, 1974.

[5] M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé, "Prosodic features in the vicinity of pauses, gaps and overlaps," in *Nordic Prosody. Proceedings of the Xth Conference*, M. Vainio, R. Aulanko, and O. Aaltonen, Eds., ed Berlin: Peter Lang, 2009, pp. 95-106.

[6] A. Rochet-Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in German spontaneous speech," presented at the Interspeech 2013, Lyon, France, 2013.

[7] A. Kendon, "Some functions of gsze direction in social interaction," *Acta Psychologica,* vol. 26, pp. 22-63, 1967.

[8] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology,* vol. 23, pp. 283-292, 1972.

[9] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: a cognitive cornerstone of conversation," *Language,* vol. 82, pp. 515-535, 2006.

[10] A. Gravano and J. Hirshberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language,* vol. 25, pp. 601-634, 2011.

[11] D. Schlangen, "From reaction to prediction: experiments with computational models of turn-taking," presented at the Interspeech 2006, Pittsburg, Pennsylvania, 2006.

[12] L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," presented at the ICASSP, Hong Kong, China, 2003.

[13] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica,* vol. 62, pp. 215-226, 2005.

[14] B. Oreström, *Turn-Taking in English Conversation*: Krieger Publishing Company, 1983.

[15] J. Local and J. Kelly, "Projection and 'silences': notes on phonetic and conversational structure," *Human Studies,* vol. 9, pp. 185-204, 1986.

[16] K. Kohler, "Prosodic boundary signals in German," *Phonetica,* vol. 40, pp. 89-134, 1983.

[17] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics,* vol. 31, pp. 251-276, 2003.

[18] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and Speech,* vol. 41, pp. 295-321, 1998.

[19] J. Fon, K. Johnson, and S. Chen, "Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech," *Language and Speech,* vol. 54, pp. 5-32, 2011.

[20] J. Edlund, M. Heldner, and J. Gustafson, "Utterance segmentation and turn-taking in spoken dialogue systems," in *Computer Studies in Language and Speech*, ed Frankfurt am Main: Peter Lang, 2005, pp. 576–587.

[21] A. Gravano and J. Hirshberg, "Turn-yielding cues in task-oriented dialogue," presented at the SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, London, UK, 2009.

[22] H. Friedberg, "Turn-taking cues in a human tutoring corpus," presented at the Annual meeting of the Association for Computational Linguistics, Portland, USA, 2011.

[23] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association,* vol. 31, pp. 139-152, 2001.

[24] M. Zellers, "Pitch and lengthening as cues to turn transition in Swedish," presented at the Interspeech 2013, Lyon, France, 2013.

[25] J. Local, J. Kelly, and W. H. G. Wells, "Towards a phonology of conversation: turn-taking in Tyneside English," *Journal of Linguistics,* vol. 22, pp. 411-437, 1986.

[26] B. Vaughan, "Naturalistic emotional speech corpora with large scale emotional dimension ratings. PhD thesis," Dublin Institute of Technology, 2011.

[27] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics,* vol. 38, pp. 555-568, 2010.

[28] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters,* vol. 6, pp. 1-3, 1999.

[29] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures,* vol. 1, pp. 77-89, 2007.

[30] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova*, et al.*, "ISO 24617-2: A semantically-based standard for dialogue annotation," presented at the LREC 2012, Istanbul, Turkey, 2012.

[31] H. Bunt, "Dimensions in dialogue annotation," presented at the LREC 2006, Genoa, Italy, 2006.

[32] C. Soria and V. Pirrelli, "A recognition-based meta-scheme for dialogue act annotation," presented at the Workshop Towards Standards and Tools for Discourse Tagging, Somerset, New Jersey, 1999.

[33] E. Grabe, B. Post, and F. Nolan, "Modelling intonational variation in English: the IViE system," presented at the Prosody 2000, Kraków, Poland, 2001.

Realization of Narrow Focus in Hong Kong English declaratives—a Pilot Study

Holly Sze Ho Fung, Peggy Pik Ki Mok

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
hollyfung@cuhk.edu.hk, peggymok@cuhk.edu.hk

Abstract

Narrow focus, i.e., focus on one word, is realized differently in native English and Cantonese. While it is signaled primarily by on-focus F0 changes such as F0 range expansion in English, it is marked essentially by lengthening of duration in Cantonese. Another difference is the pitch of the post-focus elements. While native English demonstrates post-focus F0 compression, Cantonese shows no significant post-focus pitch change. To investigate how narrow focus is realized in Hong Kong English (HKE), an emergent variety of English spoken by native speakers of Cantonese in Hong Kong, a controlled production experiment was conducted with 8 HKE speakers. Results showed that while the HKE speakers did realize foci with significant on-focus F0 range expansion, they exhibited no post-focus compression.

Index Terms: Hong Kong English, focus, post-focus compression

1. Introduction

Focus, as defined by Crystal [1], is information that is "at the center of their [speakers'] communicative interest". It can be classified as narrow or broad depending on its scope. First introduced by Ladd [2], the term "narrow focus" was defined as focus on "a particular constituent or a small set of constituents", and "broad focus" as that on an entire utterance, or any constituent larger than that of a narrow focus [3].

Early acoustic studies of focus realization in English found that narrow focus is signaled by multiple prosodic cues including raised F0 peak and mean F0, expanded F0 range, lengthened duration and increased intensity [4][5]. More recent studies, while confirming these findings, suggested a broader temporal domain of focus prosody. Instead of being solely signaled by on-focus cues, narrow focus was also found to be marked by post-focus F0 lowering and F0 range suppression, which were also referred to as post-focus compression (PFC) [6][7].

Similar to native English, narrow focus was found to be signaled by multiple cues in Cantonese including an increase in duration and intensity [8][9][10] as well as pre-focal pause insertion [8]. As for whether F0 is an acoustic correlate of Cantonese narrow focus, opinions diverged. On the one hand, Man [6] found significant F0 range expansion that was local to the focused syllable. Gu and Lee [8] found both F0 heightening and expansion in a broader scope spanning from the syllable before the sentence-medial focus to the end of the utterance, with the heightening effect more prominent on high-tone target. On the other hand, in a more recent study by Wu and Xu [10], which is more reliable regarding the larger sample size and the method of focus elicitation used, no significant on-focus or post-focus F0 variations was found.

While pitch is surely an important acoustic correlate of focus in native English, its role in focus-marking is in doubt in Cantonese. Regarding such difference, a legitimate question to ask about Hong Kong English (HKE), a non-native variety of English that emerges from the interaction between the two languages, is whether pitch is an acoustic correlate of focus in it. To answer this question, and to assess the role of transfer from Cantonese to HKE, a controlled production experiment was conducted.

2. Method

2.1. Materials

10 English and 6 Cantonese declarative sentences were used in the experiment. All the 10 English sentences (see Table 1) contain the carrier frame __ *gave a* __ *to* __, in which the empty slots were filled by different keywords. The keywords, all sonorants for continuous F0 contours, were controlled for their number of syllables and stress pattern. Half of them were monosyllabic and the other half disyllabic, all stressed on the first syllable.

Table 1. *List of English sentences (keywords underlined)*

Monosyllabic keywords	
1.	Ann gave a mole to Wayne.
2.	Lee gave a ring to Wong.
3.	May gave a ram to Lynn.
4.	Ron gave a wheel to Ray.
5.	We gave a yam to Nell.
Disyllabic keywords	
6.	Alan gave a lemon to Laura.
7.	Larry gave a melon to Luna.
8.	Mary gave a lolly to Annie.
9.	Mummy gave a warning to Molly.
10.	Willy gave a ruler to Emma.

Similarly, the Cantonese sentences also contained keywords in the sentence-initial, medial and final positions. In addition, to examine the effects of focus on different lexical tones (see Table 2), each of the sentences contained keywords of one of the six lexical tones in Cantonese. Table 3 is a list of the Cantonese sentences used.

Table 2. *Summary of Cantonese lexical tones*

	T1	T2	T3	T4	T5	T6
Tone shape	high level	high rising	mid level	low falling	low rising	low level
Tone code	55	25	33	21	23	22

Table 3. *List of Cantonese sentences (keywords underlined)*

Tones	Sentence in *Jyutping* with English translation
T1	*maau1 mi1 maai5 zo2 jyun1 joeng1 sung3 bei2 wu1 aa1*. 貓咪買咗鴛鴦送畀烏鴉。 'The cat bought tea coffee for the raven.'
T2	*waa2 min2 gin3 dou2 jyun2 jyun2 sik1 jin2 jiu2 neoi2*. 畫面見到婉婉飾演妖女。 'It is shown on the screen that Jyunjyun plays a banshee.'
T3	*aa3 jin3 keoi5 ge3 ngoi3 hou3 ling6 jan4 jim3 wu3*. 阿燕佢嘅愛好令人厭惡。 'Aajin's interest is disturbing.'
T4	*maa4 maa4 fan1 fu3 jung4 jan4 heoi3 maai5 jau4 jim4*. 嫲嫲吩咐傭人去買油鹽。 'Granny asked the maid to buy oil and salt.'
T5	*lou5 ng5 deoi5 doi6 mei5 neoi5 fei1 soeng4 jau5 lai5*. 老五對待美女非常有禮。 'Loung is very polite to beauties.'
T6	*wu6 wai6 ting3 cung4 ming6 ling6 zeon1 hang4 jam6 mou6*. 護衛聽從命令進行任務。 'The guard went for a mission on command.'

To compare the F0 and duration of focused and non-focused keywords, each of these 16 sentences were produced in four conditions, one with neutral focus (i.e., no focus) and the other three with focus in the sentence-initial, medial and final positions. To elicit these focus conditions, two sets of stimuli were prepared. The set for eliciting neutral focus contained the 16 sentences in plain font, and the other for eliciting narrow focus consisted of 48 sentences (16 sentences x 3 focus positions) with focused keywords in different positions highlighted in bold.

The reason for choosing this method over the more commonly adopted one using prompt questions was that although the latter was successful with native speakers of English in some previous studies [4][5][12][13][6][7], it did not work for the HKE speakers in our earlier pilot test. The pilot speakers (whose data are not presented here) did not realize any focus on the pieces of information being asked for, i.e., their answers were the same as those with neutral focus. Moreover, the speakers also reported that they found the prompt questions rather irritating. As a result, in this study, narrow focus was elicited instead by highlighting the keywords in bold and directly asking the speakers to emphasize them, conveying them as the most important information in the sentences, but no instruction was given to them on how they should emphasize them phonetically. Thus, the focus realized by them was particularly emphatic.

In addition, 12 practice sentences similar to the focus-eliciting stimuli were prepared to familiarize the speakers with producing narrow focus in various positions.

2.2. Speakers

3 male and 5 female native Cantonese speakers aged between 22 and 24, who acquired English as their second language, were recruited as subjects. All were undergraduates of local universities who received pre-tertiary education at local primary and secondary schools, where they were exposed to native English for 3 to 6 years from their native-speaking English teachers. Two of them have been to an English-speaking country before, one to the US for four days and the other to Australia and New Zealand for two weeks and six months respectively. As for their oral English proficiency, five attained grade C in the oral paper of Use of English (UE) in the Hong Kong Advanced Level Examination (HKALE). Among the rest, two received grade D and one grade E.

In addition to the HKE speakers, two native American English (AmE)-speaking exchange students from New York, aged 20 and 21, were recruited as control subjects. Only two control subjects were used because the patterns of narrow focus in English were already well established in the literature.

2.3. Procedures

The experiment for the HKE speakers was divided into two sessions. In the first session, the speakers were shown and recorded reading the sentences without focus. They were reminded to avoid placing emphasis on particular words in order to elicit neutral focus successfully. The sentences were arranged randomly into three blocks, the first with the Cantonese test sentences and the other two with the English sentences with monosyllabic and disyllabic keywords respectively. Each sentence was recorded twice.

The second part of the experiment began with a training session, in which the speakers were asked to read the practice sentences with foci in different sentence positions after they were told that the words in bold were the most important pieces of information to be emphasized. After becoming familiar with the procedure, they were then recorded reading the Cantonese and English test sentences with narrow foci, which were arranged into two and four blocks respectively. Each block was read in two repetitions.

As for the American English speakers, the experiment procedures were basically the same, except that they were not asked to read the Cantonese stimuli.

2.4. Data Analysis

Extraction of utterances and labeling of individual syllables of the keywords were done using Praat. For the disyllabic keywords in English with the CVCV(C) structure, syllables were segmented between the first vowel and the second consonant, i.e., CV/CV(C). For example, the keyword *lemon* was segmented as "le/mon", rather than "lem/on".

Each labeled syllable was then measured for its: 1) F0 range and 2) mean F0, which were calculated from F0 values obtained from 10 equal-distant points along the pitch contour of the target syllable.

3. Results

3.1.1. American English

Figures 1 and 2 show the mean F0 ranges and mean F0s of the disyllabic keywords produced by the two AmE speakers. (Given the admittedly small sample size, no statistical analysis was performed on the AmE data.) The abbreviation *s* stands for the stressed syllable and *us* for the unstressed syllable of a keyword. As expected, on-focus F0 range expansion occurred regardless of sentence position in AmE. On-focus F0 heightening, although insubstantial, was also observed consistently in all sentence positions. The data also confirmed the presence of post-focus compression (PFC) in native English, since both mean F0s and F0 ranges of keywords were found to decrease in the post-focus condition. (Data of the

monosyllabic keywords, which show the same pitch pattern as the disyllabic ones, are excluded here owing to page limit. Details of them are available upon request.)

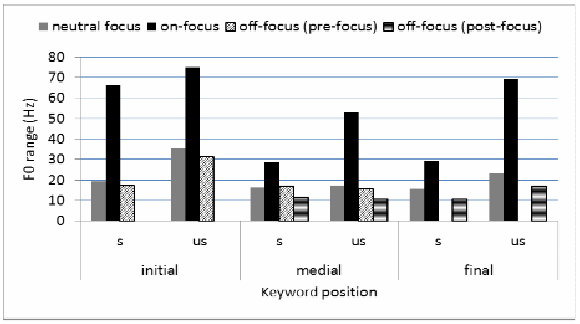


Figure 1: *Mean F0 ranges of English disyllabic keywords produced by AmE speakers*

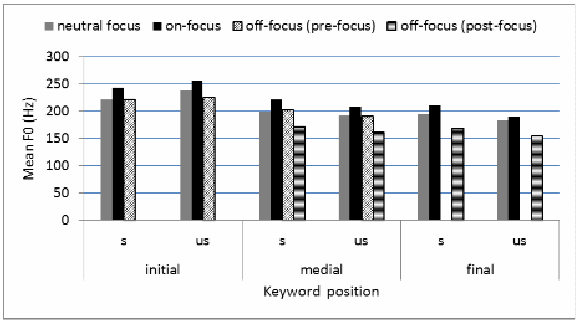


Figure 2: *Mean F0s of English disyllabic keywords produced by AmE speakers*

3.1.2. Cantonese

Among the data of all the six lexical tones collected, only those of T1, T3 and T6 were analyzed. The reason was that since we wanted to compare the Cantonese and the HKE data to evaluate the influence from the former on the latter (if any), and that HKE was suggested to be tonal with H, M and L tones [14][15][16], analysis of the three level tones in Cantonese would best suit the purpose.

One general observation of the Cantonese data is that there some insubstantial increases in mean F0 of the keywords in focus. As an example, Figure 3 shows the mean F0 of the T1 keywords produced by the female speakers. In a two-way ANOVA test, focus (neutral versus on-focus) was found to have significant main effect on mean F0 of the keywords of T3 (p=0.018) and T6 (p=0.011) produced by the female speakers, and those of T1 (p=0.027) and T6 (p=0.021) produced by the male speakers. The results suggest that the speakers demonstrated some on-focus F0 heightening, but not in a consistent manner.

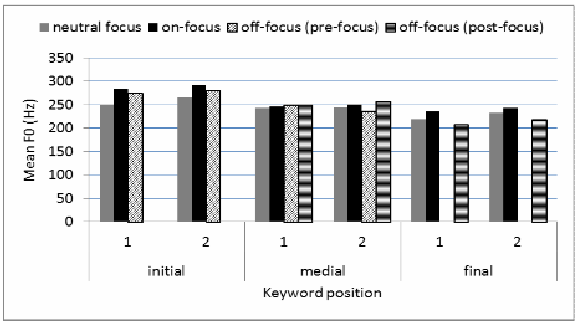


Figure 3: *Mean F0s of T1 keywords produced by female HKE speakers (abbreviations: 1—1st syllable; 2—2nd syllable of the disyllabic target)*

As for F0 range, on-focus expansion of it was found only sporadically. In addition to the result that focus had a significant main effect on F0 range only in the T1 foci produced by the female speakers (p=0.04), no strong evidence was found to support that F0 range expansion is a cue to Cantonese focus.

Neither was PFC found in the data. No significant effect of focus was found on either mean F0 or F0 range of post-focus keywords. The result confirmed Wu and Xu's [10] suggestion that PFC does not exist in Cantonese.

3.1.3. HKE

The mean F0 range of the English monosyllabic foci produced by the HKE speakers displayed in Figure 4 and Figure 5 show that for both gender groups, a keyword was produced with remarkably larger F0 range in the on-focus condition than in the neutral focus condition regardless of its position. In a two-way ANOVA assessing two main effects, namely focus (neutral focus and on-focus) and word position (initial, medial and final), focus was found to have significant effect on F0 range in monosyllabic keywords produced by both male (p=0.009) and female speakers (p=0.025), suggesting that the expansion of F0 range was focus-induced.

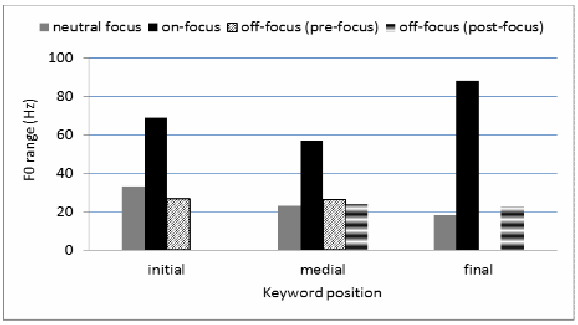


Figure 4: *Mean F0s of English monosyllabic keywords produced by male HKE speakers*

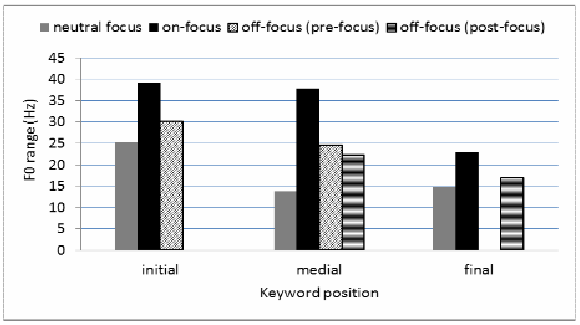


Figure 5: *Mean F0s of English monosyllabic keywords produced by female HKE speakers*

Substantial on-focus F0 range expansion was also found in the disyllabic foci, as shown below in Figure 6 and 7. In a three-way ANOVA analysis with the focus (neutral focus and on-focus), lexical stress (stressed and unstressed) and word position (initial, medial and final) as the factors, focus was again found to have a significant effect on F0 range of keywords produced by both male (p=0.009) and female (p=0.000) speakers. The results suggest that F0 range expansion is a cue to narrow focus in HKE.

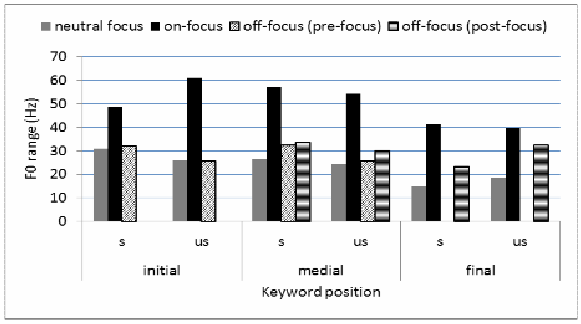


Figure 6: *Mean F0 ranges of English disyllabic keywords produced by male HKE speakers*

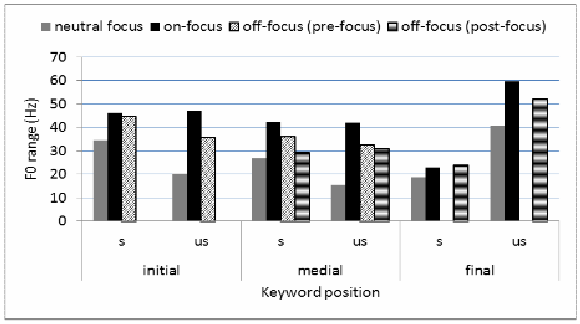


Figure 7: *Mean F0 ranges of English disyllabic keywords produced by female HKE speakers*

On the other hand, only random mean F0 heightening was found. A two-way ANOVA test found that focus had no significant effect on the rises. Therefore, mean F0 is unlikely an on-focus cue to narrow focus in HKE. The random increases of F0 might simply be the by-product of F0 range expansion.

Same as in Cantonese, no post-focus F0 lowering or F0 range suppression was found in HKE. In fact, noticeable expansion of F0 range was found in off-focus keywords, both pre-focus and post-focus ones, as shown above in Figures 6 and 7, meaning that the speakers actually had global F0 range expansion for the entire utterance with narrow focus. One possible reason for the finding is that since the subjects were asked explicitly to produce focus, they might as a result have spoken with a more exaggerated register which rendered more rise and fall in the pitch contour. Further tests are needed to verify this.

4. Discussion

Based on the results in Section 3.1.3, pitch does seem to be an acoustic correlate to narrow focus in HKE. Despite the absence of consistent on-focus F0 heightening and PFC, substantial on-focus F0 range expansion in both monosyllabic and disyllabic keywords located in all sentence positions suggests that HKE speakers do signal narrow focus with manipulation of pitch.

The absence of PFC in the HKE data can be attributed to cross-linguistic influence from Cantonese. As mentioned in Section 3.1.3, none of the HKE speakers exhibited PFC in Cantonese. If there were no influence from Cantonese, we would expect them to demonstrate PFC in their English like the two AmE speakers did. In fact, the absence of PFC in HKE is not at all surprising. Similar to HKE, English spoken by native speakers of Taiwan Mandarin, another language without the post-focal feature, was found not to have it either [17]. Its absence in HKE may provide an additional piece of evidence for that PFC in L2 English is susceptible to transfer from L1.

On the other hand, the presence of on-focus F0 range expansion in HKE cannot be explained simply by transfer from Cantonese. As mentioned, on-focus F0 range expansion occurred only sporadically and insubstantially in Cantonese. This seems to suggest that HKE speakers have two distinct intonation patterns for focus marking in Cantonese and HKE.

Based on our preliminary findings, focus intonation of HKE was found to be a "hybrid" of that of its parent languages: native English and Cantonese. While it shows on-focus F0 range expansion like native English does, it exhibits no PFC, similar to Cantonese. In other words, pitch is an acoustic correlate of narrow focus in HKE, although it is limited to the local domain, i.e., the word in focus. It has to be emphasized, though, that the conclusions are drawn from results obtained by a non-canonical way of focus elicitation involving the use of text in bold and explicit instruction to produce a focus instead of the more common one using prompt questions. Further studies employing various focus elicitation methods are needed to corroborate the results here.

5. References

[1] D. Crystal, *A Dictionary of Linguistics and Phonetics*. Oxford, UK: Blackwell Publishing Ltd., 2008.

[2] D. R. Ladd, *The structure of intonational meaning : evidence from English*. Bloomington: Indiana University Press, 1980, p. 239.

[3] D. R. Ladd, *Intonational phonology*, vol. 79. Cambridge; New York: Cambridge University Press, 1996, p. 334.

[4] S. J. Eady and W. E. Cooper, "Speech intonation and focus location in matched statements and questions," *Journal. of Acoustic Society of America.*, vol. 80, pp. 402–415, 1986.

[5] S. J. Eady, W. E. Cooper, G. V Klouda, P. R. Muller, and D. W. Lotts, "Acoustical Characteristics of Sentential Focus: Narrow v.s. Broad and Single v.s. Dual Focus Environments," *Language and Speech*, vol. 29, no. 3, pp. 233–251, 1986.

[6] Y. Xu, C. Xu, and X. Sun, "On the temporal domain of focus," in *Proceedings of International Conference on Speech Prosody 2004*, 2004, pp. 1–4.

[7] Y. Xu and C. X. Xu, "Phonetic Realization of Focus in English Declarative Intonation," *Journal of Phonetics*, vol. 33, pp. 159–197, 2005.

[8] W. Gu, K. Hirose, and H. Fujisaki, "The effect of paralinguistic emphasis on F0 contours of Cantonese Speech," in *Proceedings of International Conference on Speech Prosody 2006*, 2006.

[9] W. Gu and T. Lee, "Effects of tonal context and focus on Cantonese F0" in *Proceedings of 16th International Congress of Phonetic Sciences*, 2007, pp. 1033–1036

[10] W. Wu and Y. Xu, "Prosodic focus in Hong Kong Cantonese without post-focus compression," In *Proceedings of International Conference on Speech Prosody 2010*, pp. 1–4, 2010.

[11] V. Man, "Focus effects on Cantonese tones: An acoustic study," In *Proceedings of International Conference on Speech Prosody 2002*, pp. 2–5

[12] S. Hoskins, "The Prosody of Broad and Narrow Focus in English: Two Experiments," in *Proceedings of the 5th European Conference on Speech Communication and Technology*, 1997, vol. Rhodes, Gr, pp. 791–794.

[13] S. Jannedy, "Acquisition of narrow focus prosody," *Proceedings of the GALA '97 conference: Language Acquisition, Knowledge Representation & Processing*, 1997.

[14] W. H. Y. Cheung, "Span of high tones in Hong Kong English," *Hong Kong Baptist University Papers in Appied. Lanuage Studies*, vol. 12, pp. 19–46, 2008.

[15] L.-H. Wee, "Phonological patterns in the Englishes of Singapore and Hong Kong," *World Englishes*, vol. 27, no. 3–4, pp. 480–501, Oct. 2008.

[16] S. S. Y. Yiu, "Intonation of English spoken in Hong Kong," Hong Kong Baptist University, 2010, 2010.

[17] T. Visceglia, C. Y. Tseng, C. Y. Su, and C. F. Huang, "Realization of English narrow focus by L1 English and L1 Taiwan Mandarin Speakers," in *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* , 2011, vol. Hong Kong, pp. 2074–2077.

Altering speech synthesis prosody through real time natural gestural control

David Abelman[1]*, Robert A.J. Clark*[2]

[1]School of Informatics, University of Edinburgh, Edinburgh, UK
[2]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
robert@cstr.ed.ac.uk

Abstract

This paper investigates the usage of natural gestural controls to alter synthesised speech prosody in real time (for example, recognising a one-handed beat as a cue to emphasise a certain word in a synthesised sentence). A user's gestures are recognised using a Microsoft Kinect® sensor, and synthesised speech prosody is altered through a series of hand-crafted rules running through a modified HTS engine (pHTS, as described in [1]). Two sets of preliminary experiments are carried out. Firstly, it is shown that users can control the device to a moderate level of accuracy, though this is projected to improve further as the system is refined. Secondly, it is shown that the prosody of the altered output is significantly preferred to that of the baseline pHTS synthesis. Future work is recommended to focus on learning gestural and prosodic rules from data, and in using an updated version of the underlying pHTS engine.

The reader is encouraged to watch a short video demonstration of the work at ***http://tinyurl.com/gesture-prosody***.

Index Terms: speech prosody, pHTS, real time control, gesture

1. Introduction

Despite the significant advances made in speech technology in recent years, producing speech that is both **expressive** and **reactive** to a user's input or local environment is a significant challenge facing speech synthesis today, and one on which there has been limited research to date [1] [2].

This work aims to construct a system to alter speech prosody in a limited number of ways, based on a user's real time gestural input. Future extended systems of this type would have various potential applications. A primary use may be in text-to-speech communication aids of those with vocal disorders. More natural expression and prosody may be 'conducted' by the user in real-time, either through a set of standard natural gestures, or through a set of custom-designed gestures for those with physical disabilities (for example, eyebrow or finger movements).

Additionally, technology developed as part of this system may be incorporated within potential 'sign-language synthesis' systems of the future [3] [4]. In addition to synthesising words based on sign-language hand movements, the manner in which the gestures are performed may indicate to the system a certain expressive or emphatic style in which to synthesise the speech.

Finally, other potential applications may exist within the entertainment industry. For example, the technology may be adapted for use within synthesised singing voices, or perhaps within future 'instrument-voice hybrids' that people may wish to control through body gestures. Ultimately, any situation in which it would be useful to improve expressiveness of a voice-like synthesis in real-time would benefit from the research that this work undertakes.

2. Background

This project is built upon a modified HTS engine 'pHTS' (*performative* HTS) [1]. This technology allows HTS synthesis to be reactive to its environment - whether adapting to surrounding conditions, or being controlled expressively by a user (as is the case here). In order to make the system reactive, the phonetic context required in calculating the synthesis parameters is reduced from that of the whole sentence to a much smaller window. This change requires two main modifications. Firstly, the context used in training the model is reduced to just the current and surrounding phonemes, and the current and previous syllable. Secondly, during synthesis, the generation of parameters occurs on a sliding window of two labels.

A variety of potential applications for pHTS have been outlined and developed by the group. These include *HandSketch*, a pen-based musical instrument prototype [5], speech synthesis based on face-tracking [6], and accent interpolation through an interactive map application [7]. Meanwhile, non-pHTS examples of hand-controlled prosodic modification include [8].

Finally, [2] incorporates skeleton tracking (using Microsoft Kinect) into the pHTS system to create a reactive speech synthesiser, in which pitch and duration are controlled by the vertical position of both hands. It is found that meaningful expressiveness is difficult to simulate when pitch and duration modulations are controlled in this particular way. It is this work that this project intends to build on, incorporating gesture recognition and more constrained prosodic modification rules.

3. Design

3.1. Preliminary decisions

A number of preliminary decisions were made to constrain the scope of the project. Future iterations of the work should revisit these in order to extend the system's abilities. Decisions included:

- Limiting the system's prosodic vocabulary to contrastive emphasis, general emphasis, yes/no questions and wh-questions only
- Realisation of prosodic shifts through manual parameter shifts on a single speech database, as opposed to switching between multiple recorded databases for different effects without explicitly shifting speech parameters
- Alteration of pitch and duration only, as the two primary parametric drivers of prosody (i.e. volume, spectral energy, pause models etc. are left untouched)

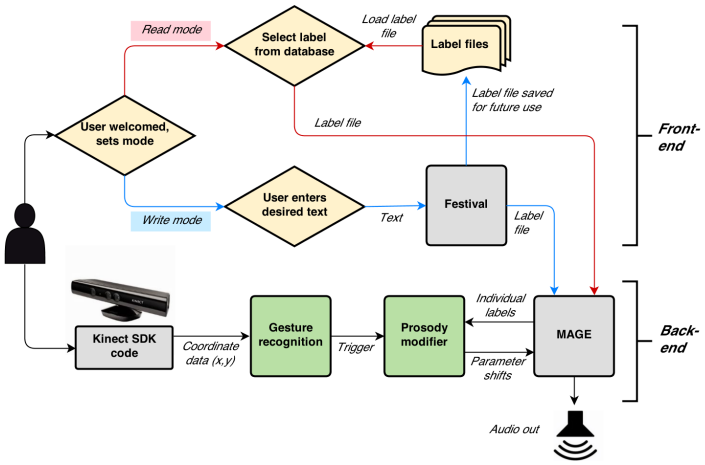


Figure 1: *System setup.*

- Magnitudes of pitch/duration shifts are manually encoded, as opposed to being learnt through data
- Gesture recognition implemented in a rule-based fashion, as opposed to being learnt through data
- Limiting the system's gesture recognition to one handed beats (contrastive and general emphasis, different hand for each), head tilts (yes/no questions and wh-questions, different side for each)
- Constricting emphasis to content words (not function words) and to stressed syllables within these words only. Future iterations may use a more flexible or intelligent natural language model to bias the prosodic effects towards the most likely syllables given the semantic meaning of the text, and the user's timing

A number of small-scale tests were carried out in order to optimise values for parameter shift magnitudes, and gesture tracking rules. Details may be found in [9]. The final set of rules implemented in terms of skeletal coordinates and pHTS parameter shifts are laid out in Section 3.3.4 of [9], with an illustrative example (contrastive emphasis) shown in this paper in Section 3.3.

3.2. System setup

A schematic for the system set up is laid out in Figure 1. The system has two modes, **read** and **write**. Both modes create a temporary label file which the backend loads into MAGE. **Write** mode passes the user's textual input (currently entered using a keyboard) to **Festival** Speech Synthesis System v2.1[1] in order to create the label file, whereas **read** mode allows the user to select a label file previously created by this process. The format of the label file required by the system is that outlined within [10].

The backend is built in C++ on top of MAGE and the work already carried out as part of [2]. OpenFrameworks[2] is used as the graphical and audio framework. The main application loop repeatedly calls an 'update' function (15 times per second), in which pre-existing **Kinect SDK code** tracks *x* and *y* coordinate data for the user's skeleton.

[1] http://www.cstr.ed.ac.uk/projects/festival/download.html
[2] http://www.openframeworks.cc/

The skeletal coordinate data are used within the **gesture recognition** stage. A simple set of if-else rules act as triggers for initiating prosodic effects, as described in [9]. Once some gesture has been recognised, the **prosody modifier** must allocate the prosodic effect to a specific syllabic unit (or set of syllabic units). Prosodic adjustments are carried out using a set of if-else rules, as also described in [9].

The parameter shifts are sent to the **MAGE** engine, and are used to shift parameter trajectories as appropriate via the pHTS engine. This functionality pre-exists within the MAGE platform code.

In addition to the resulting audio output, the application provides a visual representation of the user's skeleton as part of the application interface. Flashing text and graphical meters indicate to the user when a gesture has been recognised, and how the pitch and duration of the synthesis is being shifted.

3.3. Contrastive emphasis example

This section briefly outlines the set of rules implemented to enable the system to add *contrastive emphasis* to a synthesised sentence. Similar sets of rules exist for *general emphasis*, *yes/no questions* and *wh-questions*, and can be found in [9].

Gestural rules: As the Kinect recognises the left hand moving above the left hip, a 'contrastive window' of ∼0.5 seconds (8 frames) is triggered. (Note that due to latency issues it has not been possible to fine control the placement of this window for any hand movements above the hip, though this would be desirable in an improved system).

Prosodic rules: A pitch accent is applied if a content word's stressed syllable falls within this 0.5 second window. The pitch accent consists of a raised pitch (28%) and a reduction in speed (-10%). Following this, the remainder of the sentence is lowered in pitch (-14%) and increased in speed (14%). The final syllable of the sentence is raised in pitch (10%) to counter the default falling accent provided by pHTS.

4. Experiment 1 - Generation test

Two experiments have been carried out, the first being a generation test to investigate the accuracy to which users can control the system.

4.1. Experimental setup

In total, 12 native English speakers were tested using the system. Each user was asked to add some form of emphasis to 31 different synthesised sentences through gestural control, with each sentence being repeated eight times consecutively. Each sentence contains one or two gestures, leading to a total of 264 gestures that have been performed and tracked for each user, taking around 50 minutes in total.

A script was placed within the user's view. For each sentence the user was told which word to emphasise, and with what action (normally contrastive emphasis, as this is the most obvious to the ear). The author was able to track each attempt as being correct, early / late (missing the intended syllable but not emphasising an unintended syllable) or very early / very late (emphasising an unintended syllable). Immediately after each attempt and prior to the next, the subject was told by the author if they had gestured correctly, early or late (although it was often already clear to the user without prompting). This feedback is justified, as we would expect a fully-developed system to pro-

vide the user with some kind of explicit feedback on where their attempted emphasis fell.

The sentences are split into five primary sections. These sections are presented to each subject in the same order (avoiding learning bias over the session), but sentences within each section are presented in different orders according to various Latin Squares. Full sentence lists may be found in the original work's appendix [9].

4.2. Experimental results

A selection of findings are presented here. Two-tailed binomial tests are used to mark 95% confidence intervals in tables. The mean of the binomial distribution is set to the proportion of correct emphases out of all attempts. Chi-squared tests are used to calculate p-values for significance when confidence intervals overlap.

Spread of false positive and negative gesture timings: Considering just the *first attempt* across all sentence types in the 50 minute experiment, correct emphasis is applied 50% of the time. The user emphasises *no* word 30% of the time (gesturing only slightly too late or early), and the *wrong* word 20% of the time. These results are shown in Table 1. Note that when *all eight attempts* are considered, the application of correct emphasis improves from 50% to 65% (as users improve with practice on each sentence).

Table 1: *Spread of emphasis gesture timings (1st attempt only)*

Emphasis	% of time
Very early (wrong emphasis)	$7 \pm 3\%$
Early (no emphasis)	$4 \pm 2\%$
Correct emphasis	$50 \pm 6\%$
Late (no emphasis)	$26 \pm 5\%$
Very late (wrong emphasis)	$12 \pm 4\%$

Improvement over session: Users were requested to add emphasis onto individual words within a sentence at the start of the experiment, and onto individual words in sentences of similar rhythm after 25 and 45 minutes of practice. Results show that users do improve between 0 and 25 minutes ($p<0.01$), though not significantly between 25 and 45 minutes, suggesting that accuracy levels plateau with reasonably little experience. These results are shown in Table 2.

Table 2: *Accuracy improvement over the session*

Time since experiment start	Avg. accuracy (8 attempts)
0 minutes	$52 \pm 7\%$
25 minutes	$64 \pm 7\%$
45 minutes	$67 \pm 7\%$

Other results: Other results found include the following:

- A user emphasising *two words per sentence* will on average obtain a lower accuracy rate for the second word, in comparison to emphasising that same word *alone* in a sentence, if the words are in close enough proximity.
- Users have a significantly lower accuracy rate emphasising *words at the start and end of sentences*. Emphasising a word at the beginning of a sentence may be expected to be more challenging, as the user can be caught off-guard. There is less clear reason for words at the ends of sentences to be harder to emphasise - this may be due to a quirk of the sentences chosen within this experiment (for example, the 'rhythm' with which the synthesiser recites them).
- A user *speaking the text out loud* at the same time as gesticulating to control the synthesiser does not find his/her accuracy altered significantly, on average. It had been hypothesised that gestures may be performed more naturally at the correct moments if the user was speaking out loud whilst gesticulating. However, given this result, this wouldn't be a technique that is recommended to users in any future system.
- The *naturalness* (or unnaturalness) of the word to be emphasised does not affect a user's accuracy rate significantly.

5. Experiment 2 - Listening test

A listening test has been carried out to investigate the extent (if any) by which the output is perceived to be more natural, or have a different meaning, relative to baseline pHTS. Null hypotheses assumes a listener chooses an option from the forced choice test with equal probability - i.e. the options are equivalent. Two-tailed binomial tests are used to calculate p-values, and 95% confidence intervals are shown in tables.

5.1. Experimental setup

In total, 33 subjects were recruited for a listening test, lasting 20-30 minutes depending on the subject, under controlled conditions. All subjects identified themselves as being native English speakers, and received £6 in compensation. Each user was presented with 92 sentences split across 7 sections, and asked to select one of two options in a forced-choice test. This choice involved selecting either a preferred audio clip or a preferred textual option, depending on the question. Similarly to the generation test, sections were presented in a consistent order, but questions and options within each section were appropriately randomised. Once again, a full list of test sentences may be found in the original work's appendix [9].

5.2. Experimental results

A selection of findings regarding **contrastive emphasis** are presented here.

Perceived naturalness of contrastive emphasis: A question narrated by the author and pairs of synthesised responses were played to listeners. The sentences were of a defined form - see [9] for details. One of the responses would be a neutral pHTS synthesis, the other would have a word emphasised using the reactive synthesis system. This emphasis may be appropriate (correct word emphasised) or inappropriate (incorrect word emphasised). The participant must choose which of the two responses seems more natural. To illustrate, the user was hypothesised to prefer the 'appropriately' emphasised (first) response in this case:

'Did Jess have trout for her breakfast yesterday?'

1. *'No, Jess had **SALMON** for her breakfast yesterday.'*
2. *'No, Jess had salmon for her breakfast yesterday.'*

whereas the user was hypothesised to prefer the neutral pHTS baseline (second) response where an 'inappropriately' emphasised response was presented:

'Did Jess have trout for her breakfast yesterday?'

1. *'No, Jess had salmon for her **BREAKFAST** yesterday.'*
2. *'No, Jess had salmon for her breakfast yesterday.'*

Indeed, it has been found that listeners significantly prefer contrastive emphasis over neutral prosody when the emphasis is delivered on the appropriate word, and significantly prefer the neutral prosody over contrastive emphasis when the emphasis is delivered on an inappropriate word ($p<0.01$). Both of these results are as hypothesised, and are shown in Table 3.

Table 3: *Contrastive emphasis - naturalness*

Listener preference:	Emphasised	Neutral
Appropriately emph'd synthesis	**86 ± 4%**	14 ± 4%
Inappropriately emph'd synthesis	22 ± 7%	**78 ± 7%**

Effect on semantic interpretation of sentence by contrastive emphasis: Two textual options were presented to participants in writing, along with a single synthesised audio response. For example, a pair of textual options used was:

1. *The black dog was lying on the mat*
2. *The white mouse was lying on the mat*

with the single synthesised audio response being:

*'No, the **WHITE** dog was lying on the mat.'*

Since the emphasis is on *'WHITE'*, we would expect the listener to select the first of the two textual statements as the more appropriate, given the response. (If the emphasis had been on *'DOG'* we would have expected the user to choose the second textual option). Indeed it was found that these expected choices were made the majority of the time: the position of the synthesised emphasis significantly changes the user's semantic interpretation of the response ($p<0.01$). Results are shown in Table 4.

Table 4: *Contrastive emphasis - semantics*

Emphasis perceived to be on:	1st word	2nd word
When neutral synthesis	32 ± 7%	**67 ± 7%**
When 1st word emph'd in synthesis	**94 ± 6%**	5 ± 6%
When 2nd word emph'd in synthesis	8 ± 5%	**92 ± 5%**

Other results: Other results found include the following:

- As described above, within a defined sentence structure, appropriate contrastive emphasis is considered more natural than neutral synthesis. However, a separate section of the listening test showed that appropriate contrastive emphasis placed on the *final* syllable of a sentence is *not* perceived to be more natural by listeners. The hypothesised reason is that a final syllable emphasised contrastively needs to rise *and* fall in pitch within the same syllable. However, contrastive emphasis built into the current system only raises the emphasised syllable, resulting in an unnatural effect. This issue should be addressed in any future iterations of the work.
- Although contrastive emphasis was set as described in Section 3.3 using a 28% pitch rise on the emphasised syllable for most of the listening test, alternative values for the magnitude of the pitch rise were evaluated against one another within a short section of the test. A rise of roughly 20% was found to be optimal according to listeners' choices. Future iterations of the system may tweak parameters through tests such as these to optimise perceived naturalness.
- The experiment also evaluated listeners' perceptions of interrogative prosody, in addition to contrastive emphasis as outlined here. The reader may consult the original work for details [9].

6. Discussion

This work presented does suggest that it is possible to improve prosody of speech synthesis in real time through gestural controls. Users can control the emphasis with some accuracy, and listeners overwhelmingly prefer correctly emphasised sentences over baseline pHTS. It should be noted however that no analysis of the benefit of a correct emphasis versus the cost of an incorrect emphasis (from the listener's point of view) has been carried out within this work.

Future work should enable the user to control the system with a much superior accuracy rate to that obtained here. The largest obstacle to accurate control within this iteration of the work was relatively poor latency, caused by audio buffering. Future iterations of the work will use MAGE 2.0 [11] rather than MAGE 1.0, which the current system is based on. This will improve audio buffering times, meaning that the system has the potential to react to more granular gesture timings. For example, rather than triggering a window for potential emphasis as a user raises their wrist above their hip, the system may apply emphasis within an instant of recognising that the user's hand has reached the apex of an emphatic 'beat' trajectory. This should improve accuracy rates significantly. Additionally this would allow prosody to be affected *before* the user reaches this apex if required, for example by decreasing the speed of a syllable immediately preceding a contrastively emphasised syllable.

Future iterations of the system should implement machine learning based techniques for gestural recognition, as opposed to the current rule-based setup. This will result in a more flexible system going forward, in which new gestures can be recorded more easily, added or modified by a user, and customised to those with accessibility requirements. Similarly, learning prosodic parameter shifts through data (as opposed to the hard-coded rules currently used) will allow the system to scale more easily to a larger repertoire of available prosodic effects. A final modification that may improve the system's accuracy would be to incorporate a discourse model to aid prediction of where the user intends to apply emphases and similar prosodic effects. Even if the user's timing is slightly out, the system may then intelligently factor in prior probabilities of words most likely to be emphasised given the discourse context.

In summary, the work presented is in its early stages, but improvements such as these will lead to a novel and natural method of altering speech synthesis prosody in real time.

7. References

[1] Maria Astrinaki, Nicolas D'alessandro, Benjamin Picart, Thomas Drugman, and Thierry Dutoit. Reactive and continuous control of HMM-based speech synthesis. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 252–257. IEEE, 2012.

[2] Robert AJ Clark, Magdalena Anna Konkiewicz, Maria Astrinaki, and Junichi Yamagishi. Reactive control of expressive speech synthesis using Kinect skeleton tracking. *Information Processing Society of Japan*, 112(369):175–178, 2012.

[3] S Sidney Fels and Geoffrey E Hinton. Glove-talk: A neural network interface between a data-glove and a speech synthesizer. *Neural Networks, IEEE Transactions on*, 4(1):2–8, 1993.

[4] Robert Akl. *Evaluating appropriateness of EMG and flex sensors for classifying hand gestures*. PhD thesis, University of North Texas, 2012.

[5] Maria Astrinaki, Nicolas dAlessandro, and Thierry Dutoit. MAGE - A platform for tangible speech synthesis. In *Proceedings of the 12th Conference on New Interfaces for Musical Expression (NIME'12)*, 2012.

[6] Maria Astrinaki, Nicolas D'alessandro, and Thierry Dutoit. MageFaceOSC: Performative speech synthesis based on realtime face tracking. *QPSR of the numediart research program*, 5(1):15–16, 2012.

[7] Maria Astrinaki, Junichi Yamagishi, Simon King, Nicolas dAlessandro, and Thierry Dutoit. Reactive accent interpolation through an interactive map application. *Proceedings of the 14th Conference of the International Speech Communication Association (Interspeech 2013)*, 2013.

[8] Christophe dAlessandro, Albert Rilliard, and Sylvain Le Beux. Chironomic stylization of intonation. *The Journal of the Acoustical Society of America*, 129(3):1594–1604, 2011.

[9] David Abelman. Altering speech synthesis prosody through real time natural gestural control. *Edinburgh University MSc Thesis (http://hdl.handle.net/1842/8373)*, 2013.

[10] Heiga Zen. An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, 2006.

[11] Maria Astrinaki, Nicolas D'alessandro, Loic Reboursière, Alexis Moinet, and Thierry Dutoit. MAGE 2.0: new features and its application in the development of a talking guitar. In *Proceedings of the 13th Conference on New Interfaces for Musical Expression (NIME'13)*, 2013.

Body size projection by voice quality in emotional speech—Evidence from Mandarin Chinese

Xiaoluan Liu, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
xiaoluan.liu.12@ucl.ac.uk, yi.xu@ucl.ac.uk

Abstract

This study attempts to extend the line of research on using body size projection theory to account for emotional speech. It is predicted by the theory that anger is expressed by projecting a large body size with low pitch, rough voice and long vocal tract; happiness is expressed by projecting a small body size with high pitch, breathy voice and short vocal tract. Ten native speakers of Mandarin with drama training background recorded sentences in happy, angry, disgust and neutral emotions. We used multiple measurements to assess voice quality, formant dispersion (as an indicator of vocal tract length) and pitch. The results show clear support for the body size projection theory in voice quality, with anger and disgust associated with pressed and rough voice while happiness with breathy voice. But the results of formant dispersion and pitch demonstrate no clear directions. While the study is the first to show clear speech production support for the body size projection theory with voice quality data, the equivocal results of formant and pitch call for improvement in the method of emotion elicitation in the laboratory.

Index terms: emotional speech, body size projection theory, Mandarin Chinese

1. Introduction

Speech is one of the most important means of expressing emotions. While there have been many studies on the acoustic characteristics of emotional speech, the findings have generally been mixed [12]. One particular line of research, based on the body size projection theory, however, has taken a rather different theoretic approach to emotional speech. Originally proposed by Morton [8] for explaining animal calls, and later extended by Ohala [11] to human speech, the key idea is that vocal emotional expressions are a mechanism evolved under a selection pressure to influence the behaviour of other individuals in social interactions. For example, an angry expression is to project a large body size to scare off the opponents in case of confrontation; a happy expression is to project a small body size to attract the listener. Recently, this idea has seen support from a series of perception research in which the speech stimuli are synthetically manipulated in terms of pitch, vocal tract length and voice quality to simulate body size projection [3, 17, 18]. It is shown that listeners hear speech with synthetic parameters that project a large body size both as being spoken by a large person and as expressing anger. And they hear speech that projects a small body size as spoken by a small person and expressing happiness and friendliness [10, 17, 18].

In Xu et al. [17] an extension to the body size projection principle is proposed to incorporate additional "bio-informational dimensions" that may also serve to influence listener behaviour. These include dynamicity, audibility and association. One of these dimensions, namely, dynamicity, has been tested and shown to be relevant to the perception of a number of emotions [17, 18]. So far, however, there has not been systematic testing of either size projection or any of the other bio-informational dimensions in production studies. Thus, despite the demonstrated listener sensitivity to some of them, it is not yet known whether and how consistently speakers use any of them in the production of emotional speech. Also, the bio-informational dimensions have been tested on anger, happiness, sadness and fear [17], but not yet on disgust, which is also considered as one of the basic emotions [4].

Identifying acoustic properties that are clearly emotion-relevant from production data has not been easy, however, partly because it is generally difficult to elicit genuine emotions in the laboratory, even from trained actors and actresses. Recently, however, an emotion portrayal method has been developed for inducing emotions, and it is argued that the method can induce reliable and natural emotions in laboratory conditions because it reflects people's daily strategy to control and display emotions [12,13].

The present paper reports the results of a production study to test the bio-informational dimensions of emotion, using the emotion portrayal as the induction method, and Mandarin Chinese as the target language. Three emotional expressions are examined—anger, disgust and happiness, with the aim to find out whether the predictions of the theory can be confirmed in speech production, and how effective emotion portrayal is as a method of emotion induction in the laboratory.

2. Methodology

The basic design of the study is to have native speakers of Mandarin produce sentences with anger, disgust, happiness and neutral emotion using emotion portrayal as the induction method.

The stimuli, as shown in Table 1, consist of four Mandarin sentences with the target words *mao* and *men* imbedded in sentence-medial and sentence-final positions. The selection of *men* and *mao* is for the ease of phonetic segmentation while minimizing consonantal perturbation of F_0. To test for interaction between tone and emotion, we assigned the key word *mao* with four lexical tones: High tone *mao1* for sentence 1, rising tone *mao2* for sentence 2, low tone *mao3*

for sentence 3 and falling tone *mao4* for sentence 4. The syllable *men* following *mao* is assigned with tone 5 (neutral tone) for semantic/pragmatic naturalness of the sentences constructed. Note that *xiaomaomen* in the sentences is a compound word with three syllables, denoting the name of a brand and their album, although *mao* and *men* when used separately have their own independent meanings.

Table 1. Stimuli of the experiment. The number in each syllable represents the lexical tone: 1 for H (High tone), 2 for R (Rising tone), 3 for L (Low tone), 4 for F (Falling tone), and 5 for N (Neutral tone).

xiao3 little	*mao1* cat	*men5* particle	*yue4dui4 fa1xing2 le5*	*xiao3* little	*mao1* cat	*men5* particle
	mao2 fur				*mao2* fur	
	mao3 mortise		band release particle		*mao3* mortise	
	mao4 hat				*mao4* hat	
The band *Xiaomaomen* has released the album *Xiaomaomen*.						

Ten native Mandarin speakers with drama training background were recruited as subjects. They reported no speech or hearing problems. Emotion portrayal method was used to induce emotion, i.e., having subjects imagine themselves being in an emotional state as vividly as possible when saying each sentence [12, 13]. The recording was conducted in a sound-controlled booth. All the sentences were repeated 3 times by 10 subjects in anger, disgust, happiness and neutral emotion, resulting in 4 (*mao1/2/3/4*) * 4 (emotions) *10 (subjects) *3 (repetitions) = 480 sentences.

A customized version of Prosody Pro [16], running under Praat [1], was used to extract and analyse F0 contours, voice quality and formant dispersion (as indicator of vocal tract length [5]). Segmental boundaries were labelled by hands with visual inspection and listening validation. Given the difficulty in assessing voice quality, we obtained multiple measurements used in previous studies as follows:

H_1-H_2*, H_1-A_1*, and H_1-A_3* (H_1 and H_2 refer to the amplitude of the first and second harmonics of voiced segments; A_1 and A_3 refer to the amplitude of the first and third formants. The * symbol indicates that the measurements are estimates based on frequency bands without literally tracking harmonics and formants.);

Centre of spectral gravity—A measure for how high the frequencies in a spectrum are on average [1];

Energy of voiced segments below 500Hz and 1000Hz [14];

Hammarberg index—Maximum energy difference between the range of 0-2000 Hz and the range of 2000Hz to 5000Hz in the voiced section of the speech under examination [7];

Skewness—A measure for how much the shape of the spectrum below the centre of gravity is different from the shape above the mean frequency [1].

The above five measurements are indicators of spectral slope which is directly related to voice quality [6].

Jitter—Mean absolute difference between consecutive periods, divided by the mean period [6];

Shimmer—Mean absolute difference between the amplitudes of consecutive periods, divided by the mean amplitude [6];

Harmonicity—Harmonics-to-noise ratio measuring "the extent of acoustic periodicity expressed in dB" [6].

The above three measurements are indicators of voice roughness. Each of them indicates in one way or another the amount of aperiodicity in the voice.

Formant dispersion—Mean frequency differences between adjacent formants, which is an indicator of vocal tract length [5, 11].

The perception test by Xu et al. [18] shows that anger is associated with smaller formant dispersion than happiness. Therefore, this is a vocal dimension worth examination as well in this study.

F0 values were measured in semitones to reduce the bias towards higher pitch range over lower pitch range.

3. Results

The measurements used in statistical analyses were taken from the target syllables *mao* and *men* (both sentence medial and final across the four emotions). A series of three-way repeated measures ANOVAs were performed on the measurements of voice quality, formant dispersion and F0 of these syllables. The independent variables were emotion (anger, disgust, happiness and neutral), tone of *mao* (high, rising, low and falling), and sentence position (medial and final). A series of post-hoc Tukey HSD tests were also conducted to examine which pair of the emotions was significantly different.

The ANOVA results on tone, sentence positions and the interaction between tone, sentence positions and emotion are non-significant (hence not displayed), suggesting that tone and sentence position of the target syllable (*mao* and *men*) do not affect the acoustic characteristics of the emotional speech. In contrast, emotion is found to affect all features except formant dispersion (Table 2a-2b).

Table 2a. Means and *p* values of ANOVAs for H_1-H_2*, H_1-A_1*, H_1-A_3*, centre of gravity (COG), formant dispersion (FD), jitter (JI), shimmer (SH) and harmonicity (HA) of the four types of emotional speech (A=anger, D=disgust, H=happiness, N=neutral).

	H1-H2*	H1-A1*	H1-A3*	COG	JI	SH	HA
A	-3.1	-3.9	21.9	882.7	0.06	0.22	8.9
D	-2.6	-3.8	22.8	876.6	0.05	0.21	9.11
H	-0.1	-1.8	26.1	811.6	0.05	0.2	10.24
N	-0.3	0.7	31.1	608.4	0.04	0.19	11.58
p	<.05	<.05	<.05	<.05	<.05	<.05	<.05
F	3.84	3.92	4.01	4.06	4.35	3.65	3.98
df	3,27	3,27	3,27	3,27	3,27	3,27	3,27

With regard to H_1-H_2*, H_1-A_1*, and H_1-A_3*, it has been reported [18] that the decrease in the values of the three parameters results in not only an increase in perceived body size but also a change of perceived emotion from happiness to anger. This is in the same direction as the results of this study. As can be observed from Table 2a, the smallest values all correspond to anger, indicating an exaggerated

body size. As the values go up, the emotions produced change gradually from anger to disgust, happiness and neutral emotion.

Table 2b. Means and *p* values of ANOVAs for energy below 500/1000Hz (E<500Hz, E<1000Hz), Hammarberg index (HI), skewness (SK), formant dispersion (FD) and F0 (semitone) of the four types of emotional speech (A=anger, D=disgust, H=happiness, N=neutral).

	E <500	E <1000	HI	SK	FD	F0
A	0.22	0.65	21.09	1.1	773.8	96.07
D	0.44	0.82	22.74	1.3	762.7	89.45
H	0.33	0.7	23.46	2.31	756.6	93.93
N	0.55	0.86	27.31	2.04	760.9	92.11
p	<.05	<.05	<.05	<.05	>.05	<.05
F	4.09	4.16	3.52	3.67	2.79	3.86
df	3,27	3,27	3,27	3,27	3,27	3,27

Table 3. Results of Post hoc Tukey HSD tests, showing pairs of emotions that are significantly different (labelled by ✓) in terms of each of the parameters (A=anger, D=disgust, H=happiness, N=neutral emotion).

	A vs. D	A vs. H	A vs. N	D vs. H	D vs. N	H vs. N
H1-H2*		✓	✓	✓	✓	✓
H1-A1*		✓	✓	✓	✓	✓
H1-A3*		✓	✓	✓	✓	✓
COG		✓	✓	✓	✓	✓
E<500Hz	✓	✓	✓	✓	✓	✓
E<1000HZ	✓		✓	✓	✓	✓
HI	✓	✓	✓	✓	✓	✓
JI		✓	✓	✓	✓	✓
SH		✓	✓	✓	✓	✓
HA		✓	✓	✓	✓	✓
FD						
F0	✓	✓	✓		✓	✓

In terms of centre of spectral gravity, the prediction from perception findings [10, 18] is that it should show a decrease in values from anger to happiness. This is confirmed in Table 2a: Anger and disgust both have higher values than happiness, again indicating a larger projected body size while happiness has a low value, indicating a smaller projected body size.

Jitter and shimmer show higher values for anger than happiness (Table 2a), which, though not specifically predicted before, is in line with Morton's [8] hypothesis that aggressiveness is associated with rough voice.

Harmonicity (Table 2a) is the highest for neutral emotion and lowest for anger, with happiness in the middle having slightly higher values than disgust. Table 2b shows that energy below 500Hz and 1000Hz of happy speech is higher than that of angry speech, and that the tendency is also true with regard to Hammarberg index. As for skewness, happiness has the highest degree of skewness, followed by neutral and disgust emotion, with anger having the smallest skewness (Table 2b). These measurements are, therefore, consistent with the results of H_1-H_2*, H_1-A_1* and H_1-A_3* and hence with the body size projection theory.

The difference between anger and disgust is non-significant in most of the parameters mentioned above (Table 3), indicating a tendency of anger and disgust being positioned towards the same end of the body size dimension.

The differences in formant dispersion, however, are non-significant between different emotions (Table 2b and 3). Also, Table 2b shows that anger has a higher pitch than happiness and neutral emotion, with disgust having the lowest pitch. This is not consistent with the findings of previous perception studies.

4. Discussions

The results of voice quality are consistent with the body size hypothesis in that anger is found to project a large body size with pressed (as indicated by all the spectral tilt measurements) and rough (as indicated by higher jitter and shimmer but lower harmonicity) voice. Happiness, in contrast, is found to project a small body size with breathier and more harmonious voice. As argued in [18], with its greater spectral tilt, a breathy voice approximates a "pure tone" voice which is observed by Morton in animal calls associated with appeasement and sociability [8]. So this study is the first to show clear speech production support for the body size projection theory with voice quality data. Somewhat surprisingly, however, neutral emotion shows greater breathiness than happiness, which has not previously been predicted. It is possible that voice quality is affected not only by body size projection, but also by vocal effort, because anger, happiness and disgust are all more *activated* than neutral voice.

The non-significant difference in formant dispersion across the emotions is somewhat puzzling. It could be that the prediction of the body size projection theory is wrong, but it could also be due to the possibility that emotion portrayal method implemented in this study is not powerful enough to make speakers alter their vocal tract length when trying to produce emotional prosody. If so, it may suggest that voice quality is the most easily elicited emotion-relevant acoustic changes in speakers. But this needs confirmation from future research.

The higher pitch in anger than in happiness, as shown in Table 2b, may present a problem for the body size projection theory. But many previous studies have found that anger is associated with a higher pitch than neutral emotion [9, 12, 15]. This suggests that anger also projects high dynamicity through vocalization [17]. As speculated in [17], high dynamicity may serve to convince the listeners of the high energy the signaller possesses, thus helping to frighten them away. On the other hand, however, there are also different types of anger. As the participants in this study were instructed to portray "hot anger", nearly all of them managed to achieve it. In contrast, the happiness portrayed by the participants was more similar to calm happiness/pleasantness than to elation. This may further explain why the average pitch is higher for anger than for happiness.

The current results also show that disgust is similar to anger in terms of voice quality. This, together with the fact that the F_0 of disgust is the lowest among the four emotions, suggests that disgust involves the projection of a large body size, as has been speculated in [17].

5. Conclusions

In this study, we have used speech production experiments on Mandarin to test the body size projection theory of emotional speech. The results suggest that, firstly, tone and sentence position do not contribute significantly to the acoustic characteristics of emotional speech; secondly, disgust is similar to anger in terms of body size projection; thirdly and most importantly, among the features of emotional speech, only voice quality is fully consistent with the predictions of the body size theory: Pressed and rough voice, both projecting a large body size, is associated with anger; breathy voice, which projects a small body size, is associated with happiness. Pitch and formant dispersion did not generate results either in support or opposition to the body size projection theory. This could be due to insufficient authenticity of emotional speech under laboratory conditions, for which solutions need to be sought in future research. Overall, nevertheless, it is interesting, and in fact slightly surprising that voice quality turns out to be the vocal dimension that is the most easily elicited by the emotion portrayal method in this study. Naturally, further explorations in future research are needed.

6. References:

[1] Boersma, P. and Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program]. Version 5.3.59, retrieved 3rd Janurary 2013 from http://www.praat.org/

[2] Chen, Y. and Xu, Y. (2006). Production of weak elements: Evidence from neutral tone in Standard Chinese. *Phonetica*, 63, 47–75.

[3] Chuenwattanapranithi, S., Xu, Y., Thipakorn, B. and Maneewongvatana, S. (2008). Encoding emotions in speech with the size code—A perceptual investigation. *Phonetica*, 65, 210-230.

[4] Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6, 169-200.

[5] Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America*, 102, 1213-1222.

[6] Goudbeek M. and Scherer K. R. (2010). Beyond arousal: valence and potency/control cues in the vocal expression of emotion. *Journal of the Acoustical Society of America*, 128, 1322–1336.

[7] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities, *Acta Otolaryngologica*, 90, 441-451.

[8] Morton, E. W. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, 111, 855-869.

[9] Murray, I. R. and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93, 1097-1108.

[10] Noble, L. and Xu, Y. (2011). Friendly Speech and Happy Speech – Are they the same? In *Proceedings of The 17th International Congress of Phonetic Sciences*, Hong Kong: 1502-1505.

[11] Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica*, 41, 1-16.

[12] Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40, 227–256.

[13] Scherer, K. R. (2013). Vocal markers of emotion: comparing induction and acting elicitation. *Computer Speech and Language*, 27, 40–58.

[14] Van Bezooijen, R. (1984). *The characteristics and recognizability of vocal expressions of emotion. Dordrecht*. The Netherlands: Foris.

[15] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48, 1162-1181.

[16] Xu, Y. (2012). ProsodyPro.praat. University College London, London, UK.

[17] Xu, Y., Kelly, A. and Smillie, C. (2013a). Emotional expressions as communicative signals. In S. Hancil and D. Hirst (eds.) Prosody and Iconicity, John Benjamins Publishing Co, pp. 33-60.

[18] Xu, Y., Lee, A., Wu, W.-L., Liu, X. and Birkholz, P. (2013b). Human vocal attractiveness as signaled by body size projection. *PLoS ONE*, 8(4), e62397.

.

Scaling of Final Rises in German Questions and Statements

Jan Michalsky

Department of German Studies, University of Oldenburg, Germany

j.michalsky@uni-oldenburg.de

Abstract

Although certain intonation contours occur more frequently with German questions than with German statements, there is evidence that the semantics of intonational phonology operates on a more abstract level [1][2][3][4]. Hence, it is unlikely that there are pitch patterns in German that are exclusively used in interrogatives. Rather, intonational signaling of interrogativity can be regarded as resulting from the interaction between tonal and phonetic features. The tonal structure provides abstract semantic features, which are modified by paralinguistic features through phonetic realization [5]. This paper deals with the question which phonetic features may serve as cues to interrogativity in German. We report a reading task that was designed to elicit utterances that have phonologically identical nuclear rising pitch contours but differed by pragmatic function, serving either as a question or a statement. The observed absolute and relative scaling of nuclear and prenuclear tonal targets suggests that questions differ from statements by larger f0 excursions of nuclear rising contours, whereas the scaling of prenuclear accents does not substantially contribute to the expression of interrogativity. We conclude that phonetic cues to interrogativity in German are mainly realized through scaling and are restricted to the nuclear part of the intonational phrase.

Index Terms: intonation, interrogativity, questions, phonetic implementation, scaling, German

1. Introduction

There are two distinct levels of intonation to look at for specific features of German interrogatives: the categorical tonal structure and the continuous phonetic realization [6]. If interrogativity is regarded as a discrete pragmatic function, it may be assumed to be signaled by categorical intonational structure [7]. This view is prevalent in early descriptions of English [8], Dutch [9], and German [10], which attempt to identify pitch patterns that are linked to specific sentence types. Von Essen [10], for example, reports for German that most types of statements occur primarily with falling intonation, while certain question types occur primarily with rising intonation.

Spontaneous speech data suggest a less clear-cut distribution of pitch patterns in German: although certain intonation contours occur more frequently with specific types of questions than with statements, every sentence type may be realized with any contour of the German intonational inventory depending on the pragmatic context[11][3]. This finding is compatible with the majority of recent theories on the semantics of intonational phonology, where more abstract meanings are assigned to different intonation contours. Explanations regarding the phonology of question intonation and the difference between rising and falling contours in questions or across question types range from semantic features like *assertiveness* [12][13], over pragmatic features like *bias* [14][15] to discourse functions [11] and attitudinal features [16]. On a more abstract level, rising contours ending with a high boundary tone can be described as signaling pragmatic openness or incompleteness of the corresponding intonational phrase [17][2][3][18]. Hence, interrogativity cannot unambiguously be signaled by a specific choice of intonation contour. Rather, semantic features of intonation contours restrict the potential meanings of the associated utterance in a way that e.g. the feature of incompleteness suggests an interrogative interpretation in a certain context. This can explain why this type of intonation occurs more often with interrogative utterances like polar questions and to a certain degree wh-questions, but there is no immediate connection between the contour and the specific pragmatic function. However, if the phonology of intonation can signal certain aspects of interrogativity but not interrogativity itself the question arises whether this is the only contribution intonation can make to signal interrogativity in German, or whether there are other intonational cues, in particular at the level of phonetic realization.

According to Gussenhoven [5], intonational signaling of pragmatic functions may result from an interaction between phonological structure and phonetic realization. For example, emphatic focus may be signaled by a combination of a phonological pitch accent and an increase in its scaling. In this way, continuous variation in the phonetic realization can modify the abstract meaning of the phonological structure. This assumption is in accordance with reports that different types of increased pitch rather than a specific tonal pattern like a final rise were found as a potential universal feature of questions [19][20][21].

Evidence for the relevance of phonetic variation for the signaling of interrogativity has been found for several languages like Swedish [22][23], Finnish [24], Danish [25][26][27], and French [28] amongst others. For Dutch question intonation, Haan [29] identified several phonetic features such as higher f0 onset, raised register level, and differences in the scaling of the nuclear accent, which results in a shift in the overall global trend from declination to inclination. Eventually, phonetic effects of interrogativity were also found for German. Oppenrieder [30] observed the absence of overall declination in German questions, Brinckmann and Benzmüller [31] found differences in pitch range and f0 onset and Niebuhr et al. [32] suggest differences in shaping and alignment of prenuclear accents. Note, however, that in these experiments the potential phonetic features were not investigated independently of the phonological structure. A change of pragmatic function was always accompanied by a change in intonation contour like falling intonation in statements against rising intonation in questions [30][31][29][32]. Hence, an effect of the chosen contour on the phonetic realization cannot be excluded.

There are also several possibilities regarding the type of phonetic features involved and the domain of phonetic variation. Most research yielded effects in the scaling of tonal targets [22][23][27][24] but there is also some evidence for the

relevance of the shape of the transitions between tonal targets and of alignment relative to segmental landmarks [32]. As for the size of the affected constituent, there is evidence for the relevance of global parameters such as declination [25][26][28][30], pitch range [33][24][34], or register level [22][23], or local effects such as the scaling of the nuclear [29] or prenuclear peaks [32].

We report a reading task that was designed to investigate the phonetic effects of interrogativity in Standard German while keeping the tonal and grammatical structure constant. Since grammatically identical statements and questions can't be reliably elicited with the same intonational contour in utterance-final position we chose continuous statements and alternative questions were the nuclear rising contour occurs at the end of an intonational phrase with grammatically identical material but within an utterance. A possible generalization of the reported findings to utterance-final rising contours remains to be investigated.

In particular, we were interested in the identification of the relevant phonetic dimensions (scaling, shaping, and alignment) and domains of f0 variation (global, prenuclear or nuclear) that may be used for the distinction between questions and statements independently of grammatical and tonal structure. In the present paper we report results on the scaling of tonal targets only. Neither the phonetic variation of transitions between tonal targets nor their alignment or variation in pitch range or register level turned out to covary systematically with pragmatic function so far and remain open for further investigation.

2. Method

2.1. Speakers

The reading task was conducted with 21 speakers, 11 females and 10 males, aged between 18 and 30. The subjects were students from the University of Oldenburg and were born and raised in the northwestern part of Lower Saxony. All subjects were monolingual speakers of German.

2.2. Material

Two types of test sentences were constructed, questions and statements. Each sentence was designed to elicit two intonational phrases with the same intonation contour on the first phrase while keeping the grammatical structure identical. For the questions we used alternative questions and for the statements continuous statements with V1 word order. Subject and object items were filled with proper names to elicit two accents, a prenuclear and a nuclear one. The following examples illustrate both sentence types (for intonational annotation conventions see 2.4).

Alternative question:

(Will X nachher zu Y gehen)$_{IP}$, (oder bei Z bleiben?)$_{IP}$

H*L L*H H%|0%

Does X want to go to Y later, or stay with Z?

Continuous statement:

(Will X nachher zu Y gehen)$_{IP}$, (kann sie nicht bei Z bleiben.)$_{IP}$

H*L L*H H%|0%

If X wants to go to Y later, she cannot stay with Z.

Using a number of artificial proper names of the type *Mone* [ˈmoːnə] and *Mine* [ˈmiːnə] as accented words in the prenuclear and another set of artificial proper names of the type *Suse* [ˈzuːzə], *Söre* [ˈzøːʁə], *Neewe* [ˈneːvə] and *Narne* [ˈnaːnə] as accented words in the nuclear position we created 16 lexical variants of each sentence type. The proper names were segmentally controlled such that only voiced segments were used and all words ended in schwa to ensure a disyllabic production. We obtained a total of 32 target sentences, which were interspersed with 192 filler sentences and presented in a pseudo-randomized order.

2.3. Procedure

The test sentences were presented visually via a PowerPoint presentation with one sentence per slide. A line break was inserted after every potential intonational phrase to elicit a phrase boundary. The subjects were instructed to familiarize themselves with the sentence material in silence before they read them out aloud. Recordings were made in a sound booth in the speech laboratory at the University of Oldenburg with a portable digital recorder (Tascam HD P2) at a sampling rate of 48khz and 16bit resolution via a head mounted microphone (DPA 4065 FR).

2.4. Acoustic analysis

Only the first intonational phrase was selected for acoustic analysis. This phrase was expected to be realized with one of two possible nuclear target contours: a half-completed rise or plateau-contour L*H0% or a low-rising-contour L*HH%. Contours are represented according to the ToDI system [35] and its adaptation for German [3][18]. Equivalent notations in the classical ToBi [36] and GToBi [37] systems would be L*H-L% for the plateau-contour and L*H-H% for the low-rising-contour. The phrases contained two possible pitch accents: a prenuclear H*L accent on the subject-item in X-position and a nuclear L*H with a final boundary tone of H% or 0% on the object-item in Y-position and the following verb. Only utterances realized with the described tonal structure regarding number and position of pitch accents as a reflection of focus structure as well as choice of contour were selected for acoustic analysis.

Four points of measurement were determined as illustrated in figure 1: the beginning of the prenuclear rise, the prenuclear peak, the onset of the final rise and the offset of the final rise. From these measurements the following variables were calculated: The excursion of the rise to the prenuclear peak (*prenuclear rise excursion*), the excursion of the final rise (*final rise excursion*), the difference in the excursions (*excursion difference*) and in the peaks (*peak difference*). The excursions were determined by calculating the difference in frequency from prenuclear onset to prenuclear peak for the *prenuclear rise excursion* and from final rise onset to final rise offset for the *final rise excursion*. The *excursion difference* was calculated by subtracting the *prenuclear rise excursion* from the *final rise excursion*. The *peak difference* was calculated by subtracting the prenuclear peak from the final rise offset. For comparability of the two sexes the measurements were converted to a semitone scale.

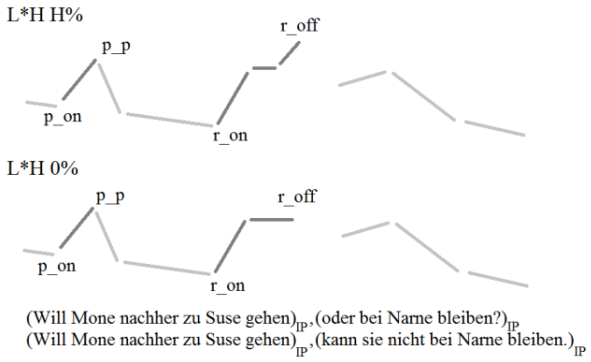


Figure 1: Points of measurement for the acoustic analysis for both target contours. (dark grey = prenuclear and nuclear rise excursion, p_on = onset of prenuclear rise, p_p = peak of prenuclear rise, r_on = onset of final rise, r_off = offset of final rise).

2.5. Statistical analysis

For the statistical analysis linear mixed effect models were used with PRAGMATIC FUNCTION (declarative / interrogative) and SPEAKER SEX (male / female) as fixed factors, and ITEM and SPEAKER as random factors. The dependent variables were *prenuclear rise excursion*, *final rise excursion*, *excursion difference*, and *peak difference*.

3. Results

Most of the intonational phrases were realized with the plateau-contour L*H0% or the low-rising-contour L*HH%. The female speakers produced 97 statements and 117 questions with plateau-contours, and 78 statements and 70 questions with low-rising-contours. The male speakers produced 159 statements and 140 questions with plateau-contours, and 65 statements and 83 questions with low-rising-contours. Four female speakers showed variation in their choice of contour while the other 7 female speakers and all 10 male speakers kept the chosen contours constant (see table1). The distribution of the two target contours shows that the choice of contour varies across speakers but doesn't systematically vary across pragmatic functions.

Speaker	Pragmatic Function	Plateau-contour	Low-Rising-contour
1	Statement	23 (67%)	11 (33%)
	Question	14 (66%)	7 (34%)
2	Statement	13 (48%)	14 (52%)
	Question	6 (40%)	9 (60%)
3	Statement	25 (81%)	6 (19%)
	Question	9 (45%)	11 (55%)
4	Statement	17 (54%)	14 (46%)
	Question	10 (45%)	12 (55%)

Table 1: Distribution of intonational contours across pragmatic functions for four female speakers.

Significant effects of PRAGMATIC FUNCTION were found for the excursion of the final rise in both L*H0% (declarative mean=5.9st, interrogative mean=7.9st, F=324.92, p<.001) and L*HH% (declarative mean=9.2st, interrogative mean=11.2st, F=109.24, p<.001) but not for SPEAKER SEX (L*H0%: female mean=6.7st, male mean=7.1st, F=0.34, n.s.) (L*HH%: female mean=8.9st, male mean=11.5st, F=2.01, n.s.), illustrated in figure 2.

No significant effects of PRAGMATIC FUNCTION could be found for the excursion of the prenuclear rise, neither for L*H0% (declarative mean=4.4st, interrogative mean=4.3st, F=0.06, n.s.) nor for L*HH% (declarative mean=4.1st, interrogative mean=4.0st, F=0.25, n.s.), and no effects of SPEAKER SEX (L*H0%: female mean=4.0st, male mean=4.6st, F=0.52, n.s.) (L*HH%: female mean=4.0st, male mean=4.1st, F=0.01, n.s.).

Consequently, there were significant effects of PRAGMATIC FUNCTION on the difference between final rise excursion and prenuclear rise excursion for L*H0% (declarative mean=1.5st, interrogative mean=3.6st, F=159.47, p<.001) and L*HH% as well (declarative mean=5.0st, interrogative mean=7.1st, F=68.10, p<.001) with no significant effects of SPEAKER SEX (L*H0%: female mean=2.6st, male mean=2.5st, F=0.001, n.s.) (L*HH%: female mean=4.6st, interrogative mean=7.5st, F=2.84, n.s.), illustrated in figure 3.

Finally, for the differences between the absolute final rise offset and the absolute prenuclear peak height, there were significant effects of PRAGMATIC FUNCTION for L*H0% (declarative mean=0.1st, interrogative mean=1.7st, F=200.51, p<.001) and L*HH% (declarative mean=3.4st, interrogative mean=5.2st, F=58.53, p<.001) with no significant effects of SPEAKER SEX in L*H0% (female mean=0.5st, male mean=1.0st, F=0.67, n.s.) but a small effect in L*HH% (female mean=2.7st, male mean=5.9st, F=7.67, p<.05), as illustrated in figure 4.

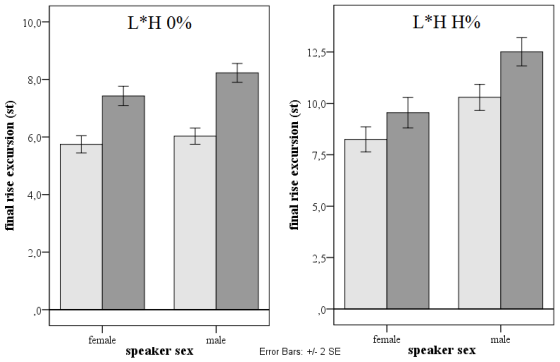


Figure 2: Phonetic effects of PRAGMATIC FUNCTION (light grey = declarative, dark grey = interrogative) and SPEAKER SEX on the *final rise excursion* in semitones for plateau- and low-rising contours.

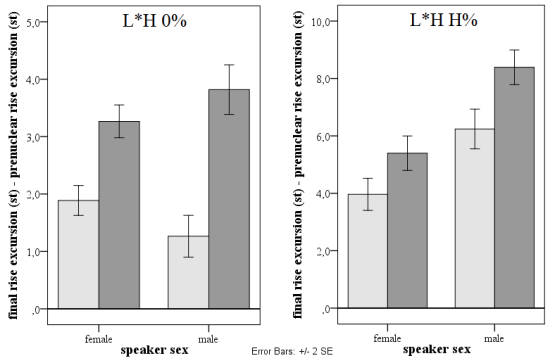


Figure 3: Phonetic effects of PRAGMATIC FUNCTION (light grey = declarative, dark grey = interrogative) and SPEAKER SEX on the *excursion difference* in semitones for plateau- and low-rising contours.

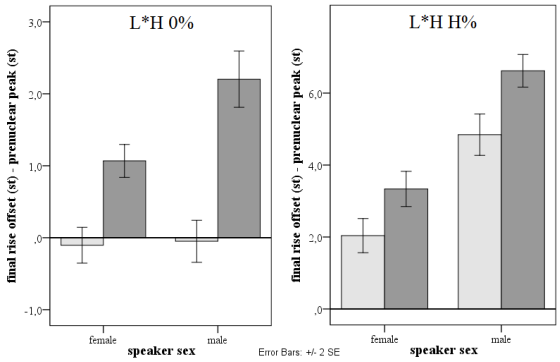


Figure 4: Phonetic effects of PRAGMATIC FUNCTION (light grey = declarative, dark grey = interrogative) and SPEAKER SEX on the *peak difference* in semitones for plateau- and low-rising contours.

4. Discussion and conclusion

The results suggest that the pragmatic function of interrogativity has significant effects on the scaling of intonation contours in German. This is in accordance with the assumption that in a typological perspective increased pitch is characteristic of interrogatives [19][20][21], as well as with experimental studies on several other languages [22][23][27][24][29].

The observed effects, however, were found to be restricted to the nuclear part of the utterance. This contradicts the majority of observations for other languages where interrogativity is found to either affect prenuclear and nuclear accents equally [22][23][27][24] or affect global parameters like declination [25][26][28], pitch range [33][24][34] or register level [22][23] in general. On the other hand, it is in accordance with Haan's [29] findings for Dutch. This suggests that Dutch and German may differ from other languages regarding the dimension of phonetic effects of interrogativity. It is, however, in contradiction to the observations of Niebuhr et al. [32] for German regarding both dimensions and types of effects. A first explanation for these differences may lie in the differences in keeping the phonological structure constant.

The variation of the nuclear part and the stability of the prenuclear part in interrogative utterances resulted in an increased difference between nuclear rise excursion and prenuclear rise excursion as well as between the absolute heights of the peaks of both rises. In other words, the local restriction of the phonetic effects also resulted in a global effect. The differences in absolute height of the peaks of both rises resulted in an even top-line for statements and an inclining top-line for questions. This is comparable to Haan's [29] upsweep, describing the amount by which the final rise exceeds the peak of the preceding nuclear accent peak. It is also compatible with the results from Brinckmann and Benzmüller [31] who observed a difference in the top-lines of statements and question-types.

Our results do not show, however, whether the actual perceptual relevant effect is 1) the difference in the excursion size of the final rise, 2) the relative difference of the excursion size of the final rise compared to the prenuclear rise, or 3) the difference in height of the peak of the final rise compared to the peak of the prenuclear rise. The first view would call for the listener's capability to compare the excursion size to some reference value in order to judge it as high enough to cue a question. The second view assumes the movement of both excursions as the perceptual cue. According to the third view, the perceptual cue would be whether the final rise's offset exceeds the prenuclear peak by a certain amount. The last two assumptions can but need not necessarily be connected because an increase in the excursion size of the final rise can be achieved by lowering the onset without increasing the height of the offset. To answer the question whether the extension of the final rise or the raising of the final offset alone provides sufficient cues to interrogativity, additional perception experiments will be carried out. It further remains to be tested whether the differences in excursion size of the final rise become more prominent when there is no prenuclear accent for comparison.

5. References

[1] Gussenhoven, C., On the grammar and semantics of sentence accents, Foris, 1984.
[2] Pierrehumbert, J. B. and Hirschberg, J., "The meaning of intonational contours in the interpretation of discourse", in P. Cohen, J. Morgan and M. Pollack [Ed], Intentions in communication, 271-311, MIT Press, 1990.
[3] Peters, J., Intonation deutscher Regionalsprachen, Walter de Gruyter, 2006.
[4] Petrone, C. and Niebuhr, O., "On the Intonation of German Intonation Questions: The Role of the Prenuclear Region", Language and Speech, 0(0):1-39, 2013. [in press]
[5] Gussenhoven, C., The phonology of tone and intonation, Cambridge University Press, 2004.
[6] Pierrehumbert, J., The Phonology and Phonetics of English Intonation, MIT, 1980.
[7] Ladd, D. R., Intonational Phonology, Cambridge University Press, 2008.
[8] Palmer, H. E., English intonation with systematic exercises, Heffer & Sons, 1922.
[9] de Groot, A.W., "De Nederlandse zinsintonatie in het licht der structurele taalkunde", De Nieuwe Taalgids, 37:30-41, 1943.
[10] von Essen, O., Grundzüge der hochdeutschen Satzintonation, Henn, 1964.
[11] Selting, M., Prosodie im Gespräch. Aspekte einer interaktionalen Phonologie der Konversation, Niemeyer, 1995.
[12] Bartels, C., The intonation of English statements and questions: a compositional interpretation, Routledge, 1999

[13] Truckenbrodt, H., "On rises and falls in interrogatives", in H. Y. Yoo and E. Delais-Russarie [Ed], Proceedings from IDP 2009, Paris, Septembre 2009, ISSN 2114-7612: 9-17, 2009.
[14] Kügler, F., "Do we know the answer? – Variation in yes-no-question intonation", in S. Fischer, R. Vogel and R. van de Vijver [Ed], Experimental studies in linguistics, Potsdam Universitätsverlag, 9-29, 2003.
[15] Kügler, F., "Dialectal variation in question intonation in two German dialects: the case of Swabian and Upper Saxon", in Proceedings of the second International Conference on Language Variation in Europe, 227-240, Uppsala University, 2004.
[16] Kohler, K., "Pragmatic and attitudinal meanings of pitch patterns in German syntactically marked questions", in G. Fant, H. Fujisaki, J. Cao and Y. Xu [Ed], From traditional phonology to modern speech processing. Festschrift for Professor Wu Zongji's 95th Birthday, 205-214, Beijing Foreign Language Teaching and Research Press, 2004.
[17] Cruttenden, A., "Falls and rises: meanings and universals", Journal of Linguistics, 17:77-91, 1981.
[18] Peters, J., "Intonation", in Duden – Die Grammatik, Kap. 2 (DUDEN-Reihe Bd. 4), 95-128, Bibliographisches Institut Mannheim, 2009.
[19] Hermann, E., Probleme der Frage. 2. Teil, Vandenhoeck & Ruprecht, 1942.
[20] Bolinger, D. L., "Intonation across languages", in J. H. Greenberg [Ed], Universals of Human Language. Band 2 (Phonology), 471-525, Stanford University Press, 1978.
[21] Ohala, J. J., "Cross-language use of pitch: An ethological view", Phonetica, 40:1-18, 1983.
[22] Hadding-Koch, K. and Studdert-Kennedy, M., "An Experimental Study of Some Intonation Contours", Phonetica, 11:175-185, 1964.
[23] Gårding, E., "A generative model of intonation", in A. Cutler and D. R. Ladd [Ed], Prosody:Models and Measurements, 11-25, Springer, 1983.
[24] Iivonen, A., "Intonation in Finnish", in D. Hirst and A. Di Cristo [Ed], Intonation Systems. A Survey of Twenty Languages, 311-327, Cambridge University Press, 1998.
[25] Thorsen, N., "A study of the perception of sentence intonation – Evidence from Danish", Journal of the Acoustical Society of America, 67(3):1014-1030, 1980.
[26] Thorsen, N., "Intonation and text in Standard Danish", Journal of the Acoustical Society of America, 77(3):1205-1216, 1985.
[27] Grønnum, N., "Superposition and subordination in intonation, a non-linear approach", Proceedings 13th International Congress of Phonetic Sciences, 2:124-131, 1995.
[28] Vaissière, J., "Language-Independent Prosodic Features", in A. Cutler and D. R. Ladd [Ed], Prosody: models and measurements, 53-66, Springer, 1983.
[29] Haan, J., Speaking of Questions: An Exploration of Dutch Question Intonation, LOT Graduate School of Linguistics, 2002.
[30] Oppenrieder, W., "Deklination und Satzmodus", in H. Altmann, A. Batliner and W. Oppenrieder [Ed], Zur Intonation von Modus und Fokus im Deutschen, 245-266, Niemeyer, 1989.
[31] Brinckmann, C. and Benzmüller, R., "The Relationship between Utterance Type and F0 Contour in German", Proceedings 6th European Conference on Speech Communication and Technology (EUROSPEECH), Budapest, Hungary, 21-24, 1999.
[32] Niebuhr, O., Bergherr, J., Huth, S., Lill, C. and Neuschulz, J., "Intonationsfragen hinterfragt - Die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax", Zeitschrift für Dialektologie und Linguistik, 77:304-346, 2010.
[33] Hirst, D. and Di Cristo, A., "A survey of intonation systems", in D. Hirst and A. Di Cristo [Ed], Intonation Systems. A Survey of Twenty Languages, 1-44, Cambridge University Press, 1998.
[34] Moraes, J. A., "Intonation in Brazilian Portuguese", in D. Hirst and A. Di Cristo [Ed], Intonation Systems. A Survey of Twenty Languages, 179-194, Cambridge University Press, 1998.
[35] Gussenhoven, C., "Transcription of Dutch Intonation", in S.-A. Jun [Ed], Prosodic typology: the phonology of intonation and phrasing, 118-145, Oxford University Press, 2005.
[36] Beckmann, M. E., Hirschberg, J. and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in S.-A. Jun [Ed], Prosodic typology. The phonology of intonation and phrasing, 9–54, Oxford University Press, 2005.
[37] Grice, M., Baumann, S. and Benzmüller, R., "German Intonation in Autosegmental-Metrical Phonology", in S.-A. Jun [Ed], Prosodic Typology: The Phonology of Intonation and Phrasing, 5-83, Oxford University Press, 2005.

Processing Prosodic Boundaries in Natural and Filtered Speech

Grace Kuo

Department of Linguistics, Macalester College, St. Paul, USA
gkuo@macalester.edu

Abstract

The prosody of an utterance can carry information that is critically important to understand the meaning of a sentence. Previous studies have shown that listeners are able to detect major prosodic boundaries in their native language in stimuli whose segmental information has been removed, such as low-pass filtered [1][2] and hummed speech [2][3][4][5]. The present boundary strength rating study is conducted on native and non-native speakers to Swedish, in an attempt to observe non-native speakers' accuracy in judging the upcoming boundary size in natural and filtered speech. 18 Taiwanese and 18 American English speakers were recruited for the rating task whose stimuli consisted of Swedish utterances from three prosodic boundary types (word boundary, phrase/tone sandhi group boundary, and Intonation Phrase boundary). In Experiment 1, participants rated the upcoming boundary strength on a slider for natural speech stimuli. In Experiment 2, they rated the boundary strength for filtered speech stimuli. The results show that both native and non-native speakers could accurately predict the upcoming prosodic boundary type in both natural and filtered speech. The acoustic analyses of duration, f0 range, f0 median, spectral tile, and harmonics-to-noise ratio reveal that both native and non-native speakers use these prosodic cues to make their judgment; however, they put different emphasis on different cues when they were presented with stimuli of different qualities (natural vs. filtered) and lengths.

1. Introduction

Previous studies of speech prosody have shown that listeners are able to predict upcoming prosodic boundaries. For example, Grosjean and Hirst [6] had the subjects listen to some part of an English sentence and asked them to predict how long the remaining sentence was. The results showed that English listeners were very accurate at predicting the amount of the rest of the sentence. However, their French listeners could only tell if a sentence ended, unable to differentiate between different amounts to come. Carlson, Hirschberg, and Swerts [7], on the other hand, found that English listeners were able to predict the strengths of the upcoming boundaries as well as Swedish listeners when the subjects were asked to express their judgment about the upcoming boundary in Swedish on a 5-point scale. This suggests that prosodic, rather than syntactic/semantic information was being used as a primary cue. In addition, the result of their follow-up study with Mandarin listeners showed that the length of the presented stimuli matters – Mandarin listeners could hear different Swedish boundaries only when presented with 2-second fragments, whereas English and Swedish listeners could differentiate the boundaries with either 2-second or one-word fragments. This finding indicates that language background affects the listeners' judgments.

Experiment 1 in the current study replicates [7]' experiment by recruiting American English and Taiwanese listeners to participate in a rating task, where half of the stimuli were from natural Swedish speech and the other half were its low-pass filtered version. It is predicted that American English listeners will accurately predict the Swedish upcoming boundaries with both 2-second and one-word fragments (as reported by [7]), and that Taiwanese listeners, speakers of another tone language, will predict the boundaries accurately only when presented with 2-second fragments (similar to [7]'s result with Mandarin speakers).

Previous study such as [1] [2] [8] showed that native listeners were able to detect major prosodic boundaries in meaningless speech materials, including re-iterant speech, low-pass filtered speech and hummed speech. Since [7] found that non-native listeners could make use of the prosodic information (in the absence of the syntactic/semantic information) to predict the upcoming boundary, Experiment 2 has listeners participate in the same rating task as in Experiment 1, but the segmental information in the stimuli was absent/reduced (low-pass filtered). It is predicted that no rating difference will be found between the natural speech stimuli and the filtered speech stimuli in that both the American English and Taiwanese listeners are able to use prosodic information to make the accurate judgments.

2. Experiment 1: Natural Stimuli

2.1. Method

2.1.1. Stimuli

The Swedish stimuli were the same stimuli [7] used in their experiment. The stimuli were obtained from a 25-minute interview with a Swedish female politician and the interview was manually annotated for perceived boundaries by three experienced transcribers. Every word was marked as being followed either by an IP boundary, a phrase boundary, or a word boundary. Sixty 2-second speech fragments followed by either of the three boundary types were chosen: 20 word boundaries (labeled as "no break" in later analysis), 20 phrase boundary ("weak break") and 20 IP boundaries ("strong break"). All stimuli came in two lengths, 2-second fragment and one-word fragment. From each 2-second fragment, the last word was extracted to be the one-word fragment.

Therefore, there were 120 utterances in total (20 items x 3 breaks x 2 fragment lengths).

2.1.2. Participants

Eighteen Taiwanese native speakers and eighteen American English native speakers participated in this experiment. None of the American English and the Taiwanese listeners had previous knowledge about or prior experience with Swedish. Neither of them had hearing or language problems according to their self-report.

2.1.3. Procedures

The subject individually judged the upcoming boundary strength for each natural utterance with an onscreen slider, whose position was manipulated by listeners from left ("small break") to right ("big break"). During the task, the subjects could choose to hear each stimulus more than once, but were encouraged to make their judgments by instinct. To minimize any possible learning effect, the stimuli were presented in a randomized order.

2.1.4. Data Analysis

The position of the slider bar was recorded by the Matlab script on a scale from 0-100. These numerical values were converted into logarithmic values to reduce the skewing in the distribution. The logarithmic strength ("log strength" hereafter) were entered into the repeated measures ANOVA, with the two within-subject factors, "Break" (no break vs. weak break vs. strong break), and "Length" (2-second vs. one-word). Since different language background might result in different rating results, English listeners' data were separated from Taiwanese listeners' data.

2.2. Results – Boundary Strength Ratings

2.2.1. English listeners

Repeated measures ANOVA reveals significant effects of "Break" ($F(2, 2154) = 74.32$, $p < .05$) and "Length" ($F(1, 2154) = 231.4$, $p < .05$). In other words, (i) listeners tended to give higher ratings for bigger boundaries, and (ii) listeners gave higher strength ratings for 2-seond fragments than for one-word fragments. In addition, listeners were able to differentiate all three breaks in either 2-second or one-word fragments. The results are shown in Figure 1. These findings were the same as the results found in [7] – English listeners were able to accurately predict all three boundaries in a nonnative language such as Swedish when presented with natural speech stimuli.

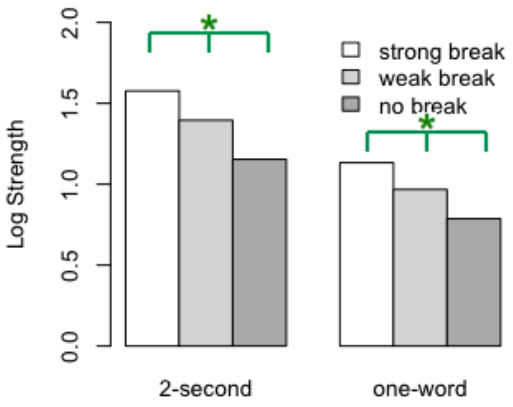


Figure 1: *English listeners' average logarithmic perceived boundary strength for Swedish Natural Stimuli. A significant difference between "Breaks" of either length is indicated with a line and asterisk above the bars.*

2.2.2. Taiwanese listeners

Significant effects were found not only in "Break" ($F(2, 2154) = 61.09$, $p < .05$) and "Length" ($F(1, 2154) = 196.88$, $p < .05$), but also in their interaction ($F(2, 2154) = 3.18$, $p < .05$). The results are shown in Figure 2. These findings were similar to [7]'s study with Mandarin listeners – tone language speakers were able to accurately predict the upcoming three boundaries only when they were presented with 2-seoncd fragments.

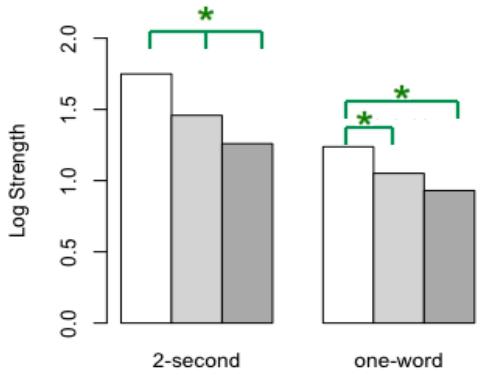


Figure 2: *Taiwanese listeners' average logarithmic perceived boundary strength for Swedish Natural Stimuli.*

3. Experiment 2: Filtered Stimuli

3.1. Method

The same 18 Taiwanese and 18 American English speakers participated in the boundary strength rating experiment. The stimuli were the low-pass filtered version of the 60 natural speech stimuli. They were generated by at a frequency cut-off of 400 Hz and 50 Hz smoothing, and the intensity was adjusted to 70 dB. The entire manipulation was done with a Praat [9] script. With low-pass filtering, most of the segmental information will be removed, yet the prosodic information, such as duration, f0, and some voice quality stay intact. Like Experiment 1, the stimuli also came in two lengths. Therefore, there were 120 utterances in total (20 items x 3 breaks x 2 fragment lengths). The experiment procedures were the same as those in Experiment 1.

In the analysis, the raw strengths are converted into the logarithmic strength, and the values were entered into the repeated measures ANOVA, which had "Break" and "Length" as the two factors.

3.2. Results – Boundary Strength Ratings

3.2.1. English listeners

The results with the English listeners show that when they were presented with filtered Swedish stimuli, they rated he three breaks differently, ($F(2, 2154) = 21.01$, $p < .05$). In addition, they were able to correctly predict all three breaks when presented with 2-second fragments, but not with one-word fragments. The results are shown in Figure 3. It seems that filtering has prevented listeners from identifying the phrase boundaries when the stimuli is short.

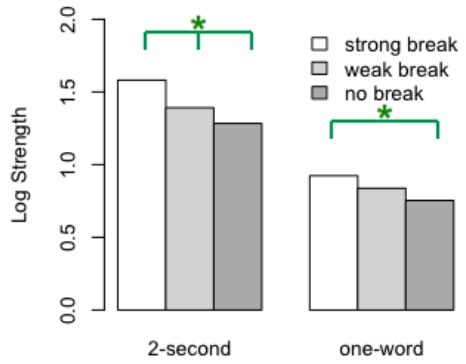


Figure 3: *English listeners' average logarithmic perceived boundary strength for Swedish Filtered Stimuli.*

3.2.2. Taiwanese listeners

Significant effects were found not only in "Break" ($F(2, 2154) = 15.67$, $p < .05$) and "Length" ($F(1, 2154) = 402.2$, $p < .05$), but also in their interaction ($F(2, 2154) = 3.31$, $p < .05$). Similar to the results with the English listeners, Taiwanese listeners were still able to accurately predict the upcoming Swedish breaks in 2-second fragments when the stimuli were low-pass filtered. However, when the filtered stimuli contained only one-word, there were no difference in ratings between the three breaks.

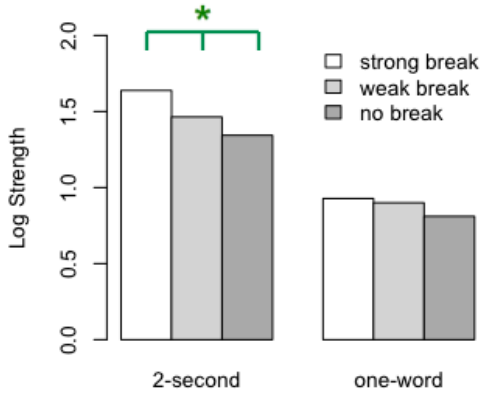


Figure 4: *Taiwanese listeners' average logarithmic perceived boundary strength for Swedish Filtered Stimuli*

4. Acoustic Correlates

In an attempt to identify the prosodic cues that could contribute to accurate boundary strength judgments in natural and filtered speech, we examined the acoustic measures from the last syllable of each stimulus: duration (=normalized vowel duration, and speech rate), pitch (including f0 range and f0 slope), harmonic amplitude/spectral tilt, harmonic-to-noise ratios, CPP, and Energy. The last syllable of each stimulus was labeled in Praat [9] and then the acoustic measures for the labeled portions were obtained using VoiceSauce [10]. The last syllable in Swedish contained usually part of a word.

As mentioned earlier, the filtered stimuli were a low-pass filtered version of the normal speech and the threshold was set at 400 Hz (as indicated with the vertical line in Figure 5), thus, any information beyond 400 Hz would have been filtered out. For the filtered stimuli, the main available voice measures were the amplitude of the first harmonic (corrected H1), HNR05 (frequency range <500 Hz) and CPP.

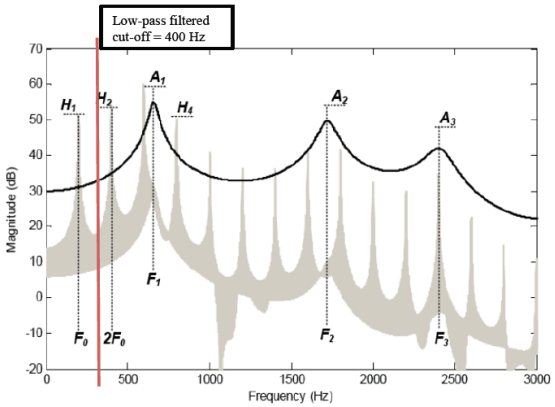


Figure 5: *An example spectrum. Adopted from [11]. The envelope in shade is the actual output from VoiceSauce [10]. The corrected values are obtained from this envelope.*

The regressions between the acoustic measures and the logarithmic boundary strength ratings are made. Multiple regression analysis was carried out for each type of stimulus (considering speech quality and the fragment length). The natural stimuli results are shown in Table 1 and the filtered stimuli results are shown in Table 2.

Table 1: *English and Taiwanese listeners listened to Swedish natural stimuli: marks show which acoustic measures contributed significantly to the regression equation.*

	English		Taiwanese	
	2-sec	1-wrd	2-sec	1-wrd
Duration				
Rate	✓	✓	✓	
F0 range	✓	✓	✓	✓
F0 median	✓		✓	✓
F0 mean	✓		✓	
H1*-H2*		✓		
HNR05				✓
CPP		✓		✓

The results show that for both English and Taiwanese listeners, when they were presented with 2-second natural stimuli, they would pay attention to durational and f0 cues. When the presented stimuli was one-word fragment, they would take voice quality into consideration.

Table 2: *English and Taiwanese listeners listened to Swedish filtered stimuli: marks show which acoustic measures contributed significantly to the regression equation.*

	English		Taiwanese	
	2-sec	1-wrd	2-sec	1-wrd
Duration			✓	
Rate	✓		✓	
F0 range			✓	✓
F0 median			✓	✓
F0 mean	✓	✓	✓	
H1*	✓	✓	✓	
HNR05	✓	✓	✓	✓
CPP				✓

The results reveal the significant correlations between the acoustic measures from the last syllable of the filtered utterances and the logarithmic boundary strength. It seems that English listeners rely on more voice quality cues for both lengths whereas Taiwanese listeners, who had better boundary strength when presented with 2-second fragments, tended to make use of more cues, including durational measures, to predict the upcoming boundaries. In addition, tone language speakers, such as Taiwanese, also used f0 range as a reliable cue when they were asked to predict upcoming boundary in a non-native language. Pitch is not only for lexical use for Taiwanese listeners.

5. Conclusion

In this study, we examined the perceived boundary strength indicated by Taiwanese and English listeners presented with Swedish natural and filtered stimuli. The distribution of the perceived boundary strengths shows that English listeners showed a three-way distinction in breaks in normal (both 2-seoncd and one-word) and filtered (only 2-second) stimuli. Taiwanese listeners also showed a three-way distinction in breaks in normal and filtered stimuli, but only when they were presented with 2-second fragments. These findings are consistent with [7]'s findings.

The acoustic analyses of normalized vowel duration, speech rate, f0 range, f0 slope, spectral tile, and harmonics-to-noise ratio reveal that non-native speakers use these prosodic cues to make their judgment; however, they put different emphasis on different cues when they were presented with stimuli of different qualities (natural vs. filtered).

6. Acknowledgements

The author would like to acknowledge Prof. Rolf Carlson, Julia Hirschberg and Marc Swerts for generously sharing their Swedish stimuli.

7. References

[1] de Rooij JJ. "Prosody and the Perception of Syntactic Boundaries", IPO ANNU Prog Rep, 10:36-39, 1975.
[2] Kreiman J. "Perception of sentence and paragraph boundaries in natural conversation", Journal of Phonetics, 10:163-175, 1982.
[3] t'Hart J., Collier, R., and Cohen, A.,"A perceptual study of intonation", Cambridge University Press, 1990.
[4] Pan, Ho-hsien, " Perceptual Tone Spaces and Taiwan Min Sandhi Rules", *personal communication*, 2011.
[5] Pannekamp, A., Toepel, U., Alter, K., Hahne, A., Friederici, AD., "Prosody-driven sentence processing: an event-related brain potential study", J. Cogn Neurosci, 17: 407-421, 2005.
[6] Grosjean, F. and Hirst, C. Using prosody to predict the end of sentences in English and French normal and brain-damaged subjects. *Language and cognitive Processes*, 11: 107-134, 1996.
[7] Carlson, R. , Hirschberg, J. and Swerts, M. Cues to upcoming Swedish prosodic boundaries – subjective judgment studies and acoustic correlates, *Speech Communication* 46: 326-333, 2005.
[8] de Rooij JJ. "Perception of prosodic boundaries", IPO Annu Prog Rep. 11: 20-24, 1976.
[9] Bowersma, P. and Weenink, D. Praat, Doing phonetics by computer (version 5.2.25) [Computer program]. Retrieved from http://www.praat.org, 2012.
[10] Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. "VoiceSauce, a program for voice analysis", *Proceedings of the 17th ICPhS*, 1846-1849. 2011.
[11] Shue, Y.-L. "The Voice Source in Speech production: Data, Analysis and Models". Doctoral dissertation, University of California, los Angeles. 2010.

Constant Tonal Alignment in Swedish Word Accent II

Malin Svensson Lundmark

Linguistics and Phonetics, Lund University, Lund, Sweden
Malin.svenssonlundmark@gmail.com

Abstract

Studies on accentual tonal alignment of intonation languages suggest that L in rising (LH) pre-nuclear accents anchors with a specific point in the segmental string, while the timing of H varies. This study investigates if lexical accents, too, exhibit a constant alignment by testing the South Swedish word Accent II. When under the strain of tempo variability the L-target was found not to be anchored with syllable onset. The results were not fully conclusive regarding H, but no clear evidence was found against anchoring of H, which could mean that H is an important phonological event in Accent II, while L is not.

Index Terms: tonal alignment, segmental anchoring, word accent, pre-nuclear accent, speech rate

1. Introduction

Over a period of 15 years there has been an on-and-off debate within intonational phonology on whether or not accentual tonal targets (L, H) are constantly aligned with the segmental string. Most studies have focused on pre-nuclear rising accents in intonation languages and have found conclusive results on the start of the rise, the L-target. So far, a language with lexical accents has not been taken into account in the recent research on constant tonal alignment.

1.1. Tonal alignment

Tonal alignment might be seen upon as a wider notion for other concepts such as timing, tonal association or segmental anchoring. The *segmental anchoring principle* presupposes that tonal targets are constantly aligned, and thus anchored at specific points in the segmental string [1]-[3]. Studies that second the principle have focused on rising pre-nuclear accents in intonation languages. Previous studies in the field of constant tonal alignment have displayed an unambiguous case of the tonal target L aligning with syllable onset in pre-nuclear accents, though the precise timing seems to vary across languages. Results include the L-target occurring just before the onset of the accented syllable ([1] for Greek; [4] for Italian), at syllable onset ([5] for Dutch; [3] for English; [2] for German), or after syllable onset ([6] for Mandarin).

While the studies show anchoring of L with the beginning of the syllable, the same consistent result does not exist for the H-target. Some studies found H anchoring after syllable offset [1], [2], [6], or somewhere late in the syllable [3]. Caspers and Van Heuven [5] found that the end of the rise, the H-target, varied considerably under time pressure and thus rejected that it did anchor with the segment. However, they did also consider whether or not the variation had to do with segmental structure.

It can be concluded that the L-target appears to be more stable than the H-target in rising pre-nuclear accents in intonation languages. Niemann et al. [4] has suggested that the cross-linguistic variation of anchoring of L is systematic and can be explained by different phonological structures between the intonation languages. The inconclusive results on H have been highlighted by Niebuhr et al. [7] who addressed the effect of individual speaker strategies. They also proposed that instead other related features are responsible for the consistent results of segmental anchoring.

1.2. Swedish accents

In the prosodic typology of Swedish intonation provided by the Lund Model [8], [9], the two Swedish word accents are assumed to be represented by a fall associated with the stressed syllable in a prosodic word, where the two accents differ in the timing of the fall. There is a regional variation between the Swedish dialects. For example in the South Swedish dialect (South) both accents are timed considerably later than in the Central Swedish dialect (Svea): in South Swedish the high level in Accent I is associated with the stressed vowel and in Accent II with offset of the stressed syllable.

The original dialect typology has later been revised by Bruce [9], who identified, for all dialects, an LHL tonal gesture from which bitonal gestures are extracted; either a fall, H+L, or a rise, L+H. Bruce generalized for Accent II an association of a fall in the dialects with an early timing of the accents (Svea and Göta) and a rise in the dialects with a late timing (South, Gotland, Dala, North). For the South Swedish dialect, a late timed dialect type, Bruce made the specific assumption of a fall, an H+L pattern, for Accent I and a rise, an L+H pattern, for Accent II. The rise in South Swedish Accent II has indeed been shown to be relevant from a perceptual point of view [10].

As if by chance, there is a rough phonetic match between the timing of the lexical Accent II in South Swedish and the pre-nuclear accents in the already mentioned studies on tonal alignment in intonation languages. The research question formulated here is whether or not additional phonetic features are similar such as if L and H are anchored with a segment, as is assumed by the segmental anchoring principle.

The present study is a production study where the hypothesis of segmental anchoring is tested on the Swedish word Accent II. Speech rate is used as an experimental tool and is based on the idea that speakers will try to retain primary features of phonological properties, while they will let other features be modified under time pressure [5]. Speech rate has been used successfully by a number of researchers in studies concerning tonal alignment [3], [5], [6]. Because the Lund Model (revised by Bruce [9]) assumes that both L and H in the rising L+H gesture of Accent II are phonologically relevant, anchoring of L and H is expected.

2. Method

2.1. Speakers and recording

The material was initially collected for a different study in which two age groups were recorded. For this study only the older speakers were tested due to technical issues. There were

seven speakers, four males and three females, and the average age of the speakers was 72 years. All speakers were voluntary and spoke the same variety of the South Swedish dialect. A criterion for speaker selection was that they had all lived most of their lives in the same area in the northeastern part of the South Swedish region. Moreover, their parents also had to have lived most of their lives in the area.

All of the recordings were made in people's homes. An IMG Stage boundary microphone (table-microphone) with phantom power was used (ECM-302B) since it is non-invasive and the speakers were expected to be naïve with no prior recording experiences.

The material was read twice by each speaker at three different speech rates: normal, slow and fast. The recording leader set the pace of the speech rate with the leading question and the speaker was asked to answer the question and to follow the speech rate of the recording leader.

2.2. Speech materials and data processing

The materials consisted of three test sentences with the same test word: *många*, meaning 'many' in English. The materials were mixed with 37 further sentences not investigated here. The test word, in its three sentence contexts, fits the following criteria: an unbroken tonal curve, a word Accent II, identical segmental surroundings ([9 syllables] bisyllabic target word [2 syllables]) and that neither syllable and vowel onset, nor syllable and vowel offset coincided. The leading question sought to that the target word *många* [ˈmɔŋːa] 'many' occurred before nuclear accent in all three sentences. The F0 contour of an example sentence can be seen in Figure 1.

The author performed segmentation and annotation in Praat [11]. Since each speaker was recorded twice, the material consisted of 126 items (3 sentences x 2 repetitions x 3 speech rates x 7 speakers). Each target word was segmented into syllables, and in addition, the boundaries of the accented vowel were determined. The boundary between the two syllables was defined as the temporal midpoint of the long, ambisyllabic consonant. The tonal curve has been semi-automatically annotated for the tonal targets L and H by the author (Figure 2). Extracted measures were the start and the end of the rise (L and H), syllable onset and syllable offset.

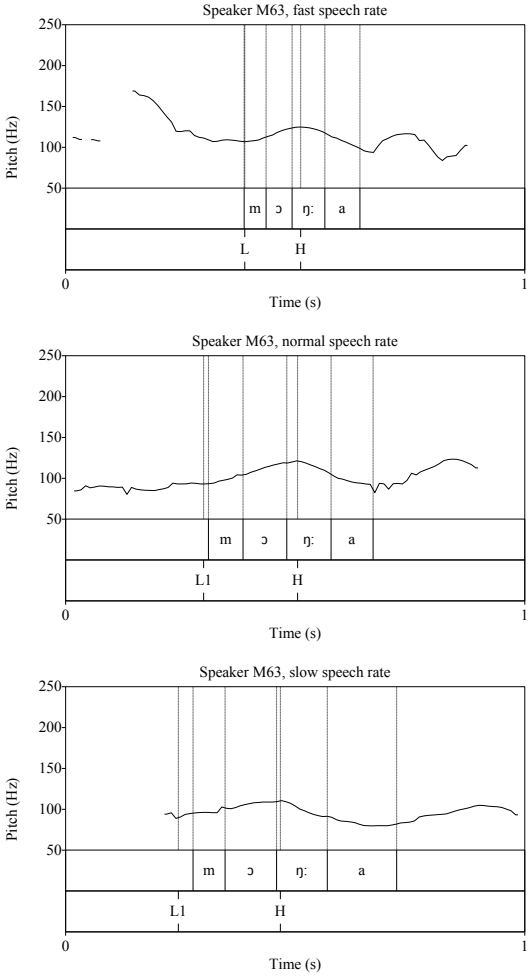


Figure 2: *The segmented target word [ˈmɔŋːa] in three different speech rates, spoken by male speaker M63. The low (L/L1) and the high target (H) of the tonal curve has been annotated.*

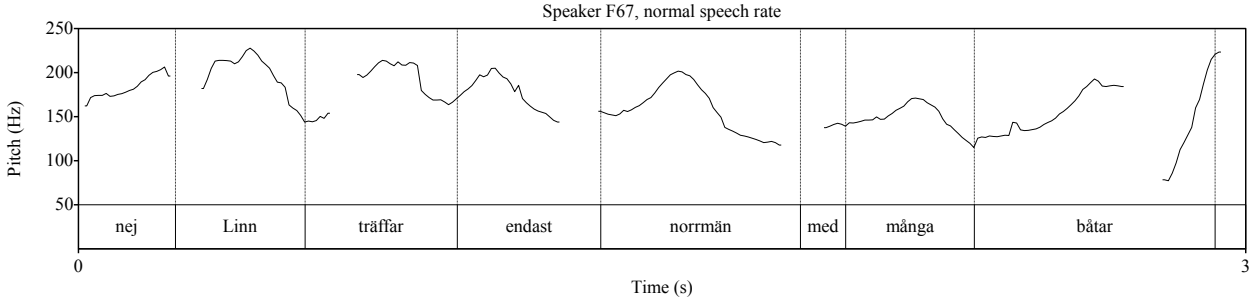


Figure 1: *One of the three target sentences in the material as spoken by female speaker F67 in normal speech rate. The sentence translated into English is: 'No, Linn only meets Norwegians with many boats'.*

3. Results

Average syllable duration shows a difference in speech rate between the recordings (Figure 3). An ANOVA confirmed that speech rate had a significant effect on syllable duration ($F = 66.490$, df 1, $p < .001$), concluding that the rate manipulation was successful. Since the segments are affected by speech rate, the temporal distance between the tonal targets L and H should also be affected by speech rate, if they are anchored with the segmental string. An ANOVA was run and showed a significant effect of speech rate on the temporal distance between L and H (henceforth, rise time) ($F=17.129$, df 1, $p = .006$) resulting in shorter rise times for faster speech.

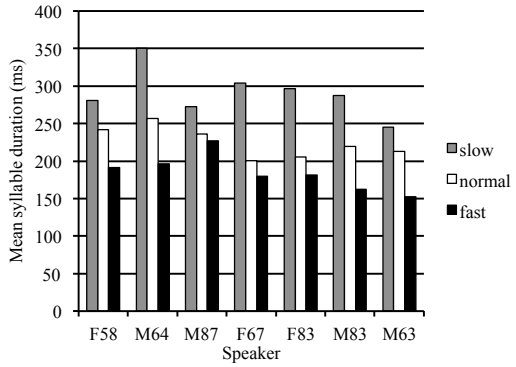


Figure 3: *Average syllable duration for each speaker in each speech rate.*

If anchoring of the tonal targets with specific points in the segmental string occurs this would necessitate a correlation between segment duration and distance between tonal targets. This was tested by means of a Correlations Pearsons (2-tailed) test. There appears to be a weak to moderate positive linear relationship ($R= 0.433$, $N= 74$), which indicates a correlation. However, it does not seem to be a convincingly strong correlation (Figure 4).

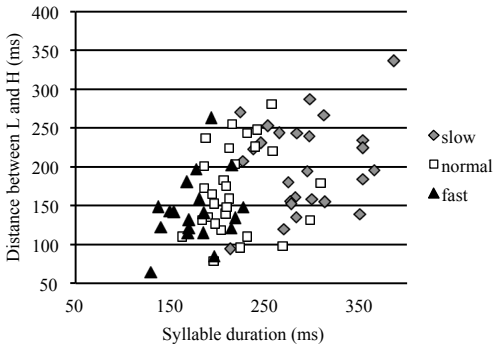


Figure 4: *Scatter plot of the relationship between syllable duration and the distance between tonal targets L and H.*

To test whether the weak to moderate relationship might indicate that either only one or neither of the targets is anchored, two new measures were calculated: the distance between L and syllable onset, and the distance between H and syllable offset, where anchoring is measured as distance in milliseconds. An ANOVA was first run with speaker as random sample. To account for missing data, an average value was first calculated for each speaker across the available items for each condition. The ANOVA showed no significant effect of speech rate on the distance between the tonal target L and syllable onset ($F =0.460$, df 1, $p = .523$). An ANOVA was also run on the distance between H and syllable offset, showing no significant effect of speech rate on anchoring of H ($F = .702$, df 1, $p = .434$). However, Figure 5 sends a different message and displays a large variance of anchoring of L between the speakers comprising alignments before as well as after syllable onset. The H-target also varies considerably in its alignment, but somewhat less than the L-target (see also Table 1). Notably, the H-target is consistently aligned before syllable offset, on average around 40 ms for both normal and fast speech (Table 1).

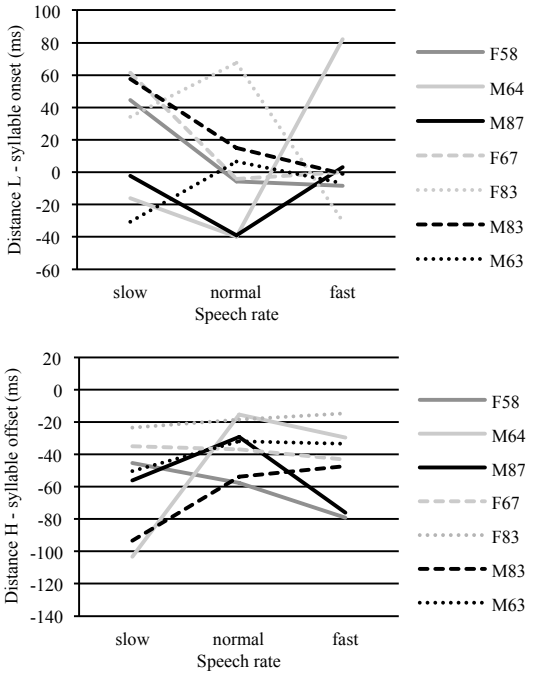


Figure 5: *Distance between tonal target and segment boundary, average value for each speaker. The first graph shows distance between L and syllable onset, the second between H and syllable offset. A negative number indicates that timing is before the boundary.*

Table 1: *Distance between tonal target and segment boundary (ms), average value for each speaker. Standard deviation in parentheses. A negative number indicates that the target is before the boundary. * Only one item in this condition.*

	L - syllable onset			H - syllable offset		
	slow	**normal**	**fast**	**slow**	**normal**	**fast**
F58	45 (47)	-6 (97)	-9 (1)	-45 (18)	-58 (14)	-79 (32)
M64	-16 (22)	-40 (*)	82 (*)	-103 (49)	-16 (*)	-29 (*)
M87	-2 (43)	-39 (44)	3 (*)	-56 (32)	-29 (48)	-76 (*)
F67	62 (56)	-4 (9)	0 (4)	-35 (21)	-37 (10)	-43 (17)
F83	34 (*)	67 (39)	-30 (28)	-24 (*)	-19 (13)	-15 (18)
M83	57 (26)	15 (38)	-1 (45)	-93 (9)	-54 (20)	-47 (7)
M63	-31 (36)	7 (29)	-7 (8)	-50 (26)	-32 (25)	-33 (31)
All	24 (51)	9 (54)	-6 (36)	-65 (37)	-38 (23)	-40 (27)

The anomaly of only one available item in some conditions for three of the speakers (M64, M87 and F83) can be seen in Table 1. In order to avoid a type II error, additional ANOVAs were made with target words as random sample.

The ANOVA with target word as sample showed that speech rate did in fact have a significant effect on the distance between L and syllable onset ($F = 14.095$, df 1, $p = .013$) suggesting that L does not seem to be constantly aligned at or close to the syllable onset. This conclusion is also supported by Figure 6, which, again, displays a large spectrum of alignments both before and after syllable onset. An ANOVA on the distance between H and syllable offset was also calculated which shows a low p-value; however not statistically significant ($F = 5.159$, df 1, $p = .072$). Speech rate appears to not affect the possible anchoring of H. Average value and standard deviations for each target word are shown in Table 2.

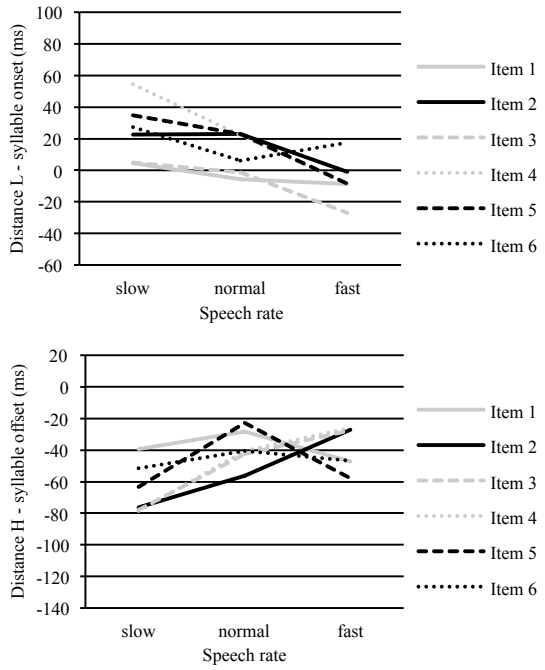


Figure 6: *Distance between tonal target and segment boundary, average value for each target word. The first graph shows distance between L and syllable onset, the second between H and offset. A negative number indicates that timing is before the boundary.*

Table 2: *Distance between tonal target and segment boundary (ms), average value for each target word. Standard deviation in parentheses. A negative number indicates that the target is before the boundary.*

	L - syllable onset			H - syllable offset		
	slow	normal	fast	slow	normal	fast
1	4 (22)	-6 (37)	-9 (37)	-39 (32)	-28 (25)	-47 (37)
2	23 (65)	23 (51)	-1 (18)	-76 (25)	-56 (19)	-27 (37)
3	5 (50)	-2 (59)	-27 (34)	-78 (54)	-42 (19)	-28 (17)
4	54 (56)	22 (78)	-9 (11)	-79 (39)	-40 (21)	-26 (22)
5	35 (50)	23 (59)	-8 (9)	-63 (33)	-23 (17)	-58 (16)
6	27 (64)	6 (60)	18 (59)	-52 (30)	-40 (29)	-47 (18)
All	24 (51)	9 (54)	-6 (36)	-65 (37)	-38 (23)	-40 (27)

4. Discussion

This study seems to not support the anchoring of L in the L+H rise of South Swedish Accent II. The rise is surely an important feature of the word accent [10], but if the start of the rise is not constantly aligned it is possible that L is not a phonological event. The end of the rise, the timing of H, might be an important phonological feature. Independent of syllable duration H was aligned within the syllable, on average 40 ms before syllable offset in the case of normal and fast speech, which supports the Lund Model and the accent typology that incorporates the South Swedish dialect. The study would benefit from relative measures of syllable duration, not the least to further establish the location of H, but also to shed light on the variability of L which affected by speech rate occurs both before and after syllable onset (Figure 5 and 6). Future studies would also benefit from addressing the precision of tonal peak measures (a problem in many comparative studies [1]-[3], [5], [6]), and use alternative methodology, such as Tonal Center of Gravity (TCoG) [12].

The data displayed a great variability both between and within speakers. Ladd et al. [3] also observed a similar degree of variability. By excluding certain speakers that seemed to use a different strategy to define pitch accent, they were able to find support for segmental anchoring. It might be that constant alignment is a strategy only for some speakers or that the same speaker uses different strategies for alignment. The proposition by Niebuhr et al. [7] to include speaker strategy in studies on tonal alignment is thus a valid suggestion.

The auxiliary hypothesis that primary features of phonetic properties will try to be retained by speakers, while other features will be allowed to be modified by time pressure might be the case for the normal and the fast rate. The slow rate, however, seemed to divert from the others, which can be seen in the scatter plot with the slow rate being much more scattered than the normal or the fast rate (Figure 4). The anomalies found on slow speech rate have been reported in other studies as well, where difficulties with the slow speech rate seem to have brought forth additional prosodic features to enable the, perhaps, unnaturally slower speech [3]. Even though the results of the study confirmed that the manipulation of speech rate was successful, a future use of speech rate as an experimental tool needs to be further investigated.

The coincidence was pointed out that the rise of the pre-nuclear accent in intonation languages phonetically roughly matched the lexical Accent II in South Swedish. The results, however, did not confirm a phonological match. The start of a rising pre-nuclear accent in an intonation language needs to be anchored, but this does not seem to be the case for a South Swedish Accent II rise. Since evidence was found against L anchoring with syllable onset, the results do not support the revised Lund Model of a LH gesture. Further studies on the anchoring of the LHL tonal gesture in Swedish Accent II are suggested.

5. Acknowledgements

This paper is based on an unpublished master thesis. While working with an earlier version of the thesis I was supervised by professor Gösta Bruce and associate professor Hugo Quené. I am very grateful for having had their excellent supervision in the initial steps of the study. I also want to extend a big thank you to my supervisor Gilbert Ambrazaitis for his invaluable help, advices and enthusiasm.

6. References

[1] Arvaniti, A., Ladd, D. R. and Mennen, I., "Stability of tonal alignment: the case of Greek prenuclear Accents", Journal of Phonetics, 26:3-25, 1998.

[2] Atterer, M. and Ladd, D. R., "On the Phonetics and Phonology of 'Segmental anchoring' of F0: Evidence from German", Journal of Phonetics, 32:177-197, 2004.

[3] Ladd, D. R., Faulkner, D., Faulkner, H. and Schepman, A., "Constant 'Segmental anchoring' of F0 movements under changes in speech rate", The Journal of the Acoustical Society of America, 106(3):1543-1554, 1999.

[4] Niemann, H., Mücke, D., Nam, H., Goldstein, L. and Grice, M., "Tones as Gestures: the Case of Italian and German", ICPhS XVII Proc., Hong Kong, 1486-1489, 2011.

[5] Caspers, J. and Van Heuven, V. J., "Effects of time pressure on the phonetic realization of the Dutch Accent-lending pitch rise and fall", Phonetica, 50:161-171, 1993.

[6] Xu, Y., "Consistency of tone-syllable alignment across different syllable structures and speaking rates", Phonetica, 55:179-203, 1998.

[7] Niebuhr, O., D'Imperio, M., Fivela, B. G. and Cangemi, F., "Are there 'Shapers' and 'Aligners'? Individual differences in signaling pitch accent category", ICPhS XVII Proc., Hong Kong, 120-123, 2011.

[8] Bruce, G. and Gårding, E., "A prosodic typology of Swedish dialects", in E. Gårding, G. Bruce and R. Bannert [Ed], Nordic prosody, Papers from a symposium, 219-228, Lund, 1978.

[9] Bruce, G., "Components of a prosodic typology of Swedish intonation", in T. Riad and C. Gussenhoven [Ed], Tones and Tunes, Volume 1: Typological Studies in Word and Sentence Prosody, 113-146, Berlin, 2007.

[10] Ambrazaitis, G. and Bruce, G., "Perception of south Swedish word Accents", Working papers, 52:5-8, Lund, 2006.

[11] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer [Computer program]", version 5.2.03. Online: http://www.praat.org/, accessed on 24 November 2010.

[12] Barnes, J., Veilleux, N., Brugos, A. and Shattuck-Hufnagel, S., "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology", Laboratory Phonology, 3(2): 337-383, 2012.

A simplified version of the OpS algorithm for pitch stylization

Antonio Origlia, Francesco Cutugno

LUSI-Lab, Dept. of Electrical Engineering and Information Technology
University of Naples "Federico II" - Italy
antonio.origlia@unina.it, cutugno@unina.it

Abstract

In this work we present a new version of our previously published Optimal Stylization (OpS) algorithm for pitch stylization. Here we give a better perceptual representation of the pitch curve for linguistics research. While the OpS algorithm produced good stylizations for naive listeners, when deployed in a prosodic analysis tool, we observed that, under specific conditions, important details were missed in the stylized curve to an expert's ear. Changes introduced in the dynamic tonal perception model to solve these problems resulted in a simpler and more robust model. We show how the new version of the OpS algorithm is able to recover these situations while not significantly altering the original OpS curves.

Index Terms: pitch stylization, tonal perception, prosody

1. Introduction

Prosodic research focuses on messages transmitted through the use of intonational strategies. While the $F0$ curve is indeed the main correlate of intonation, it does not represent what it is actually *heard* by the human ear. In [1, p. 25], it was stated that *No matter how systematically a phenomenon may be found to occur through a visual inspection of F0 curves, if it cannot be heard, it cannot play a part in communication*. This led to the definition of *stylization* as an approximation of the $F0$ curve by means of linear segments. In [1, p.42], this was defined as a sequence of segments that *[. . .] should eventually be auditorily indistinguishable from the resynthesized original and it must contain the smallest possible number of straight-line segments with which the desired perceptual equality can be achieved*.

Among the attempts to produce an account of the intonational account, the MOMEL algorithm [2] has been widely used in the literature. This algorithm does not produce a *proper* stylization as its goal is to produce a model of the macroprosodic component, which can be used together with the microprosodic component the algorithm produces to rebuild the original pitch curve. In this sense, a stylization should be intended as a lossy filter for microprosody while the output of the MOMEL algorithm does not discard the microprosodic component. Nevertheless, the macroprosodic profile obtained with MOMEL is usually considered as reference for stylization algorithms.

In [3], the concept of dynamic tones, or glissandos, was used in order to produce a stylization of the pitch curve. The Prosogram, a perceptually motivated representation of the pitch curve [4] is based on this algorithm. This representation includes a segmentation of the considered utterance into syllables to represent the pitch curve in terms of glissandos and static tones. In [5, 6], the concept of syllables was used again to position the linear segments used in the stylization. In [7], the pitch stylization problem was treated as an optimization problem for the first time by using a Dynamic Programming algorithm designed to optimize the position of a predefined number of segments estimated on the basis of the findings presented in [8]. As a quality measure, this algorithm used the statistical closeness between the stylized curve and the original one.

In [9], we presented the OpS framework along with an investigation of the possibility of using prominence information to reduce the number of points used to stylize non-prominent areas. The OpS algorithm uses a *divide et impera* strategy to balance a cost measure, based on the number of points used by the stylized curve, and a quality measure. In [9], we showed that statistical closeness does not necessarily reflect the results of the listening test so, in [10], we presented an updated version of the algorithm using a tonal perception model to compute the quality measure. While this model has the same basis of [3], it is dynamic in the sense that it uses the findings of [4] and the indications coming from the experiments with the OpS version using prominence annotation to avoid using rigid thresholds. Also, by retaining the generic OpS framework, it explicitly takes into account the cost of the curve during computation, closely following the definition of stylization.

In this paper, we summarize the parts of the OpS algorithm that have been modified to obtain the new version, we highlight the problems that the algorithm had in retaining certain classes of details that are important for linguistics research and we present the changes we introduced. Qualitative and quantitative tests performed on the same corpus we used for the objective tests in our previous works show that these changes do not alter the OpS curve on a large scale. By means of a case study, we show that the details we were interested in recovering are correctly represented by the new version of the algorithm.

2. The OpS algorithm

In [10], we presented a new version of the OpS algorithm substituting the original quality measure $q(S, \bar{S})$ with a new measure based on a tonal perception model, following the approach of [3]. This tonal perception model was dynamic in the sense that it did not use rigid thresholds to model the human ear's capabilities of perceiving dynamic tones (glissandos). This was achieved by considering the effect energy movements have on tonal perception (i.e. [11]) by taking as reference the Spectral Constraint Hypothesis (SCH) [12] and by relying on a continuous value to describe the *glissando likelihood* Γ_g of a pitch movement based on the findings of [4]. The reader is referred to [10] for details regarding the computation of the Γ_g value. For reasons of space, in this paper we summarize only the parts that we modified to obtain the new version of the algorithm.

First of all, we describe the segmentation strategy adopted during the first phase of the *divide et impera approach*. Given a generic pitch curve, the algorithm splitted this curve into two

subcurves sharing and an endpoint by choosing the first local maximum, if present, and choosing the midpoint otherwise. This is because, by considering them later during the backtracking phase, tonal peaks were implicitly considered more important than other points, as their removal influenced large portions of the final curve.

During the backtracking step of the *divide et impera* schema, the removal of the point shared by two adjacent subcurves $A = [a_1, \ldots, a_n]$ and $B = [b_1, \ldots, b_m]$, with $a_n = b_1$ was evaluated. The two possible mergings of the two subcurves were either the curve where the shared point was kept $S = [a_1, \ldots, a_n, b_2, \ldots, b_m]$ or the curve where it was not $\bar{S} = [a_1, \ldots, a_{n-1}, b_2, \ldots, b_m]$. The overall quality function $F(S)$, computed as the balance between perceptual equality and cost, was compared with $F(\bar{S})$. If $F(\bar{S})$ had a higher value than $F(S)$, the $\bar{S}$ curve was passed to the next backtracking step. For the sake of simplicity, in the following formulas we assume that the two curves being compared have the same number of points. Of course, when the removed point s_i is considered in the $\bar{S}$ curve, $\bar{s}_i$ corresponds to the value obtained by linearly interpolating s_{i-1} and s_{i+1} in t_{s_i}.

To compute the quality of the stylized curves, the algorithm considered the accumulated difference between the original curve and the proposed one in terms of glissando likelihood for each of the segments in the curve, evaluating the risk of introducing a glissando where a static tone was found and vice versa. The distance D between a generic segment $[s_i, s_{i+1}]$ and its stylized counterpart $[\bar{s}_i, \bar{s}_{i+1}]$ was computed as:

$$D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}]) = \\ D_{acc}([s_i, s_{i+1}]) + \Gamma_g([s_i, s_{i+1}]) - \Gamma_g([\bar{s}_i, \bar{s}_{i+1}]) \quad (1)$$

The quality of the $\bar{S}$ curve with respect to the S curve was computed as the weighted mean glissando likelihood accumulated distance over the segments of the $\bar{S}$ curve. The weights were the time portions of the complete curve represented by each segment, thus obtaining:

$$q_g(S, \bar{S}) = \\ \sum_{i=1}^{n-1} \left((1 - |D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])|) \frac{t_{s_{i+1}} - t_{s_i}}{t_{s_n} - t_{s_1}} \right) \quad (2)$$

The model presented in [10] also takes into account a $q_d(S, \bar{S})$ measure related to differential glissando perception checking that a glissando, if present, is kept as it was. In this paper we concentrate on $q_g(S, \bar{S})$ as the changes we made to this function are simply replicated in $q_d(S, \bar{S})$.

Concerning the cost measure, in [9] it was computed as the ratio between the number of points used by the stylized curve and the number of points found in the original curve transformed with a sigmoid function. After introducing the dynamic tonal perception model, in [10] we reported a problem causing the quality measure to rapidly dominate the cost measure in long, continuous pitch curves and causing the insertion of more target points than necessary. In [10] we counterbalanced this effect by introducing an empirically determined α modifier to augment the weight of the cost factor depending on the length of the curve in the cost function. The cost function used by the OpS algorithm is:

$$c(S, \bar{S}) = 1 - \left(\frac{1}{1 + exp(\frac{-(x^\alpha - 0.5)}{0.1})} \right) \quad (3)$$

where x is the ratio between the number of points used by the stylized curve and the number of points used by the original one.

3. Observed problems

The results of the perceptual tests reported in [9, 10], in which naive listeners were recruited, indicated that the stylization proposals of the OpS algorithm performed, in terms of quality, in a similar way with respect to other approaches. The OpS algorithm had the advantage of being parameter independent and it was able to use less points by explicitly taking into account a cost measure during computation. In [13], we included the OpS algorithm in the Prosomarker tool: an instrument designed to give a perceptual account of the pitch curves and to describe the synchronization of the pitch targets with automatically detected segmental events (syllable boundaries and nuclei). While using this tool to describe simple intonation phenomena, we were able to trace a number of recurring situations in which the OpS algorithm was not able to capture specific classes of details from the curve that appeared to be critical to an expert linguist's ear.

1) In [10] we found that giving priority to local minima if no local maxima can be found in the curve during the splitting phase did not seem to introduce improvements. Not having this rule introduced the possibility that a local minimum was evaluated very early during backtracking. As we have seen, this implicitly assigns less importance to the point because the impact of its removal is evaluated on a limited portion of the curve. This was systematically noticed by the human experts evaluating the quality of the OpS curves while testing Prosomarker, as they were able to detect small discrepancies both in timing and in tonal level of lowering targets in the resynthesis with respect to the original utterance.

2) The quality measure dominating the cost one in long pitch segments was not completely addressed by the introduction of the α parameter. Continuous pitch segments, longer than the ones we tuned α on, were found in other corpora: in these segments the effect was strong enough to make the α weighting useless. The presence of the α parameter is also less motivated from a theoretical point of view than the rest of the model, thus making the framework less reliable than we intended.

3) When local maxima split the curve in two subcurves that are very unbalanced in length, the algorithm was unable to adequately protect the smaller part of the curve. The quality of the longer subcurve was considered more important than the quality of the shorter subcurve that, subsequently, was often overstylized. This was caused by the weighting of each segment dependently of the fraction of time it stylized.

4. The SOpS algorithm

We now present the updates to the OpS algorithm we introduced in order to address the problems we highlighted in the previous section. The final model we obtain is simplified with respect to the preceding version. For this reason, we will refer to the updated version of the OpS algorithm as the Simplified Optimal Stylization (SOpS) algorithm.

To address problem 1, we reintroduced the splitting rule giving priority to local minima if no local maxima can be found. By evaluating these points later during the backtracking phase, the SOpS algorithm is able to protect low targets better than the OpS algorithm. Problems 2 and 3 were both related to the measure we used to evaluate shared endpoints removal during backtracking. Specifically, having the whole subcurves influ-

ence the quality measure introduced the problems related to differences in the curves' length. However, the removal of the shared endpoint, while generically influencing the quality of the two curves' mergings, is more specifically related to the quality of the two neighboring *segments*. Back to the preceding example, given the A and B curves, the removal of the shared point $a_n = b_1$ only influences the quality of the $[a_{n-1}, a_n]$ and $[b_1, b_2]$ segments. Therefore, having the quality evaluation of the curves $[a_1, a_{n-1}]$ and $[b_2, b_m]$ contributing to the evaluation introduces an identical factor on both sides of the comparison operator. Eliminating this factor makes the algorithm take into account only the neighboring segments quality. By weighting equally these two segments, we also remove the effect of longer movements being considered more important than shorter ones. Equation 2 is reformulated as

$$q_g(S, \bar{S}) = \{2 - |D([s_{i-1}, s_i], [\bar{s}_{i-1}, \bar{s}_i])| - |D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}])|\}/2 \quad (4)$$

Also, by considering local minima earlier in the splitting phase of the *divide et impera* schema, the *midpoint split* rule is applied to segments that are either quasi-linear or parabolic. In the first case, small differences are introduced by removing points while, in the second case, the *midpoint split* rule rapidly produces quasi-linear segments. This way, early evaluated points are more concerned with small details mainly depending on energy and pitch interactions, while lately evaluated points are more related with the description of larger prosodic events. Because of this distinction, it is not necessary to retain the fine details produced by the early backtracking steps up to the points controlling medium/long range pitch movements. Since the changes introduced by removing these points become very evident by delaying their evaluation to the latest steps of the backtracking process, the influence of the fine details on the decision process is not relevant. We therefore modified Equation 1 so that it does not keep track anymore of the preceding stylization steps obtaining the new formulation

$$D([s_i, s_{i+1}], [\bar{s}_i, \bar{s}_{i+1}]) = \Gamma_g([s_i, s_{i+1}]) - \Gamma_g([\bar{s}_i, \bar{s}_{i+1}]) \quad (5)$$

Concerning the cost measure, in [9] we used the sigmoid transformation of the ratio between the number of points used by the stylization and the number of points used by the original curve so that [9, p. 1994] values of the cost measure at one end of the scale would not have been very different. As the impact revealed itself to be negative with respect to the evaluation of the quality/cost balance, we now consider the untransformed ratio represented as x in Equation 3 as cost measure.

5. Test material

For the presented evaluations we employed the 382 files of the prominence annotated TIMIT subset used in [14] to test automatic methods for prominence detection. This dataset was chosen for the curves cost evaluation we presented in [9] because we needed a large set of prominence annotated speech samples to evaluate the impact of using prominence information in a pitch stylization task. The same dataset was used for the cost related tests in [10]. In this work, we use the same dataset for qualitative, other than cost, evaluation because our goal is to check that the new approach does not introduce detectable changes on a large scale as we are interested in recovering only the details the OpS algorithm was missing. The dataset consists of 382 files containing 20 minutes of read speech extracted from the TIMIT corpus.

6. Results

From the quantitative point of view, we considered the number of points used by the SOpS algorithm with respect to OpS. The SOpS algorithm, on the considered dataset, uses 3.46 points per second (Pps) while the OpS algorithm uses 3.59 Pps. Table 1 shows a summary of the cost test between OpS, SOpS and an older version of the OpS algorithm employing manual prominence annotation called OpSProm [9].

Table 1: Cost test results.

	OpS	SOpS	OpSProm
Points per second	3.59	3.46	3.47
Total points	4118	4007	4009

A paired t-test indicated that the difference in Pps between OpS and SOpS is not statistically significant ($p > 0.01$). However, close inspection of the pitch curves where the OpS algorithm introduced more points than necessary showed that the SOpS algorithm does not suffer from this problem. The amount of reduction observed (0.13 Pps) and the actual p-value (0.012) are coherent with the goal we had of reducing the number of points used only in specific areas. The performance of the SOpS algorithm in terms of Pps is much more similar to the one we obtained with the OpSProm algorithm. A paired t-test between the Pps measures obtained by SOpS and OpSProm confirms this ($p > 0.9$) with greater certainty with respect to the result we presented in [10], where we stated (p. 205) that the difference between OpS and OpSProm, while not significant ($p > 0.01$), was to be taken carefully as the actual p-value was 0.0142.

From the qualitative point of view, a Wilcoxon test on the differences between curves generated by the two algorithms showed that the location shift is not significant ($p > 0.4$). The size of the considered dataset makes it safe to assume that no significant differences can be found between the curves proposed by the two algorithms on a large scale. This result confirms that the modifications introduced by the SOpS algorithm do not alter the stylized curve up to a statistically detectable degree. Close inspection of the cases on which the new model is intended to perform better, however, show that the details the OpS algorithm was not able to retain are correctly modeled by the SOpS algorithm.

7. Case study

In Figure 1, we show the detail of a pitch contour, the stylization proposed by the OpS algorithm (dashed line) and and the alternative proposed by the SOpS algorithm (dotted line) along with the energy profile. While the two algorithms perform identically on the first movement, the final rise/fall sequence is described differently. Since the curve's portion after the peak is much shorter than the rest of the curve, protecting the final lowering movement was considered not valuable enough by the OpS algorithm. This decision is encouraged by the tonal perception model as the rising movement preceding the final fall is synchronized with a rising energy profile, thus lowering the modeled glissando perception capability. The influence of sections that do not depend by the point being evaluated also plays

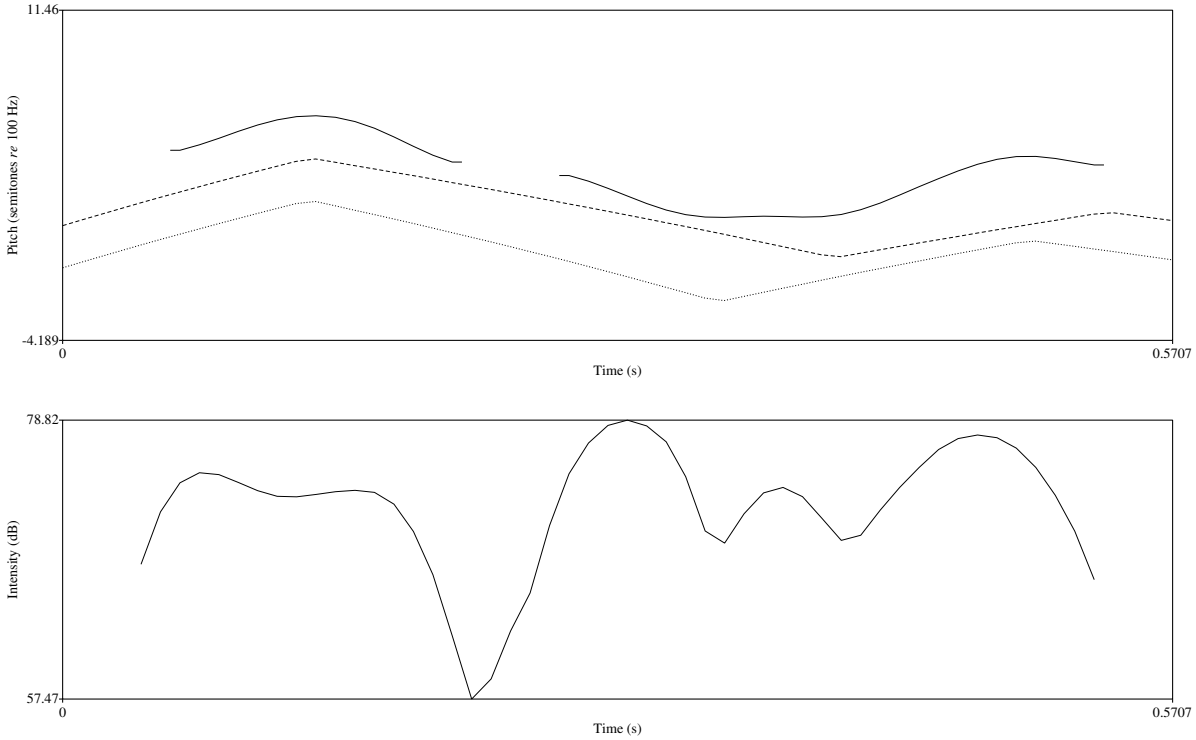


Figure 1: A pitch contour (solid line) along with the OpS stylization (dashed line) and the SOpS one (dotted line). The stylized curves are shifted by 2 Semitones each with respect to the original one for visualization purposes. Along with the pitch curve, the energy profile of the considered speech fragment is shown.

a role, as discussed before. The SOpS algorithm, by considering only the neighboring subcurves and by weighting them equally, is able to protect the final movement when evaluating the peak point, as expected because of the synchronized falling energy contour. The turning point before the rise is shifted 60ms earlier because of the segmentation strategy giving more importance to local minima. This improves the representation of the subcurve synchronized with the falling energy movement. The following pitch rise, synchronized with a rising energy contour, is more stylized than before, so no points are added. From perceptual inspection, this choice appears to improve the overall quality of the curve used in the example. The audio files of the original utterance from which the provided example is extracted are attached to this paper together with the resynthesis obtained with the OpS and SOpS curves. The magnitude of the changes the SOpS algorithm introduces with respect to the OpS curves are, in general, similar to the ones shown in the example. This explains why the similarity test based on statistical closeness is not able to detect a significant difference between the two algorithms. Being these changes important for an expert listener, however, we are in line with out observation that statistical closeness measures are not good estimators of a stylized curve's quality [9].

8. Conclusions and future work

We have presented the SOpS algorithm, an evolution of the OpS algorithm achieving better precision in stylizing specific details of the pitch curve that are important for an expert's ear. The SOpS algorithm is based on a simplified version of the dynamic tonal perception model used by the OpS algorithm. While the curves produced by the SOpS algorithms do not differ in a statistically relevant way from the original OpS curves in a qualitative and quantitative sense, we have shown that close inspection of the details we were interested in recovering are correctly represented by the SOpS stylization, thus obtaining a better representation of the *perceived* intonational profile that can be used in prosodic research. Concerning the number of control points used, we have shown that the new algorithm obtains results more similar to the ones we reported by using manual prominence annotations without altering the produced curves in a significant way. Therefore, the perceptual model behind the SOpS algorithm produces representations of the pitch curve that are both more precise and essential than the ones produced by its predecessor on fine intonational details. The simplified version of the dynamic tonal perception model will make it easier, in the future, to introduce the full range of changes indicated by the SCH, as we are currently considering energy only. The SOpS algorithm is implemented as a Python module of the Prosomarker tool, which is freely available for research purposes.

9. Acknowledgements

This work was supported by the European Community, within the FP7 SHERPA IP #600958 project. We would like to thank all the people who used the Prosomarker tool and gave us the necessary feedback to improve the OpS algorithm.

10. References

[1] J. t'Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: An Experimental-Phonetic Approach*. Cambridge: Cambridge University Press, 1990.

[2] D. Hirst and R. Espesser, "Automatic modelling of fundamental frequency using a quadratic spline function." *Travaux de l'Institut de Phonetique d'Aix-en-Provence*, vol. 15, pp. 75–85, 1993.

[3] C. D'Alessandro and P. Mertens, "Automatic pitch contour stylization using a model of tonal perception," *Computer Speech and Language*, vol. 9, no. 3, pp. 257–288, 1995.

[4] P. Mertens, "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model," in *Proc. of Speech Prosody*, 2004.

[5] M. Wypych, "Automatic pitch stylization enhanced by top-down processing," in *Proc. of Speech Prosody [Online]*, 2006.

[6] S. Ravuri and D. Ellis, "Stylization of pitch with syllable-based linear segments," in *Proc. of ICASSP*, 2008, pp. 3985–3988.

[7] P. K. Ghosh and S. Narayanan, "Pitch Contour Stylization Using an Optimal Piecewise Polynomial Approximation," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 810–813, 2009.

[8] D. Wang and S. Narayanan, "Piecewise linear stylization of pitch via wavelet analysis," in *Proc. of the European Conference on Speech Communication and Technology*, 2005, pp. 1–4.

[9] A. Origlia, G. Abete, C. Cutugno, I. Alfano, R. Savy, and B. Ludusan, "A divide et impera algorithm for optimal pitch stylization," in *Proc. of Interspeech*, 2011, pp. 1993–1996.

[10] A. Origlia, G. Abete, and F. Cutugno, "A dynamic tonal perception model for optimal pitch stylization," *Computer Speech and Language*, vol. 27, pp. 190–208, 2013.

[11] M. Rossi, "Interactions of intensity glides and frequency glissandos," *Language and Speech*, vol. 21, pp. 384–394, 1972.

[12] D. House, "Differential perception of tonal contours through the syllable," in *Proc. of ICSLP*, 1996, pp. 2048–2051.

[13] A. Origlia and I. Alfano, "Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification," in *Proc. of LREC-2012*, 2012, pp. 997–1002.

[14] F. Tamburini and P. Wagner, "On automatic prominence detection for german," in *Proc. of Interspeech*, 2007, pp. 1809–1812.

Interpersonal factors affecting tones of question-type utterances in Japanese

Hiroaki Hatano [1], *Carlos T. Ishi* [1], *Miyako Kiso* [1]

[1] ATR Intelligent Robotics and Communication Labs.

{hatano.hiroaki, carlos, miyakokiso}@atr.jp

Abstract

The purpose of this paper is to clarify the interpersonal factors affecting phrase final tones of question-type utterances in daily conversations. We extracted question-type utterances ending with final particles from our Japanese dialogue speech database and classified them into two categories according to the degree of information request. Prosodic features were then analyzed by focusing on phrase final F0 movement and pitch reset. Analysis results indicated that F0 rising and falling degrees increase when the speaker expresses an attitude of intimacy to the dialogue partner, such as in conversations among family members and infant-directed speech. In addition, the presence of pitch reset in the phrase final was found to have functions of relieving the speaker's tension, when the dialogue partners have distant relationship.

Index Terms: question, intonation, interpersonal relationship

1. Introduction

In daily conversations, question is one of the most commonly-used speech acts. In general, a question is premised on the presence of an interlocutor. Therefore, it is a speech act peculiar of dialogues, being a key to understand the details in dialogue communication.

In general, it is important for raising the intonation of question-type utterances in Japanese. For example, it is stated that if the pitch of phrase final is lowered, it sounds like a cross-examination [1]. Therefore, an incorrect use of the phrase final intonation may cause misunderstandings in communication. Further, Japanese learners should acquire the phrase final intonation of questions [1].

There are many researches on prosodic analysis of questions in natural conversations. For example, in [2] the prosody of questions has been investigated in natural Swedish conversations, revealing that prosody differs according to the different types of questions. It was reported that yes-no questions generally have a falling intonation, whereas wh questions generally have a rising intonation in Swedish. In [3], terminal intonations of questions extracted from conversations between a doctor and a patient in Dutch have been examined. It was found that terminal F0 rises have higher values in the order of wh questions < yes/no-questions < declarative questions, in male speech. We also have investigated questioning prosody extracted from daily conversations in Japanese [4][5][6]. In our previous studies, we pointed out that the occurrence rates of non-rising tones increase in question types where the speaker assumes that the interlocutor does not necessarily hold the answer information.

These researches are mainly focused on the types of questions (e.g. yes-no questions, wh questions and so on). Although these approaches are effective to reveal prosodic variability of questions, one also should concern with another factor, namely interpersonal relationship between the dialogue partners. It is well-known that infant-directed speech (**IDS**) has some characteristic prosody relative to adult-directed speech (**ADS**) [7]. In [7], the prosodic modifications (such as higher mean-f0, f0-minimum and f0-maximum, greater f0-variability, shorter utterances and longer pauses) of IDS have been analyzed. In [8], it has been revealed that the voice quality parameter "NAQ" varies according to dialogue partner. As seen above, it is clear that the interpersonal relationships affect the speech prosody in various ways.

We have reported that phrase final tones of questions are varied according to the hierarchical relations or familiarity [4][5][6]. For example, if the interlocutor has higher status relative to speakers (e.g. senior staff), F0 of phrase final syllables are significantly lower than that of lower status (e.g. junior staff) or equal footings (e.g. colleague). These previous studies, however, were not conditioned about question types and phrase final part-of-speech information.

In this study, we aim to clarify the interpersonal factors which affect the prosodic modification of question-type utterances. Our previous research had made the result with combined all phrase finals, such as final particles, auxiliary verbs and without suffixes. Under that condition, it was unclear whether the difference in intonation was caused by morphological or interpersonal factors. Thus, in the present work, we constraint our analyses to questions ending with final particles.

2. Materials and methods

2.1. Speech data

We used a Japanese conversational speech database recorded in ATR/IRC labs, which was also employed in our previous studies [4][5][6]. The database has 69 dialogue sessions including 31 speakers (12 adult males, 15 adult females, 2 young child males and 2 young child females). "Young child" means pre-elementary school (age under 6) in this study. Each dialogue session has 10 to 15 minutes, face-to-face communication and no specific tasks (topic-free conversations). The total duration of speech data is about 900 minutes. Different types of interpersonal relationships between the dialogue partners are included, for example, mother/father-son/daughter, superior-subordinate, friends, first meets, and so on. Some speakers participated in multiple sessions talking with different interlocutors.

Simultaneous recordings of speech and EGG (electro-glottograph) signals are available in this database. Sampling rates are 16 kHz/16 bits. Audio data is recorded using directional microphone (Sanken CS-1). In part of the dialogue sessions, headset microphone data is also available.

2.2. Extraction of question-type utterances

The speech utterances in the database are segmented in phrase units based on pauses and clear pitch resets between phrases. 2~4 native Japanese speakers evaluated each utterance from the standpoint of turn-taking function (turn-keeping, turn-yielding, backchannel, and fillers). In the present work, we employed the turn-yielding utterances where 2 or

more annotator's judgments agreed. As a result, we got 4231 utterances from the database.

For each of the utterances, question types were annotated by 3 native speakers (research assistants). We used the same label set of question types used in our previous research [5][6]. The agreement rates (in terms of Kappa coefficients) among the question type labels by each pair of annotators were .77, .75 and .68. We use the utterances where 2 or more annotators agreed for the subsequent analysis.

We then separated these question types according to the degree of information request. The category of question types expressing higher degree of information request includes: *Yes-No questions* (n=1258), *Information request* (451), *Subjective feedback request* (317), *Repetition request* (39) and *Counter-questions* (16). The category of question types expressing lower degree of information request include: *Quiz-type questions* (n=8), *Agreement request* (979), *Open-type questions* (159), *Backchannel-type questions* (377), and *Self-questions* (161). We call the former category as "**HIR**: Higher degree of Information Request" and the latter category as "**LIR**: Lower degree of Information Request" hereafter. By the above procedure, we got 2081 utterance in HIR and 1684 utterance in LIR (total size is 3765).

2.3. Parameterization of phrase final tones

There are many acoustical candidates and categorizing method for phrase final tones. For example, the difference between the average pitch of the first half and the second half of the question is used in [1] for rough estimate of the rising or falling intonation. J_ToBI and X-JToBI are well-known prosodic labeling schemes [11][12]. The latter is particularly adjusted for spontaneous speech descriptions and used in the "*Corpus of Spontaneous Japanese*". In the present study, however, we employ a set of parameters proposed in [9][10] for description of phrase final tones, on the grounds that these parameters are based on human's perception. Five tone categories are used in the present work, namely Rise, Flat, Fall, Reset-Flat and Reset-Fall. Added to this, parameters can be automatically extracted by the procedures described below.

For phrase final duration, an automatic procedure was first realized, by using power and spectral change constraints [9]. And then the errors in the automatic segmentation were manually corrected. The newly segmented boundary intervals are used as segmental duration of the phrase finals.

For the pitch-related parameters, F0 values are first estimated based on a conventional method of picking peaks in the normalized autocorrelation function. F0 was extracted for both speech and EGG signals available in the database. However, the speech signals are used only when the EGG signals are not available in the database. For speech signal, the auto-correlation of the LPC inverse-filtered residue of the pre-emphasized signal was used, while for the EGG signal, the autocorrelation of a high-pass filtered signal with 70Hz cutting frequency (for removing DC and low frequency movements) was used. All estimated F0 values are then converted to a musical (log) scale.

The phrase final is split in two segments of equal length, and representative F0 values are extracted for each segment. Several candidates for the representative F0 values have been tested in [9]. Here, we use the ones that best matched with perceptual scores of the F0 movements. For the first segment, an average value is estimated using F0 values within the segment (*F0avg2a*). And for the second segment, a target value is estimated as the F0 value at the end of the segment of a first order regression line of F0 values within the segment (*F0tgt2b*). A variable called *F0move* is defined as the difference between *F0tgt2b* and *F0avg2a*, quantifying the amount and direction of F0 movement within the syllable.

F0move = F0tgt2b - F0avg2a

In this study, phrase finals are categorized as rise pitch movements when *F0move* > 1 semitone & duration > 100 ms, fall pitch movements when *F0move* < -1.3 semitone & duration > 100 ms, and flat pitch movements otherwise.

A parameter called *F0reset* is another important factor in categorizing the phrase final tones. This parameter indicates the presence or absence (or degree) of pitch reset between the phrase final and the syllable prior to the phrase final. The degree of pitch reset is defined as follows:

F0reset = F0avg2a - F0avg_p

F0avg_p is an average F0 value of the final portion of the syllable preceding the phrase final. *F0avg_p* is estimated from four reliable F0 values obtained by back-tracking and searching from the phrase final start point. A pitch reset is judged to be present when *F0reset* > 1 semitone.

The thresholds above were based on pitch movement perception experiments. From the utterances where agreement was obtained for the question type annotations, the ones where F0 could not be extracted and is unclear were removed from the analysis, resulting in a total of 2,226 utterances. We randomly picked 758 utterances for perception experiments. 3 annotators labeled each utterance according to the 5 tone categories (Rise, Fall, Flat, Reset-Fall and Reset-Flat). The agreement rates (in terms of weighted Kappa coefficients: Rise > Reset-Flat > Flat > Fall > Reset-Fall) among the tone category labels of the 3 annotators were 0.75, 0.61 and 0.57. The automatically classified tones under the above criteria were also checked for agreement with annotator's decision. The agreement rates (in terms of weighted Kappa coefficients) between above criteria and the 3 annotators were 0.68, 0.65 and 0.52.

Extracted phrase final tones according to the degree of information request, HIR and LIR, are shown in Table 1.

Table 1. Distribution of the phrase final tones classified according to degree of information request (HIR/LIR: Higher/Lower degree of Information Request). The numbers indicates the occurrences.

Tone	*HIR*	*LIR*	*Sum*
Rise	749 (62.8%)	319 (30.9%)	1068 (48.0 %)
Reset-Flat	88 (7.4%)	136 (13.2%)	224 (10.1 %)
Flat	25 (21.0%)	257 (24.9 %)	508 (22.8 %)
Fall	76 (6.4%)	152 (14.7 %)	228 (10.2 %)
Reset-Fall	29 (2.4%)	159 (16.4 %)	198 (8.9 %)
Sum	1193	1033	2226

2.4. Classification of phrase final morphemes

Linguistic information about the part-of-speech and morphemes appearing at phrase finals was taken into account when verifying the influence of tones. For example, the occurrence rate of rising tones at the last syllable of questions

was about 30 % in phrases ending with final particles, while it was over 60% in phrases not ending with final particles (e.g. nouns) [5].

For classification of phrase final morpheme and part-of-speech, we used the free part-of-speech and morphological analyzer software "MeCab" [13] (the dictionary used was UniDic [14]). Because of the spoken style of the utterances in the database, we checked and corrected the output from MeCab which mainly attempt to analyze written language. In this study, we selected the phrases ending with final particles.

Table 2 shows the distributions of the phrase final particles (top 3) according to the degree of information request, HIR and LIR. The total numbers of final particles were 391 in HIR and 660 in LIR.

Table 2. Distributions of the phrase final particles (top 3) according to the degree of information request (HIR and LIR). The number indicates the occurrences.

HIR		*LIR*	
/no (N)/	235 (60.1%)	/yone/	117 (17.7 %)
/ka/	115 (29.4%)	/ne/	110 (16.7 %)
/Qke/	14 (3.6%)	/na/	98 (14.8 %)

2.5. Criteria for analysis of interpersonal relationship

The aim of this section is to clarify if phrase final prosody could change according to the distance in the interpersonal relationship between the dialogue partners. For the subsequent analysis, we select the speakers according to the interpersonal relationship under the following criteria.

1. Speakers who have dialogue sessions with both young child and adult speakers.
2. Speakers who have dialogue sessions with his/her family member and others (only adult speakers).
3. Speakers who have dialogue sessions with acquaintances and with someone else for the first meeting (only adult speakers).

In case 1, we attempt to verify how *F0move* of phrase finals vary at questioning utterances of IDS and ADS in HIR and LIR. Under this condition, we got 5 speakers in HIR and 4 speakers in LIR. In cases 2 and 3, we attempt to verify how *F0move* or *F0reset* vary according to the degree of closeness (intimacy) in their interpersonal relationships. The purpose of case 2 is to verify whether the prosodic features of IDS appear in similar situations or not. Generally speaking, talking with a young child is a stress-free situation and the speaker's attitude tends to be friendly. This situation is thought to be similar to when talking with family members. Under this condition, we got 4 speakers in HIR and 5 speakers in LIR. The purpose of case 3 is to consider whether or not the speaker changes the prosody of questions when talking with a completely unknown person (for the first meeting). Under this condition, we got 7 speakers in HIR and LIR.

3. Analysis Results and Discussions

3.1. *F0move*: for young children or adults

Fig. 1 shows the average *F0move* in semitones when the interlocutor is young child or adult, in HIR and LIR conditions.

As **Fig. 1** indicates, when the interlocutor is a young child, *F0move* in the final particle is significantly higher than that of adult at HIR (Welch Two Sample t-test; t (48.1) = -2.8, p < .01). On the other hand, *F0move* for young child is significantly lower than that for adult at LIR (t (58.1) = 2.1, p < .05). No significant differences were found in *F0reset* in both conditions (HIR: *F0reset* for young child is -0.46, for adult is 0.98, t (37.9) = -0.4, *ns*; LIR: for young child is 1.97, for adult is 2.02, t (62.6) = 0.1, *ns*).

These results mean that the typical prosody to young child at HIR questions is exaggerated in rising tones. At LIR, on the other hand, we can notice that the falling tones are exaggerated for young child. The typical question prosody of LIR ending with final particles is a reset-flat (F0move > -1.3) when the interlocutor is an adult, while it becomes a reset-fall (F0move < -1.3) when it is a young child.

These results agree with a past study on Infant-directed speech (IDS), where it was mentioned that IDS has greater F0 variability than adult-directed speech (ADS) [7].

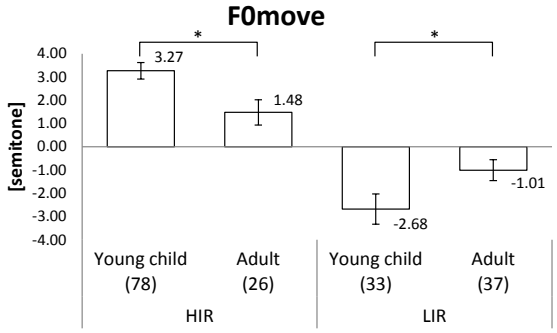


Fig. 1 Average *F0move* for young child/adult at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

3.2. *F0move*: for family members or others

Fig. 2 shows the average *F0move* in semitones when the interlocutor is family member or others, in HIR and LIR conditions.

As can be seen in **Fig. 2**, when the interlocutor is a family member, *F0move* of the final particle is significantly higher than that in others, at HIR (t (66.8) = 2.7, p < .05). On the other hand, *F0move* for family member is significantly lower than that for others at LIR (t (120.4) = 2.0, p < .05). *F0reset* is significantly different at LIR but not at HIR (HIR: *F0reset* for family member is -0.73, for others is 1.64, t (57.4) = -1.6, *ns*; LIR: for family member is 3.70, for others is 1.59, t (113.0) = 2.0, p < .05).

Fig. 2 indicates a similar tendency to **Fig. 1**. F0move is exaggerated for family members at HIR and LIR conditions. It is interesting to note that tone categories at both conditions for others are flat (i.e. -1.3 < *F0move* < 1). This result suggests that questions for others have the tendency to become a reset-flat tone. In other words, flattened F0 movements (i.e., repressed rising and falling degrees) in final particles appear in a relatively formal dialogue situation. In our previous study about questioning tones [4], it has been reported that if the interlocutor has higher status relative to the speaker (e.g. senior staff), *F0move* of phrase final syllables were significantly lower than that of lower status (e.g. junior staff)

or equal footings (e.g. colleague). And in [6], we showed that the occurrence rate of rise tone of question is lower when the speaker feels concern for the interlocutor (e.g. elderly) than that of "no concern" situation. These previous studies, however, were not conditioned on the part-of-speech of phrase finals. Generally speaking, speakers are required a formal speech style when talking with someone who has higher status or when feeling concern for the interlocutor. The results of the present work, where the part-of-speech of phrase finals is constrained to final particles, also support the results above.

It has been pointed out in [15] that a rising intonation at phrase finals in Japanese indicates a relatively heavy attitude of speakers. In other words, rising tone has a strong attitude of requesting an answer to the interlocutor. From this point of view, it seems reasonable to suppose that it is easier to speakers exposing their attitudes for family members. Consequently, the degree of *F0move* at HIR for family members becomes higher than that for others. In addition to the features of rising tones, the degree of falling can also be interpreted as showing strong attitudes at LIR condition (i.e. requesting agreement).

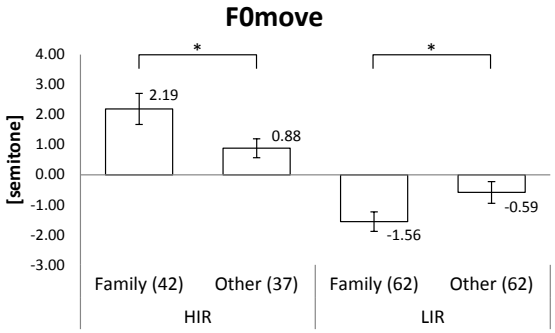


Fig. 2 Average *F0move* for family-member/others at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

3.3. *F0reset*: for acquaintances or first meets

Fig. 3 shows the average *F0reset* when the interlocutor is an acquaintance or a person of first meeting, in HIR and LIR conditions.

As **Fig. 3** indicates, when the interlocutor is a person of first meeting, *F0reset* is over 1 semitone (i.e. pitch reset is present) at both conditions (HIR and LIR). On the other hand, *F0reset* is lower than 1 semitone for acquaintance at both conditions. These differences are significant (HIR: t (91.5) = 2.1, $p < .05$; LIR: t (212.2) = 2.0, $p < .05$). In contrast, *F0move* didn't show significant differences in both conditions (HIR: *F0move* for acquaintance is 0.77, person of first meeting is 0.70, t (82.4) = -0.1, *ns*; LIR: for acquaintance it is -0.62, and for person of first meeting it is -1.23, t (213.1) = -1.7, *ns*).

As we mentioned in the previous section, speakers tend to be mindful of the questioning manner, according to the interpersonal relationship (to raise or drop in phrase final tone at HIR or LIR). Although significant differences did not appear, values of *F0move* are in the range of flat tones at HIR and LIR conditions (i.e. -1.3 < *F0move* < 1). It is thought that the effect of repressing the degree of rising or falling tones in relatively formal situations is reflected in this result.

In addition to this observation, **Fig. 3** showed another aspect of tone functions. Speakers clearly use pitch reset in questions at HIR and LIR conditions for completely unknown people (*F0reset* > 1). It has been described in [15] that pitch reset ("ukiagari-tyo" in their term) at phrase finals means speaker's light attitude. Although the participants of recordings can stop the dialogue at any time, they usually try to be friendly by continuing the conversation in smooth manner. Considering this situation and the description in [14], pitch reset at phrase final particles in question utterances can be interpreted as having functions of relieving the tension or showing a friendly attitude of the speaker.

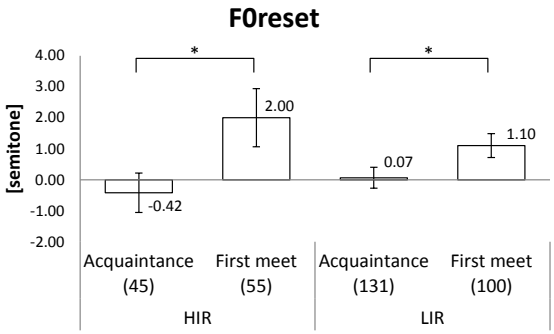


Fig. 3 Average *F0reset* for acquaintance/first meet at HIR and LIR questions. The numbers in parenthesis indicate the numbers of occurrences.

4. Conclusions

We conducted quantitative analysis on free conversational dialogue database, for clarifying the functions of phrase final tones in questions, from an interpersonal relationship viewpoint. Question-type utterances were classified into two categories on the basis of degree of information request (HIR and LIR). The analysis results indicated that:

1. The degree of F0 rising at HIR questions and the degree of F0 falling at LIR questions are exaggerated when talking with a young child, in comparison to when talking with an adult.
2. Similarly, the degree of F0 rising at HIR questions and the degree of F0 falling at LIR questions are exaggerated when talking with a family member, in comparison to when talking to others.
3. The degree of F0 reset at the phrase final syllable is higher when talking with a person of first meeting at both HIR and LIR questions, in comparison to when talking with an acquaintance.

It is inferred from these results that an increase of raising degree at HIR question and lowering degree at LIR question can be regarded as an attitude of intimacy to the interlocutor. In addition, the presence of pitch reset at the final particles can be interpreted as having functions of relieving the speaker's tension.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 23680019, 25884099. We also thank all research assistants of the ATR group who contributed with the annotations.

6. References

[1] Ayusawa, T., "Acquisition of Japanese accent and intonation by foreign learners", *Journal of the Phonetic Society of Japan*, Vol.7 No.2, 47-58, 2003 (in Japanese).

[2] Strömbergsson, S., Edlund, Jens., and Hous, David., "Prosodic measurements and question types in the Spontal corpus of Swedish dialogues", Proceedings of *INTERSPEECH* 2012., 2012.

[3] Heuven, V. J., van, Hann, J. and Kirsner, R. S., "Phonetic correlates of sentence type in Dutch: Statement, question and command", Proc. ESCA International Workshop on Dialogue and prosody, 35-40, 1999.

[4] Hatano, H., Arai, J. and Ishi, C. T., "Analysis of factors which contribute to choice of questioning prosody in natural conversation", Proceedings of *The Spring Meeting of the Acoustical Society of Japan.*, 429-430, 2013 (in Japanese).

[5] Hatano, H., Kiso, M. and Ishi, C. T., "On the factors which contribute to decision of phrase final intonation of questioning utterance in natural conversation", Proceedings of *The Twenty-Seventh General Meeting of the Phonetic Society of Japan.*, 59-64, 2013 (in Japanese).

[6] Hatano, H., Kiso, M. and Ishi, C. T., "Analysis of factors involved in the choice of rising or non-rising intonation in question utterances appearing in conversational speech", Proceedings of *INTERSPEECH 2013*., 2013.

[7] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies. and Fukui, I., "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", J. Child. Lang., 16:477-501, 1989.

[8] Campbell, N. and Mokhtari, P., "Voice Quality: the 4th prosodic dimension", Proceedings of the *ICPhS'03*, 2417-2420, 2003.

[9] Ishi, C. T., "Perceptually-related F0 parameters for automatic classification of phrase final tones", *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No.3, 481-488, 2005.

[10] Ishi, C. T., "The function of phrase final tones in Japanese: Focus on turn-taking", *Journal of the Phonetics Society of Japan.*, Vol.10 No.3, 18-28, 2006.

[11] Venditti, J., "The J_ToBI model of Japanese intonation", in Jun, S. A [Ed] *Prosodic typology: The phonology of intonation and phrasing*, 172-200, New York: Oxford University Press, 2005.

[12] Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J., "X-JToBI: An extended J_ToBI for spontaneous speech", Proceedings of *ICSLP2002*, 1545-1548, 2002.

[13] Downloadable: http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html, accessed on 12 Dec 2013.

[14] Downloadable: http://sourceforge.jp/projects/unidic/, accessed on 12 Dec 2013.

[15] Kawakami, S., "Bunmatsu nado no joosyootyoo ni tsuite", Kokugo-kenkyuu, 16, 2-46, 1963 (in Japanese).

Prosodic correlates of perceived quality and fluency in simultaneous interpreting

George Christodoulides[1], *Cédric Lenglet*[2]

[1] Centre Valibel, Institute for Language & Communication, University of Louvain, Belgium
[2] Department for Specialized Translation and Terminology, FTI-EII, University of Mons, Belgium
george@mycontent.gr, cedric.lenglet@umons.ac.be

Abstract

This study explores the relationship between prosodic features specific to simultaneous interpreting and the listeners' perception of the fluency and accuracy of interpreting, as well as their comprehension of the source speech. Two groups of participants (47 subject experts and 40 non-experts) listened to a 20-minute lecture in German, along with its interpretation into French under two conditions (the actual interpretation, or a read-aloud rendition of the same text by the same interpreter) and answered comprehension and rating questions. The prosodic features of the two conditions were analysed, and differences regarding the temporal organisation of speech, disfluencies, pitch register and the interface between prosody and syntax emerged. Our results suggest that interpreting-specific prosodic features affect the perception of fluency, which in turn affects the perception of accuracy. However the impact on listeners who enjoy relevant contextual knowledge is less pronounced.

Index Terms: speech perception, quality of simultaneous interpreting, fluency

1. Introduction

Simultaneous conference interpreters facilitate multilingual communication in political, technical and other meetings. Typically, they work in soundproof booths, translating speech in real-time so that the participants can follow the debate in their language, without interruption. They are expected to "communicate the speaker's intended messages as *accurately, faithfully, and completely* as possible (…) and to *be clear and lively in [their] delivery*" [1, original emphasis]. Since prosody conveys important information in human communication (e.g. information status, focus, intent, emotion), the interpreter is expected to decode such information from the prosodic structure of the source language (SL) and encode it in parallel constructions in the target language (TL). This 'translation' of prosodic information, along with the linguistic information, from SL to TL inspired studies on the correspondence, or alignment, of prosodic patterns across languages (e.g. [2] where an algorithm is proposed to find equivalences between clusters of intonation patterns in a parallel bilingual corpus).

Beyond these conscious *choices* made by the interpreter, however, the prosodic features of SI are influenced by two *constraints*: the reformulation process of translation and the high cognitive load induced by the task itself [e.g. 3]. SI strategies such as stalling (i.e. waiting for enough input before producing a translation or committing to a syntactic structure) and anticipation (i.e. predicting part of the speaker's input) [4: 201] affect the temporal structure of interpreters' speech. Speech produced under high cognitive load presents specific characteristics, including a lower articulation rate, longer silent pauses, an increased number of filled pauses, corrections and restarts, as well as alterations in voice quality [5, 6, 7].

Surveys among users of simultaneous interpreting suggest that they consider accuracy (or fidelity) to be a crucial quality criterion, whereas prosodic features, such as intonation or accent, are not deemed paramount [8]. However, many users cannot assess the interpreters' accuracy because of their lack of knowledge of the source language. Instead, the users' perception of SI quality may depend on the prosodic features of the interpreters' speech, including intonation, hesitations and pauses. Moreover, the interpreters' liveliness appears to influence the listeners' understanding of the speech content [9, 10]. In other words, the core objective of SI, namely to "produce the same effect on [the listeners] as the original [speech] does on the speaker's audience" [10: 155], might depend not only on what the interpreters say, but also on *how* they say it.

The link between quality perception and prosody is all the more important in SI, as previous research indicates that SI has a distinctive prosodic profile. It results from the interplay of the choices and constraints outlined above, as a speaking style determined both by the situational context and by individual characteristics [c.f. 11, 12]. A particular prosodic profile for SI has been observed in studies for at least the following language combinations: Hebrew to/from English, with a large number of "low-rise non-final pitch movements" [13: 231]; English to German, with "long pauses [and a] high proportion of final pitch movements that indicate a continuation" [14: 72]; and English to French, with less numerous and longer silent pauses, and a narrower pitch range compared to the source speech [15].

Fluency has been regularly used as a quality criterion in expectation surveys on SI [8], although it is a polysemous concept. In ordinary language, fluency refers to general (often foreign) language proficiency, whereas a more technical definition associates fluency with speech flow and absence of disfluencies such as pauses, hesitations and repetitions [16: 537]. Experimental research has shown that these temporal features do not only influence the perception of the interpreter's fluency, but also that of its intonation [8: 67]. Listeners asked to rate "fluency" seem to blend temporal features with intonation (e.g. pitch variation). Consequently, it seems appropriate to merge these parameters in a perceptual study.

Does the particular prosody of SI have an impact on its perception? To date, most studies on prosody and quality in SI are based on carefully doctored speeches [8, 9]. Consequently, their findings cannot be linked directly to the perception of the *authentic* prosody of SI. Shlesinger [13] conducted a small-scale experiment on the impact of authentic SI prosody on 15 listeners' understanding of speech content, comparing excerpts of speeches produced under two different conditions: read aloud from a script and interpreted simultaneously. In a

listening test, the subjects' scores in the "read-aloud" condition were approximately twice as good. She concluded that SI intonation affected meaning and perception, but she argued that this effect would be counterbalanced in the case of authentic conference participants with the relevant contextual knowledge [13: 234]. Our experiment aims to answer the following research questions: Does simultaneous interpreting from German into French have particular prosodic features? If yes, do these features influence the listeners' objective and subjective understanding of the speech content, and their perception of the interpreter's fluency and accuracy?

2. Method

2.1. Perceptual Experiment

Our goal was to create a situation as close to authentic SI as possible. The original speech is an abridged presentation on investment strategy by a German fund manager; its duration is 20 minutes. German was selected as the source language in order to increase the likelihood that French-speaking listeners rely on the interpreter only. A professional conference interpreter (male, French native speaker, 6 years of experience) interpreted the German presentation into French in a state-of-the-art interpreting booth. The recording of this interpretation was transcribed; punctuation was added at syntactically-complete clause boundaries; discourse markers and interjections were included in the transcription (only filled pauses, e.g. 'euh', were omitted). The same interpreter read the transcript, after rehearsing it, and was recorded in a booth. We thus obtained two different prosodic profiles by the same speaker: under authentic SI conditions; and prepared reading, without the cognitive constraints of SI. The two versions were synchronized to the video of the original presentation using *Praat, Audacity* and *AviDemux*.

The experimental design is a conference simulation adapted from [9]. The subjects watch the video of the German presentation and listen to an interpretation into French. An interpreter pretends to work in a booth at the back of the room. This creates the impression of a live interpretation, whereas actually, the subjects are listening to one of the recordings, according to the experimental condition they were assigned to.

The subjects were 87 French-speaking university students: 47 students of economics and 40 translation students. Students in economics were chosen because of their specialized knowledge and their greater availability than professional economists. Translation students were chosen in order to control the influence of prior thematic knowledge. The subjects were matched for academic performance (based on grade records) to control memory and prior knowledge. The resulting pairs were randomly distributed between two experimental conditions: interpreted and read-aloud speech.

We use a listening comprehension test and an assessment questionnaire, which we both pretested extensively. The listening test consists of 3 multiple-choice and 4 half-open questions and assesses the comprehensibility of the speech with a listening score (interval scale). In the assessment questionnaire, the subjects are asked to rate on a 7-point ordinal scale how fluent the interpreter's delivery was (Fluency), how well they think they understood the lecture (Subjective Comprehension) and how accurately they reckon the interpreter rendered the speech (Accuracy).

2.2. Linguistic and Prosodic Analysis

The two recordings (SI and Read) were orthographically transcribed in Praat [17]. We obtained a phonetic transcription as well as an automatic segmentation of words, syllables, phones and pauses, automatically using EasyAlign [18]; the alignment was corrected manually. A 'delivery' tier was added to annotate articulation-related (schwa, creaky voice, liaison and elision) and paralinguistic phenomena (audible breath, noises). Part-of-speech tagging and multi-word unit detection were obtained automatically using DisMo [19] and subsequently verified manually. We applied an annotation scheme for disfluencies based on [20]: i) single-token disfluencies: filled pauses, hesitation-related lengthening, lexical false starts and intra-word pauses; ii) structured disfluencies: repetitions (of one or more words), deletions, substitutions, insertions, and complex combinations of the above.

To process our data we used *Praaline* [21], a toolkit that interfaces with *Praat* and runs a cascade of scripts and/or external analysis tools, each of which may add features to an annotation level (e.g. syllables, words etc.), stored in a relational database. We applied *Prosogram's* [22] two-step algorithm for pitch stylisation: for each syllable, vocalic nuclei are detected based on intensity and voicing, and then the f_0 curve on the nucleus is stylised into a static or dynamic tone, based on a perceptual glissando approach. Syllabic prominence was estimated with *ProsoProm* [23], *Analor* [24], and a manual perceptual annotation was also performed (cf. 2.3). Segmentation into accentual and intonational phrases was performed by an expert annotator (taking into account all prominence scores); furthermore, perceptually-motivated prosodic boundaries were calculated based on the approach proposed in [25]. Several aggregate measures were calculated using *ProsoReport* [12]. In order to study the interface between prosody and syntax, a three-level syntactic annotation was added. First, an annotation into minimal chunks based on the phrasal tag-set of the French Treebank [26] was added manually. We also applied the model for syntactic annotation into functional sequences and dependency clauses detailed in [27] and obtained segmentation into Basic Discourse Units whenever the major prosodic boundaries and the dependency clause boundaries coincide. In total, the two recordings are 42-minutes long (1256 seconds each), and contain 8760 syllables, 1335 silent pauses, and 6143 tokens (words).

2.3. Evaluation of automatic tools

As a corollary study, we evaluate the performance of the above-mentioned automatic tools. Results for prominence detection are shown in Table 2; in line with [28] there was a fair agreement between the human annotator and the tools, and between the tools themselves (consistently lower for the SI).

Table 1. *Evaluation of prominent syllable detection*

Both conditions	**ProsoProm vs. Analor**	**ProsoProm vs. Manual**	**Analor vs. Manual**
Precision	97.1%	81.9%	59.3%
Recall	39.9%	49.1%	86.4%
Correct	77.4%	84.4%	81.6%
F-measure	56.6%	61.4%	70.3%
Cohen's kappa	**0.447**	**0.524**	**0.576**
Interpreting κ	0.394	0.456	0.561
Reading κ	0.478	0.568	0.581

A comparison between the (manually corrected) segmentation into accentual and intonation phrases (AP/IPs), and the perceptually-motivated prosodic boundaries (PBs) proposed by [25] can be seen on Table 2. As expected, the perceptual PBs are coarser than the hierarchical segmentation into AP/IPs (which, for French, is based on the prominence of the last syllable of an each unit).

Table 2. *Comparison of prosodic boundary detection*

Sylls with PBs vs. IP/APs	IP boundary		AP boundary	
	Yes	No	Yes	No
No PB	427	**6773**	1262	**5938**
Minor PB	244	264	381	127
Intermediate PB	55	72	73	54
Major PB	**921**	1	**922**	0

The precision of the POS annotation (DisMo) was 96.3 %, while the syntactic annotation was performed manually.

3. Results

3.1. Global Prosodic Features

A selection of global prosodic features is shown in Table 3.

Table 3. *Global prosodic measures.*

Measure	SI	Read
Articulation ratio (%)	72.6	62.7
Articulation rate (syll/s)	4.91	5.27
Speech rate (syll/s)	3.59	3.33
Speech segments (runs)	493	858
… with average length (syll)	9.2	4.9
Var. coefficient of vowel duration	0.089	0.022
Var. coefficient of syllable duration	0.079	0.042
Median pitch (Hz)	127	152
Pitch range (semitones)	7.8	14.2
Pitch trajectory (semitones/s)	15.13	22.47

We note a higher **articulation rate** (syllables per second *excluding* pauses) under the Reading condition. The interpreter made more silent pauses in the Reading condition (cf. 3.2.); this is reflected in the lower articulation ratio, and the lower speech rate (sylls/s *including* pauses). Speech segments (continuous stretches of speech separated by silent pauses >250ms) are considerably more numerous and shorter under the Reading condition than in SI. These measures indicate that the interpreter **over-segmented** his speech in the Reading condition (short utterances and extensive use of pauses). The observed difference between the variance coefficients of vowel duration and of syllable duration indicate that under the SI condition, the interpreter accelerated and decelerated his articulation more frequently than under the Reading condition. Finally, **pitch range and pitch trajectory** are smaller under the SI condition, compared to Reading; this indicates that the latter was a livelier rendition of the text.

3.2. Silent Pauses and Disfluencies

A Mann-Whitney U test on average **silent pause length** indicates that it is longer under the SI condition ($p < 0.001$). We modelled silent pause length as a mixture of log-normal distributions, following the methodology in [29] and [30]:

$$f(x) = \sum_{i=1}^{N} \pi_i \Lambda_i(\mu_i, \sigma_i^2, x)$$

Three component distributions are identified (using a Bayesian Information Criterion), and their parameters estimated using the Expectation-Maximisation algorithm (Table 4, Figure 1). Cut-off values *t* (local maxima of the model's uncertainty function) are used as thresholds to categorise pauses as 'short', 'medium' or 'long' (instead of *ex-nihilo* fixed thresholds).

Table 4. *Log-normal mixture model of silent pause length (in ms), 1.3 < σ < 1.8, N = 3.*

Pause type	SI			Read		
	π	M	t	π	μ	t
Short	44%	195	283	39%	136	203
Medium	32%	568	1037	48%	581	602
Long	24%	1570		13%	1221	

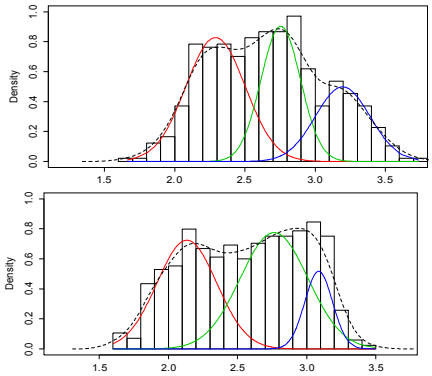


Figure 1. *Density plots of log (silent pause length) and component distributions (SI: top, Reading: bottom).*

Regarding **disfluencies**, under the SI condition the interpreter produced 272 filled pauses vs. only 8 under the Reading condition. Other types of disfluencies were almost inexistent in Reading. Under the SI condition false starts, repetitions and deletions (in order of frequency) were observed. In total, **9.8%** of the tokens were disfluent in SI, compared to **0.4%** in Reading.

3.3. Prosody-Syntax Interface

Basic Discourse Units (BDUs) in [27] are proposed as "the segments that speakers and listeners use to interpret the discourse they are engaged in". Based on the observation that listeners use both prosody and syntax as cues to information structure, BDUs are defined as segments that run between the points where major prosodic boundaries and dependency clause boundaries **coincide**. In a congruent BDU, one intonation unit (IU) contains one dependency unit (DU); in intonation-bound BDUs, one IU contains several DUs; in syntax-bound BDUs, one DU packs several IUs. Regulative BDUs contain only discourse markers or adjuncts. A mixed-boundary BDU contains more than one DU and more than one IU, and is the product of a lack of synchrony between prosodic and syntactic boundaries. Table 5 shows the distribution of BDUs of different types under the two conditions.

Table 5. *Number and average duration of BDUs per type and condition.*

Condition / BDU type	SI		Read	
	%	Avg dur (s)	%	Avg dur (s)
Congruent	21.3	2.57	20.2	2.04
Regulative	21.9	1.24	24.4	0.85
Intonation-bound	5.3	6.43	0.4	1.98
Syntax-bound	30.2	8.20	55.0	5.60
Mixed-boundary	21.3	11.79	0	

We observe that in the SI condition, there was frequently a **mismatch between prosodic and syntactic boundaries**. These are typically cases in which the interpreter constructs a phrase incrementally, pausing inside syntactic units. Figure 2 shows how the three different types of silent pauses are distributed between and within syntactic units (chunks and functional sequences) and BDUs.

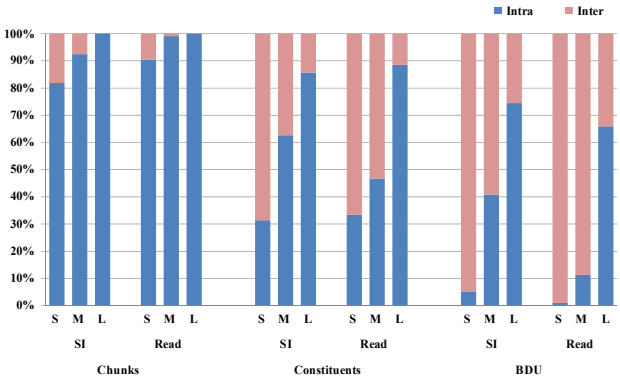


Figure 2. *Silent pauses of different types (S: short, M: medium, L: long), within (intra) and between (inter) syntactic units.*

We observe that in the SI condition, medium-length pauses *within* constituents and *within* BDUs occur more frequently than in the Reading condition (+15% and +30% respectively). This is in line with the high percentage of mixed-boundary BDUs observed. The high proportion of syntax-bound BDUs with a relatively short duration under the Reading condition is another indication of over-segmentation.

3.4. Perception of Quality and Fluency

The questionnaire data were coded and processed with *IBM SPSS Statistics*. Missing observations were excluded. The mean Listening Score, which measures Objective Comprehension (the sum of the correct answers in comprehension questions; max = 17), and the median Accuracy, Fluency and Subjective Comprehension ratings (1 = best), broken down by subject groups and experimental conditions are shown in Tables 6 and 7. Highest scores and best ratings are in boldface.

Table 6. *Mean Listening Score per group (17 = max)*

Group	SI	Read	Both
Translation students (TRAN)	8.47	**9.39**	8.94
Economics students (ECON)	7.95	**8.39**	8.18
Both groups	8.18	**8.83**	8.52

Table 7. *Median quality and subjective comprehension ratings per group (1 = best)*

Group	Rating	SI	Read	Both
TRAN	Accuracy	2	2	2
	Fluency	3	**2**	2
	Subjective Compr.	5	**4.5**	5
ECON	Accuracy	3	3	3
	Fluency	3	3	3
	Subjective Compr.	3	3	3
Both	Accuracy	2	2	2
	Fluency	3	**2**	3
	Subjective Compr.	4	4	4

The translation students' median ratings of fluency and subjective comprehension are better in the Reading condition. Across all groups, there is a moderate correlation between the experimental condition and the fluency ratings, which turns out to be significant in a Spearman correlation test (Translation: $r = -0.506$; $p = 0.001$; Economics: $r = -0.323$; $p = 0.027$; Translation + Economics: $r = -0.393$; $p < 0.001$; two-tailed). In other words, **the subjects who listened to the read-aloud speech tended to rate fluency better**.

Concerning the fluency ratings without regard to the experimental condition, there is a moderate and significant correlation between fluency ratings and subjective comprehension among the students of economics ($r = 0.480$; $p = 0.001$, two-tailed). There is a slightly stronger significant correlation between fluency and accuracy ratings across all groups (Translation: $r = 0.490$; $p = 0.002$; Economics: $r = 0.496$; $p = 0.001$; Translation + Economics: $r = 0.520$; $p < 0.001$; two-tailed).

4. Conclusions and Perspectives

In this paper we presented a perceptual study based on a conference simulation, striving to create a situation as close to authentic conditions as possible.

With respect to our first research question, our findings confirm previous studies regarding the **particular prosodic characteristics of SI**. In the SI condition, the interpreter produced long silent pauses, frequent filled pauses and several reformulation-related disfluencies. The articulation rate was more variable (i.e. more accelerations and decelerations), and the pitch range and trajectory were both narrower in SI, indicating that the same person rendered the text in a more lively fashion, when freed from the cognitive constraints of interpreting. The main effect of SI was observed in the prosody-syntax interface, with often mismatched prosodic and syntactic boundaries, and more intra-unit pauses. The combination of these prosodic features has had an effect on the perceived fluency rating. As previous research has shown that "some disfluencies may be considered felicitous by listeners" when used for communicative purposes [31], it will be interesting to explore the contribution of each prosodic factor in fluency ratings, in a future study.

With respect to our second research question, the results lend additional support to the claim that **the perception of the interpreters' accuracy is linked to that of their fluency**, thus confirming previous experimental findings [e.g. 8, 9]. The differences in listening scores (objective comprehension) between the experimental conditions are less pronounced among students of economics. This seems to support Shlesinger's claim that the prosody of interpreting has less impact on the listeners who enjoy relevant contextual knowledge. One explanation could be that the translation students processed the speech at a more superficial level and hence, were more affected by perturbations of the prosodic structure of the speech. The students of economics could use their prior knowledge to process the speech content at a deeper level and make inferences to compensate for disturbing prosodic variations. Admittedly, the higher mean listening score of translation students is unexpected. We hypothesize that these students benefited from their capacity to capture the gist of speeches in their notes thanks to an elaborate note-taking technique they develop in introductory courses to conference interpreting. In a future study, perceptual and prosodic data could be correlated to test the effect of each prosodic factor on perceived quality and fluency.

5. Acknowledgements

We would like to thank Prof. Liesbeth Degand (University of Louvain) for her help with the syntactic annotation; Dr. Mathieu Avanzi for the annotation of accentual and intonation units; our anonymous interpreter colleague who volunteered for the recordings; our subjects; and the teaching and technical staff at UMons, who made this experiment possible. The second author is supported by a UMons 100% research grant.

6. References

[1] AIIC, "Practical guide for professional conference interpreters", Online: http://aiic.net/page/628/practical-guide-for-professional-conference-interpreters/lang/1#5, 1990/2004.

[2] Agüero, P.D., Adell, J., and Bonafonte, A., "Prosody generation for speech-to-speech translation", Proceedings of the ICASSP, 2006.

[3] Seeber, K. and Kerzel, D., "Cognitive load in simultaneous interpreting: Model meets data", International Journal of Bilingualism 16(2): 228–242, 2012.

[4] Gile, D., Basic Concepts and Models for Interpreter and Translator Training, Revised edition, Amsterdam and Philadelphia, John Benjamins, 2009.

[5] Jameson, A., Kiefer, J., Müller, C., Grossmann-Hutter, B., Wittig, F. and Rummer, R., "Assessment of a user's time pressure and cognitive load on the basis of features of speech", in M. Crocker and J. Siekmann [Eds] Resource-adaptive cognitive processes, 171–204, Berlin, Springer, 2009.

[6] Tet Fei Yap, "Speech production under cognitive load: Effects and classification", Diss., The University of New South Wales, 2012.

[7] Schuller, B., Batliner, A., "Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing", Wiley, 2013.

[8] Collados Aís, Á, Macarena Pradas Macías, E., Stévaux, E. and García Becerra, O. [Eds], La Evaluación de la Calidad en Interpretación Simultánea: Parámetros de Incidencia. Granada, Comares, 2007.

[9] Holub, E. and Rennert, S., "Fluency and intonation as quality indicators", Paper presented at the Second International Conference on Interpreting Quality, Almuñécar, Spain, 2011.

[10] Déjean le Féal, K., "Some thoughts on the evaluation of simultaneous interpretation", in D. and M. Bowen [Eds], Interpreting: Yesterday, Today, and Tomorrow, 154–160, Binghamton, State University of New York Press, 1990.

[11] Léon, P., Précis de phonostylistique, Parole et expressivité, Nathan Université, Paris, 1993.

[12] Goldman, J.-Ph., Auchlin, A. and Simon, A.C., "Description prosodique semi-automatique et discrimination des styles de parole" in H.-Y. Yoo and E. Delais-Roussarie [Eds], Actes d'IDP 2009, 207-221, Paris, September, 2009.

[13] Shlesinger, M., "Intonation in the production and perception of simultaneous interpretation", in S. Lambert and B. Moser-Mercer [Eds], Bridging the Gap: Empirical Research in Simultaneous Interpretation, 225–236, Amsterdam and Philadelphia, John Benjamins, 1994.

[14] Ahrens, B., "Prosodic phenomena in simultaneous interpreting: A corpus-based analysis", Interpreting 7(1): 51–76, 2005.

[15] Christodoulides, G., "Prosodic features of simultaneous interpreting", in P. Mertens and A.C. Simon [Eds], Proceedings of the Prosody – Discourse Interface Conference, 33–37, Leuven, Belgium, 2013.

[16] Chambers, F. "What do we mean by fluency?", System 25(4): 535-544, 1997.

[17] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", online at http://www.praat.org

[18] Goldman, J.-Ph. EasyAlign: an automatic phonetic alignment tool under Praat, Proceedings of InterSpeech, Florence, Italy, 2011.

[19] Christodoulides, G., Avanzi, M. and Goldman, J.-Ph., "DisMo: A morphosyntactic, disfluency and multi-word unit annotator: An evaluation on a corpus of French spontaneous and read speech", Paper submitted to LREC, 2014.

[20] Shriberg, E., "To 'errrr' is human: ecology and acoustics of speech disfluencies", Journal of the International Phonetic Association 31(1): 153–169, 2001.

[21] Christodoulides, G., "Praaline: integrating tools for speech corpus research", Paper submitted to LREC, 2014.

[22] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model" in B. Bel and I. Marlien [Eds], Proceedings of Speech Prosody 2004, Nara, Japan, 23–26 March, 2004.

[23] Goldman, J.-Ph., Avanzi, M., Simon, A.C. and Auchlin, A., "A continuous prominence score based on acoustic features", Proceedings of InterSpeech 2012, 9–13 September, 2012.

[24] Avanzi, M., Lacheret, A., and Victorri, B. "ANALOR. A Tool for Semi-Automatic Annotation of French Prosodic Structure", Proceedings of Speech Prosody 2008, 119–122, 2008.

[25] Mertens, P. and Simon, A.C., "Towards automatic detection of prosodic boundaries in spoken French" in P. Mertens and A.C. Simon [Eds], Proceedings of the Prosody – Discourse Interface Conference, 81–87, Leuven, Belgium, 2013.

[26] Abeillé, A., Clément, L. and Toussenel, F., "Building a treebank for French" in A. Abeillé [Ed] Treebanks: Building and using parsed corpora, 165–188, Kluwer, Dordrecht, 2003.

[27] Degand, L., Simon, A.C., "On identifying basic discourse units in speech: theoretical and empirical issues", Discours (4), Online: http://discours.revues.org/5852, 2009.

[28] Avanzi, M., Rousier-Vercruyssen, L., Schwab, S., Gonzalez, S., and Fossard, M., "C-PROM-Task: A New Annotated Dataset for the Study of French Speech Prosody", Proceedings of TRASP, Aix-en-Provence, 27–30, 2013.

[29] Goldman, J.-Ph., François, T., Roekhaut, S., Simon, A. C., "Étude statistique de la durée pausale dans différents styles de parole", Actes des 28èmes Journées d'Étude sur la Parole (JEP), Association Francophone de la Communication Parlée, Mons, Belgium, 25-28 May, 2010.

[30] Little, D., Oehmen, R., Dunn, J., Hird, K., Kirsner, K., "Fluency Profiling System: An automated system for analyzing the temporal properties of speech", Behavior Research Methods 45 (1): 191–202, 2012.

[31] Moniz, H., Trancoso, I., Mata da Silva, A.I., "Classification of disfluent phenomena as fluent communicative devices in specific prosodic contexts", Proceedings of Interspeech 2009, 1719-1722, ISCA, Brighton, UK, September 2009.

Rhythm analysis in Arabic L2 speech

Ghania Droua-Hamdani[1], Sid-Ahmed Selouani[2], Yousef A. Alotaibi[3]

[1]Speech Processing Laboratory, CRSTDLA, Algiers, Algeria
[2]LARIHS Laboratory, University of Moncton, Shippagan, Canada
[3]College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

gh.droua@post.com, sid-ahmed.selouani@umoncton.ca, yaalotaibi@ksu.edu.sa

Abstract

This paper investigates rhythm speech metrics in Modern Standard Arabic. Recordings of native (L1) and non-native (L2) speakers were obtained from the West Point corpus. The experiment examined the rhythm metric properties of L2 speech using Pairwise Variability Indices and Interval Measures. The application of the Control/Compensation Index to the corpus is also described. Variations in rhythm metrics are detailed by focusing on between-speaker differences such as gender of speakers. L1 and L2 speakers (females and males) were compared through measurement of the duration of short and long vowels.

Index Terms: rhythm metrics, Control/Compensation Index, L2 speakers, gender of speakers, Modern Standard Arabic, English

1. Introduction

Recent studies have developed a battery of metrics to show rhythm differences and similarities between languages. These same metrics have also been used to study second language acquisition, by examining the impact of the first language (L1) on the rhythm of the second language (L2). A large number of studies have investigated the effect of L1 use on L2 production: Korean English [1]; Singapore English [2]; German L2 influenced by Chinese, English, French, Italian and Romanian L1 [3]; Norwegian as L2 [4]; English, Spanish, Dutch and French [5].

Rhythm refers to the temporal organization of speech. The rhythm models developed are based on the acoustic durations of vocalic and consonantal intervals in vocal signals. The most popular algorithms performed are: Interval Measures (IM) and Pairwise Variability Indices (PVI). The IM approach involves computing three separate measures from the segmentation of speech signals (global utterance) into vocalic (V) and consonantal (C) units (ΔV, ΔC and %V) [6]. The time-normalized metric measures (VarcoV/C) were introduced when it was observed that the consonantal interval measure is inversely proportional to speech rate [7]. The PVI algorithm differs from IM and VarcoV/C models in that it focuses on the temporal succession of the vocalic and consonantal intervals instead of the global utterance [8]. The model suggests that the rPVI should be used for the consonantal intervals, while the nPVI (normalized Pairwise Variability Index) should be used for the vocalic intervals, which are more prone to be affected by speech rate.

More recently, a new proposal of rhythm metrics has been put forward [9]. The model is inspired by previous studies on syllable compensation, which state that controlling languages (syllable-timed) show low levels of compensation at intra- and inter-syllabic levels in comparison to compensating languages (stressed-timed) that are thought to show higher levels of compensation. The rhythm metric suggested is the Compensation and Control Index (CCI) that may be used for computing the level of compression (lengthening or shortening) allowed in a language according to the context. In order to show the intra-syllabic comportment, the CCI takes into account all the segments composing each vocalic and consonantal interval. The formula used in CCI computation consists of a modification of the rPVI algorithm where each vocalic or consonantal interval is divided by the number of phonological segments that are included in this interval. A new version of the CCI formula has been developed [10], but for the purpose of this experiment, we used the original formula, shown below, where *m* stands for V or C intervals, *d* for duration, and *n* for number of segments within the relevant interval.

$$cci = \frac{100}{(m-1)} \sum_{k-1}^{m-1} \left| \frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}} \right|$$

The model aims to offer a better representation of the rhythmic tendencies of natural languages, suggesting that geminate consonants and long vowels separately count for two different segments. According to the model, the controlling languages should present similar tendencies of C and V local durational fluctuations scattered along the bisecting line. Conversely, compensating languages should fluctuate more in the V than in the C segments. They should be clustered under the bisector as shown in Figure 1 [10].

The experiment, described in this paper, focused on the rhythm properties of Arabic L2 speech by gauging the influence of L1 on L2 rhythm. A comparison of Arabic L2 with different languages is also given. The rhythm metrics that were examined include: three interval measures (%V, ΔV, and ΔC), two time-normalized indices (VarcoV, VarcoC), two pairwise variability indices (nPVI-V, rPVI-C), and two compensation and control indices (CCI-V, CCI-C).

The paper is organized as follows: Section 2 summarizes the main characteristics of the Arabic language. Section 3 gives an overview of the speech data and speakers used in the study. Section 4 describes the results and discussion about different rhythm experiments on Arabic L2. Section 5 concludes this work.

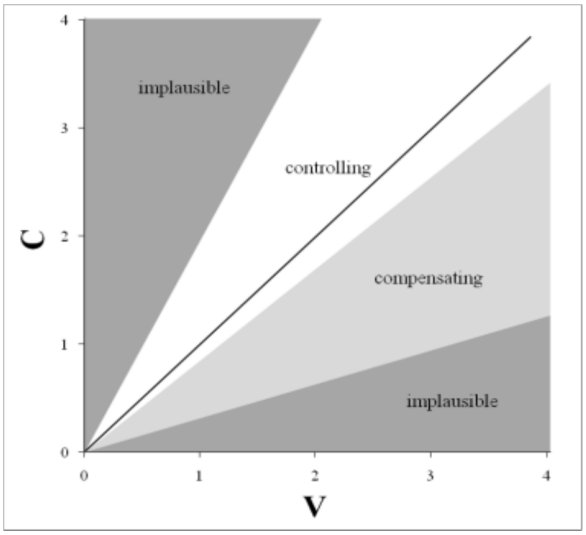


Figure 1: *Schematic representation of the major rhythmic types according to the CCI model [10].*

2. Modern Standard Arabic

Arabic is part of the Semitic language family, and spoken by more than 340 million speakers from the Middle East to North Africa. Modern Standard Arabic (MSA) is used in most written documents as well as in formal spoken interactions in all Arabic countries.

MSA uses a distinct phonological set that includes 28 consonants and six vowels: three short vowels /a/, /u/ and /i/ and three long vowels /a:/, /u:/ and /i:/, that are the counterparts of the short vowels. Regarding the number of vowels, Arabic is much less complex than English, which has twelve vowels. English and Arabic rely on different types of phonological contrasts in both vowels'quality and quantity [11].

There are two distinct consonant classes in MSA: pharyngeal and emphatic consonants. In addition to these, the language is characterized by two distinctive features that are fundamental in avoiding semantic ambiguity: long vowels and gemination. Gemination, or consonant elongation, occurs when a consonant is pronounced for a perceptibly longer period of time than a short consonant. All Arabic consonants may be geminated.

The Arabic language consists of two kinds of syllables: open syllables (CV and CV:) —and closed syllables (CVC, CV:C and CVCC). Arabic vowels never occur initially. Every vowel must be preceded by a consonant (which may include the glottal stop [ʔ]). All Arabic syllables must contain at least one vowel [12].

3. Corpus description and speakers

Speech material used for the study was taken from the West Point corpus, which was collected and processed by the Department of Foreign Languages at the United States Military Academy at West Point and the Center for Technology-Enhanced Language Learning (CTELL) (Linguistic Data Consortium (LDC) [13]). West Point speech files were recorded from 110 speakers. The corpus consists of collections of four main Arabic scripts including a total of 258 sentences. Script 1 is spoken by 75 L1 speakers of Arabic (41 males). Scripts 2–4 are read by 35 L2 speakers (25 males). The L2 speakers' recordings were captured at a sampling rate of 16 bit at 22.05 kHz. The recordings were collected at normal speech rate. The age of the speakers is not mentioned.

To reduce factors that can compromise the reliability of rhythm analysis, a large sample of speakers was used and the measurement effects across speech materials were controlled for. Speech material from 29 speakers reading either five sentences from scripts 1 or 2 was used. In total, 145 recordings were used in the analysis. Table 1 shows the number and gender of speakers in the sample.

Table 1. *Distribution of native and non-native speakers in Modern Standard Arabic (MSA) corpus.*

Native speakers (L1)		Non-native speakers (L2)	
male	female	male	Female
5	10	6	8
15		14	

To avoid variability due to the segmentation procedure, one researcher manually carried out all segmentations for the speech corpus. Recordings were analyzed using Praat software. All vowels and consonants were segmented by inspection of speech waveforms and wideband spectrograms by one researcher. Vowel and consonant durations were extracted using a customized script on the boundary label files. Rhythm metrics were computed for each sentence for each speaker.

4. Results & Discussion

4.1. IM & PVI metrics for L1 and L2 speakers

The first experiment studied rhythm metrics of L2 speakers in comparison to L1 speakers by computing different IM and PVI metrics. Table 2 reports the average values of each of the seven rhythm metrics applied to the data. Figure 2 reveals that all the vocalic metrics: vocalic interval measure, vocalic time-normalized interval, vocalic percentage, vocalic pairwise variability index (ΔV, VarcoV, %V and nPVI-V) are higher for L1 speakers in comparison to L2 rhythm values. As can be seen from the same plot, the consonantal interval measure and the consonantal pairwise variability index (ΔC and rPVI-C) for L1 speakers also present higher scores than L2 metric values. However, the VarcoC value for L2 speakers is close to the L1 measure.

Table 2. *Modern Standard Arabic (MSA) rhythm metrics (ms) for L1/L2 speakers.*

Metrics	L1 spk.	L2 spk.
%V	42.41	40.12
ΔV	49.42	45.58
ΔC	53.29	43.90
VarcoV	65.14	58.64
VarcoC	50.87	49.79
rPVI-C	74.84	68.71
nPVI-V	56.22	52.17

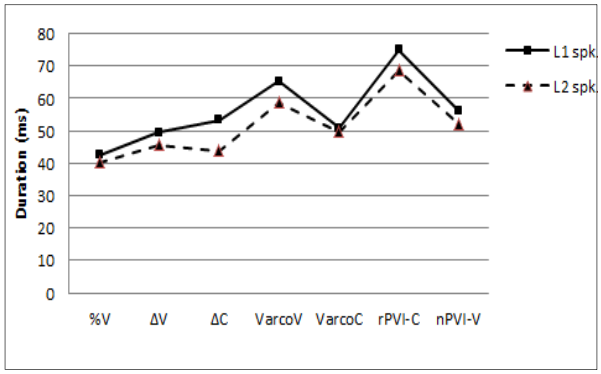


Figure 2: *Mean rhythm metrics of Modern Standard Arabic (MSA) L1 and L2.*

Statistical analysis (One-way ANOVA) shows significant effects between L1 and L2 rhythm values for three metrics: %V, ΔC and ΔV ($F(1.143)=13.17$, $p<0.001$; $F(1.143)=15.79$, $p<0.001$; $F(1.143)=8.71$, $p=.004$; 95% confidence intervals). These results show that L1 speakers present higher vocalic proportions compared to L2 speakers. This fluctuation in vocalic intervals suggests that non-native speakers reduce the vocalic intervals. Likewise, consonantal intervals of L2 speakers are less lengthened than their L1 counterparts. This finding suggests that the pronunciation of MSA consonants by non-native speakers decreases in duration compared to that of native speakers. Considering all the metrics employed, it seems that this reduction, especially in vocalic duration, helps L2 speakers to sustain syllable structures when producing MSA.

4.2. CCI metrics of L1 and L2

The CCI algorithm aims to describe the intra-syllabic behavior of languages. In the second experiment, CCI metrics were computed (vocalic/consonantal compensation and control Index; CCI-V and CCI-C) for each sentence of MSA produced by L1 and L2 speakers. Statistical analysis (One-way ANOVA) shows significant differences in CCI-C and CCI-V values between L1 and L2 rhythm ($F(1,143)=4.75$, $p=0.031$; $F(1,143)=29.78$, $p< 0.001$).

Regarding the CCI chart, the average values of each of CCI-C and CCI-V were computed for both native (L1) and non-native (L2) speakers. Comparison of the vocalic/ consonantal CCI metrics shows that L1 and L2 are separated by the bisector. As shown in Figure 3, L2 Arabic falls close to the bisecting line in the controlling area, whereas L1 Arabic is below the bisector in the compensating part.

Arabic is a stress-timed language, like English and German [14]. According to the CCI model, stress-timed languages should show higher levels of compensation between vowel and consonant segments, whereas controlling languages should present similar tendencies of C and V local durational fluctuations. However, the results reveal that Arabic L1 presents more compensating CCI while the opposite happens to Arabic L2 production, which appears to be more controlling then compensating. MSA L1 fluctuates more in the V than C segments compared to MSA L2, where the variation tendency is similar for vowels and consonants.

Some studies claim that rhythm metrics do not manage to clearly distinguish between L1 and L2 speakers [15]. However, the main findings of this experiment show that the deviation is noticeably observed in the case of Arabic L2 and L1 speakers when CCI metrics are used.

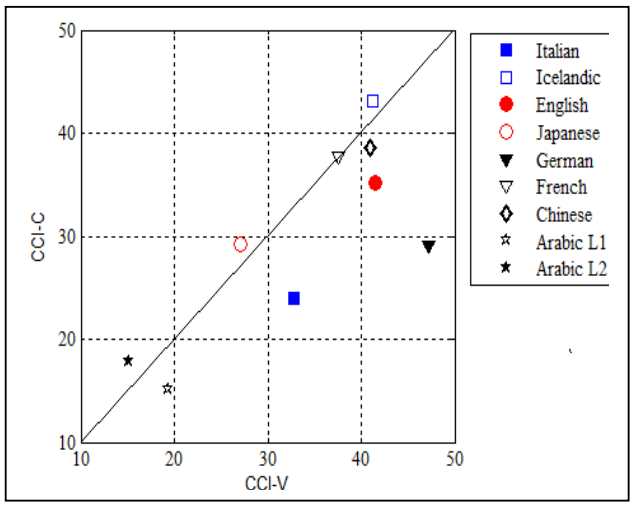


Figure 3: *MSA L1 and L2 among controlling and compensating languages.*

Figure 3 illustrates a comparison of Arabic L1 and L2 with a set of languages including English, French, Italian, Japanese, Chinese, Icelandic and German [16]. The graph shows that Arabic L2 falls in the controlling area, similar to French, which is considered one of the controlling languages. In contrast, Arabic L1 is grouped with the compensating languages, like English and German, which are clustered under the bisector. These results show that in the case of MSA, it is possible to distinguish L2 from L1, and Arabic L1 from other languages.

In this study, the L1 of non-native Arabic speakers was English. When L1 and L2 share the same rhythm typology, it may be difficult to distinguish between L1 and L2 speakers [15]. However, findings reported in Figure 2 show that distinction between L1 and L2 speakers is possible using rhythm metrics, even although Arabic and English are both stress-timed languages.

4.3. Vowel duration in L2

In light of the substantial differences between English and Arabic vowel systems (in terms of quantity) and IM, PVI, VarcoV/C and CCI–V findings, an investigation of the acoustic characteristics of MSA vowels produced by English speakers was conducted. The purpose was to show how durations fluctuate in the vocalic rhythm metric of L2 speakers. One-way ANOVA was conducted to test the effect of the origin of speakers on vowel duration values. The results show a highly significant effect ($F(1, 1347)=29.41$, $p<0.001$). Table 3 presents the comparison of the mean duration values in milliseconds of short (v) and long (V) vowels of L1 and L2 speakers.

Table 3. *Average durations of short (v) and long (V) MSA vowels by L1/L2 speakers (standard deviations are given in parentheses).*

	v (ms)	V (ms)	V/v
L1 spk.	86.62 (36.46)	170.29 (43.27)	1.97
L2 spk.	83.95 (36.41)	139.39 (43.45)	1.66

The results indicate that durations of L1/L2 short vowel values are close to each other. For long vowels, the data do not exhibit a unified pattern for both groups. L1 speakers have longer durations compared to L2 individuals. The results also show that the vowel length contrast (the ratio of long-to-short vowel durations) is greatest in the L1 group, while L2 speakers have a lower ratio. ANOVA was conducted to test the effect of vowel length (short/long) on L1/L2 speakers. The results show a significant effect of origin of speaker on vowel length [$F(1, 1345)=32.42$, $p<0.001$]. These findings suggest that the difference in duration of long vowels is mainly related to the L1 of speakers. The fluctuation observed in vocalic rhythm metric values of L2 compared with L1 speakers is due to the reduction of long vowels rather than short vowels. These results suggest that L2 speakers struggle to produce the long vowels as produced by L1 speakers.

To further examine between-speaker differences according to gender, a comparison of the mean duration of short/long vowels of female and male speakers for both L1 and L2 was performed. Results are summarized in Table 4.

Table 4. *Average durations of long and short MSA vowels by L1/L2 according to gender (standard deviations are given in parentheses).*

	v (ms)	V (ms)	VL1/VL2
L1 female	88.90 (36.99)	168.59 (45.05)	1.23
L2 female	88.09 (37.66)	136.58 (44.87)	
L1 male	82.19 (34.94)	174.10 (39.05)	1.22
L2 male	78.48 (34.02)	143.28 (44.87)	

The results show that L1 and L2 females have similar duration values for Arabic short vowels, while L1 males produce slightly longer durations than their L2 counterparts when producing short vowels. For long vowels, both L2 females and males show similarly reduced duration compared to the L1 females and males respectively (the ratio of VL1/VL2). Gender was the final independent variable tested in this experiment. Results of the ANOVA show a highly significant effect of gender on vowel length ($F(1, 1341)=8.51$, $p=0.004$). L2 speakers struggle to produce long vowels that are similar to those of L1 speakers in terms of quantity (duration), and this is true for females and males. This can be explained by the influence of the L1 of non-native speakers, in this case the phonetic system of English on that of the L2, Arabic. The L1 of non-native speakers influences the acquisition and subsequently the production of Arabic, the L2.

5. Conclusions

This study examined variation in rhythm of L2 speakers of MSA. Several experiments were conducted to show rhythm properties of Arabic L2 using IM and PVI algorithms. CCI models were also used to describe L2 production at the intra-syllabic level. The main results show that rhythm metrics of L2 speakers present many differences in terms of phoneme duration (vowels and consonants) compared to L1 speakers. To explain the fluctuation in vocalic rhythm metrics, a comparison of vowel duration (short/long) between L1 and L2 speakers was performed. The results show that the fluctuation in vocalic rhythm metrics is due to a reduction in duration of long vowels for both male and female L2 speakers.

6. Acknowledgements

This work is supported by the NPST program at King Saud University, Project Number 10INF1325-02.

7. References

[1] Jang, T.Y., "Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers' English pronunciation", In Proc. of the 2nd International Conference on East Asian Linguistics, 2009.

[2] Ling, L-E., Grabe, E. and Nolan, F., "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English", Language and speech, 43(4):377–401. 2000.

[3] Gut, U., "Non-native speech rhythm in German", In Proceedings of the ICPhS conference, 2437–2440, 2003.

[4] van Dommelen, W., "Temporal patterns in Norwegian as L2", Trends in linguistics studies and monographs, 186:121, 2007.

[5] White, L. and Mattys, S.L., "Calibrating rhythm: First language and second language studies", Journal of Phonetics, 35(4):501–522, 2007.

[6] Ramus, F., "Acoustic correlates of linguistic rhythm: Perspectives". In Speech prosody, 115–120. Citeseer, 2002.

[7] Dellwo. V., "Rhythm and speech rate: A variation coefficient for deltaC", Language and language processing, 231–241, 2006.

[8] Grabe, E. and Low, E.L., "Durational variability in speech and the rhythm class hypothesis", Papers in laboratory phonology, 7:515–546, 2002.

[9] Bertinetto, P.M. and Bertini, C., "On modelling the rhythm of natural languages", Proc. 4th International Conference on Speech Prosody, 427-430, Campinas, 2008.

[10] Zhi, N., Bertinetto, P.M., and Bertini, C., "Modelling the speech rhythm of Beijing Chinese in the CCI framework", Proc. 17th ICPhS, HongKong, 2011.

[11] Kopczynski, A. and Meliani, R., "The vowels of Arabic and English", Papers and studies in contrastive linguistics, 27:183–192, 1993.

[12] Alotaibi, Y. and Selouani, S. A., "Evaluating the MSA West Point speech corpus", International Journal of Computer Processing of Languages, 22(4): 285–304, 2009.

[13] Linguistic Data Consortium LDC. http://www.ldc.upenn.edu.

[14] Selouani, S-A., Alotaibi, Y. A. and Pan, L., "Comparing Arabic rhythm metrics among other languages", In Audio, Language and Image Processing (ICALIP), 2012 International Conference, 287–291, 2012.

[15] Gut. U., “Rhythm in L2 speech”, Speech and Language Technology, 14/15, 2012.
[16] Mairano, P. and Romano, A., “Rhythm metrics for 21 languages”, Proceedings of ICPhS XVII, Hong Kong, 17–21, 2011.

Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation

Rasmus Dall[1], *Junichi Yamagishi*[1 2], *Simon King*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, United Kingdom
[2]National Institute of Informatics, Tokyo, Japan

r.dall@sms.ed.ac.uk, jyamagis@inf.ed.ac.uk, simon.king@ed.ac.uk

Abstract

In this paper we present evidence that speech produced spontaneously in a conversation is considered more natural than read prompts. We also explore the relationship between participants' expectations of the speech style under evaluation and their actual ratings. In successive listening tests subjects rated the naturalness of either spontaneously produced, read aloud or written sentences, with instructions toward either conversational, reading or general naturalness. It was found that, when presented with spontaneous or read aloud speech, participants consistently rated spontaneous speech more natural - even when asked to rate naturalness in the reading case. Presented with only text, participants generally preferred transcriptions of spontaneous utterances, except when asked to evaluate naturalness in terms of reading aloud. This has implications for the application of MOS-scale naturalness ratings in Speech Synthesis, and potentially on the type of data suitable for use both in general TTS, dialogue systems and specifically in Conversational TTS, in which the goal is to reproduce speech as it is produced in a spontaneous conversational setting.

Index Terms: speech synthesis, evaluation, naturalness, MOS, spontaneous speech, read speech, TTS

1. Introduction

In speech synthesis research there are two generally used methods for evaluation, namely intelligibility and naturalness. Intelligibility is a metric which has robust measures such as semantically unpredictable sentences (SUS) [1] and synthesis systems perform well compared to natural sentences [2, 3]. Naturalness on the other hand is a less defined concept, although it is generally always used e.g. in the Blizzard challenges [2, 4, 5]. It is also used to evaluate prosody and is the focus of this paper.

Naturalness is normally evaluated as a Mean Opinion Score (MOS) where participants rate the quality of the synthetic speech on a 5-point scale ranging from 1-Very Unnatural to 5-Very Natural. The scale itself has not been much investigated, however the Blizzard 2008 [2] evaluation gave support to the scale being treated, by listeners, as an interval rather than ordinal scale by comparing it to scores obtained using an unnumbered slider. While systems tend to perform well on intelligibility they are generally lacking behind natural speech in terms of naturalness. One assumption made in several conversational speech synthesis studies is, that spontaneous conversational speech is more natural than read speech [6–8]. Thus, it is assumed, synthesis based on conversational speech will similarly increase the system's naturalness. However, it has not been shown that people actually find conversational speech more natural than read speech, and earlier studies using spontaneous recordings have not managed to increase the perceived naturalness of synthetic speech [6, 9]. People can distinguish the two modes of speech with high accuracy despite lexical equivalence [10], so it is likely that people will be able to pick up upon and judge according to this distinction when asked. This study attempts to test this by obtaining naturalness ratings of natural speech from the same speakers, of speech produced spontaneously in a conversation and when reading aloud. We hypothesise, as has been done before, that conversational speech is considered more natural.

It is also likely that 'naturalness' as a concept is underspecified. That is, we do not have an exact definition of what naturalness is. In fact differing studies give participants differing instructions. The Blizzard 2013 evaluation [11] instructs participants to give a score which "should reflect your opinion of how natural or unnatural the sentence sounded. You should not judge the grammar or content of the sentence, just how it sounds." In contrast [12] explains the meaning of naturalness as if it is "likely that a person would have said it this way?" (p.470). The two stand in contrast to each other, the one asking to disregard grammar and content, and the other to judge the 'way' it was said - including content and grammar. If listeners do find it to be underspecified then people's perceptions should be be influenced by their expectations of what naturalness means in any given context. We therefore attempt to influence the prior expectations of listeners by slight variations in instructions to bias them toward either conversational or read speech, and compare this to the general case with no further instructions.

Note that there are genuine worries about the ecological validity of MOS-scale naturalness tests of isolated sentences presented in very controlled noise environments. It is not the purpose of this paper to attempt to rectify these, but rather to explore current means and enable further detail in their application. Section 2 describes our first listening test, in Section 3 we attempt to separate audio and text and Section 4 discuss the overall implications, before concluding in Section 5.

2. Naturalness Ratings of Spontaneous and Read Speech

A simple way of testing if there is a preference for conversational over read utterances is to mimic the standard naturalness test setup. In such a procedure the common instruction is for the participant to listen to one sentence at a time, rating how natural they find the sentence. That is people are only told to rate what sounds 'natural' with no further qualification. If naturalness is an underspecified concept it should be possible to influences people's ratings by slightly changing the given instructions, and as we are concerned with the difference between conversational and read speech we attempt to influence people's perceptions in these directions. Instead of closely matching the content of these sentences by rating the same sentences either spoken in a conversation or read aloud (see Section 3), it was decided to ini-

Read	Spontaneous
Challenge and errors both go well.	It's kinda ridiculous, but it was funny at the time.
Author of the Danger Trail Philip Steel etc.	When I was younger I... loved uhm Ang Lee.
How funny is your funniest joke?	Absolutely, I'm sure there are evil kings with rotten voices.
Officials have no evidence yet that the plane could have been sabotaged.	And at the point where it goes into the park, the tunnel goes underneath at that point.

Table 1: Example sentences.

tially use sentences representative of the respective styles to see if a difference was to be found in a fairly unconstrained setting.

2.1. Data

Studio recordings of conversational and read-aloud data from two differing speakers, one male and one female, was used as the stimuli. For each speaker 30 conversational and 30 read sentences were selected. For the read sentences the female data included mainly read news text and the male data was the first 30 sentences of the Arctic prompts [13]. The conversational utterances were chosen from recordings of the speakers having an unscripted conversation with an experimenter. The sentences were chosen so as to be complete sentences with no initial or final disfluency, although disfluencies were allowed in the sentences. Where the read-prompts had a distinct third-person perspective most conversational sentences in the database were first person. To reduce this mismatch, conversational sentences were chosen to generally be about something rather than the speaker him/herself. Sentences in both conditions were also matched for length with the shortest being about 2s long and the longest about 6s. Table 1 provides a few example utterances and audio samples are available.[1]

2.2. Method

32 paid native speakers of English were recruited, mainly students at the University of Edinburgh. 11 participants rated general naturalness (GenNat), 10 conversational naturalness (ConvNat) and 11 participants reading naturalness (ReadNat). Participants were instructed to rate the sentences in the standard TTS paradigm and they were instructed to "Listen to each sentence and rate it according to how natural you find the sentence from a scale of 1 - Very Unnatural to 5 - Very Natural" in the GenNat case, in the ConvNat the sentence "if you were having a conversation" was added between "sentence" and "from"; in the ReadNat case "if somebody was reading aloud" was added in the same place. This difference in instruction was the only difference between conditions. Each participant rated all 120 sentences once, in a randomised order of presentation for each participant. Each participant also rated an additional 5 sentences as a trial run to get accustomed to the methodology. After the trial run participants were encouraged to ask clarifying questions before proceeding to the main part of the test. The test was performed in a soundproof room with the participants wearing good quality headphones. The test took about 15 minutes to complete. There were three groups of participants (GenNat, ConvNat and ReadNat) and two types of audio (conversational or read).

[1] http://rasmus.dall.dk/SP2014Samples.zip

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	660	660	600	600	660	660
Mean	2.98	4.23	2.62	4.04	3.67	3.74
SD	1.192	1.131	1.291	1.189	1.182	1.466
p	p<0.0001		p <0.0001		p=0.352	

Table 2: Condition descriptives. The shown significances are between spontaneous and read sentences for each condition.

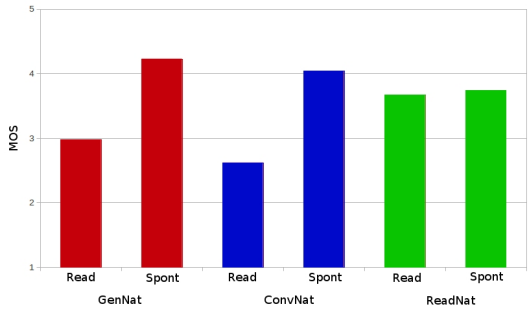


Figure 1: Overall ratings by category.

2.3. Results

As noted in Section 1 we have evidence that the 5-point MOS scale is used as an interval scale and not in an ordinal fashion, therefore we can meaningfully compare the means instead of the medians of the ratings [14]. No null responses were recorded and all ratings were used in the analysis. A significant difference was found between the read (M=2.98, SD=1.192) and conversation (M=4.23, SD=1.131) sentences in the GenNat group (t(1318)=19.644, p <0.001), this was also the case for ConvNat (Read: M=2.62, SD=1.291; Conv: M=4.04, SD=1.189; t(1198)=19.848, p <0.001) but not the ReadNat condition (Read: M=3.67, SD=1.182; Conv: M=3.74, SD=1.466; t(1318)=0.93, p=0.352). In other words, when asked to rate what they found natural with no further instruction, or instructions toward conversation, participants preferred the spontaneous utterances, however there was no preference when rating naturalness for reading aloud. See Table 2. Across instruction conditions one-way ANOVA's were run for each speech type. An effect for both read (F(2,1917)=122.285, p <0.001) and spontaneous utterances (F(2, 1917)=25.509, p <0.001) were found. Bonferroni correction showed all differences to be significant at the p <0.001 level for the read speech and for the spontaneous speech all differences were significant at the p <0.001 level except GenNat and ConvNat which was significant at p <0.05. Thus different instructions gave different ratings. It is possible that the findings are speaker specific or gender specific. Repeating the tests by speaker we find that the effects are slightly smaller for the male speaker and larger for the female, however both speakers exhibit the same tendencies with the same significant differences suggesting that, at least in this small sample, neither speaker or gender affects the results.

3. Separating Acoustics and Text

While we see a difference in a fairly unconstrained setting, it is clear that the content of the read and conversational sentences was quite different despite ensuring that each spontaneous utterance was "complete". It is therefore possible that the prefer-

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	248	246	249	249	247	246
Mean	2.79	4.29	2.99	4.09	3.36	4.07
SD	1.292	0.915	1.292	0.938	1.114	1.145
p	p<0.0001		p <0.0001		p <0.0001	

Table 3: Descriptives for the audio data. The significances are between spontaneous and read sentences for each condition.

	GenNat		ConvNat		ReadNat	
	Read	Spont	Read	Spont	Read	Spont
N	399	399	399	400	395	397
Mean	3.36	3.72	2.73	3.81	3.76	3.07
SD	1.385	1.286	1.280	1.213	1.139	1.404
p	p<0.001		p <0.0001		p <0.0001	

Table 4: Descriptives for the textual data. The significances are between spontaneous and read sentences for each condition.

ences found are not due to differences in articulation or speech mode - but rather due to differences in content. The opposite, however, is also possible, that is, the content has nothing to say and only the acoustic differences matter. In order to tear this apart further we need to isolate the two possibilities. This is possible in the following way, firstly in order to test whether it is purely the content of the utterance which affect people's perception, we can elicit ratings from people based on text only. That is by comparing normal written text - e.g. from newspapers or novels - with transcriptions of conversational speech we can avoid the acoustic component entirely and focus purely on the content. Secondly we can isolate the acoustic component by recording a speaker in a conversational setting and then, at a later time, ask the same speaker to re-read transcriptions of their own earlier utterances. The content of the utterances will be the same however the mode of speech will differ. In this way we can tear apart the effects of content and mode.

3.1. Data

One acoustic and one textual dataset was obtained. The acoustic data consisted of studio recordings of 50 sentences initially produced in a longer conversation by a female speaker with one of the experimenters. From this conversation 50 complete (as above) sentences were identified and transcribed. The speaker was then, a few days after the first recording, asked to re-read the sentences by having them given as prompts.[2] The textual data consisted of 120 sentences. Half were taken from transcriptions of spontaneous data and the other from written sources. The transcribed data was obtained 50/50 from two generally available corpora of spontaneous data (AMI [15] and Switchboard [16]). The written data contained 30 sentences from the Arctic [13] scripts and the last 30 sentences were from News data taken from prompts used in the Edinburgh Voicebank Project [17]. For both types, novels and news, names and quotes were avoided as none were included in the spontaneous and their length matched to the spontaneous in terms of numbers of words. The choice of using various sources for both written and spontaneous data, and the inclusion of disfluencies, was to enable analysis of the possibility of internal variation depending on the style of the textual data but this analysis is not presented here due to space constraints, however we note that it

[2]Samples are available at http://rasmus.dall.dk/SP2014Samples.zip

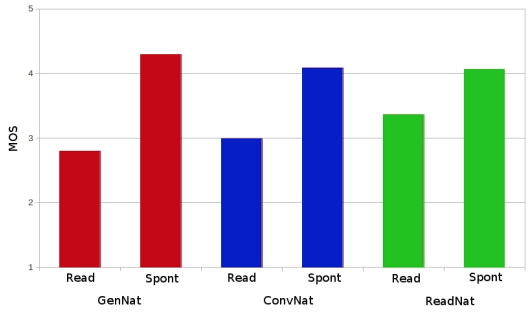


Figure 2: Naturalness ratings for the audio.

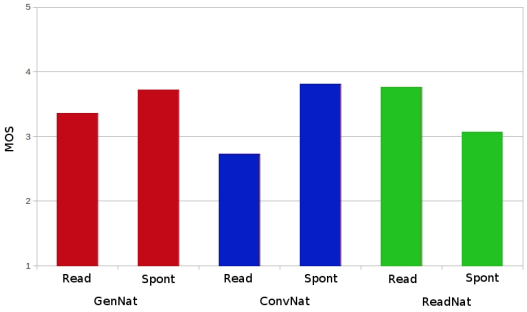


Figure 3: Naturalness ratings for the text.

does not significantly affect the presented results. An example sentence of each type can be found in Table 5.

3.2. Method

30 paid native speakers of English, mainly students at the University of Edinburgh, were recruited to take part. The general method was similar to the first experiment except as noted below. As before, each participant was assigned one of three groups - general naturalness (GenNat), conversational naturalness (ConvNat) or reading naturalness (ReadNat). The test had two sections. Section 1 consisted of the 50 audio samples and 4 test samples, two spontaneous and two read. Section 2 contained the 120 textual samples and 6 test samples, one of each text type. Except for test samples all presentation was randomised for each participant. In section 1 participants were asked to rate for naturalness according to their group as in experiment 1. In section 2 participants were asked to imagine that the sentence was either "spoken aloud" (GenNat), "said in a conversation" (ConvNat) or "read aloud" (ReadNat), and then judge how natural the sentence would be. In total the test took about 15 minutes to complete.

3.3. Audio Results

15 responses (1%) null responses were excluded. For the GenNat (t(492)=14.864, p <0.0001) and ConvNat (t(496)=10.837, p <0.0001) groups we see a repetition of the previous results with spontaneous speech being significantly preferred over read prompts (Table 3). Contrary to earlier we now have a significant difference for the ReadNat group (t(491)=6.888, p <0.0001) - that is *spontaneous* speech is significantly preferred over read speech (see Figure 2). Again one-way ANOVA's were run for each speech type across groups. Here we find that no difference exists for read speech (F(2, 746)=2.693, p=0.068) - ratings

Source	Example
AMI	Yeah, but you can appreciate the way they look.
SB	I do try and regulate how much exercise I get a week.
Arctic	Unconsciously, our yells and exclamations yielded to this rhythm.
News	The current deployment is designed as a deterrent.

Table 5: Example textual sentences. SB = Switchboard.

of reading naturalness did not change with instructions. However for the spontaneous speech a significant difference was found (F(2, 746)=12.197, p <0.0001) and Bonferroni correction showed the read group to be significantly (at p <0.01) different to the general and conversational group, no difference existed between those (p=0.154). In other words, instructions toward rating for reading naturalness changed peoples perception toward a higher preference for read speech.

3.4. Text Results

11 responses (0.5%) null responses were excluded. In both the GenNat (t(797)=3.877, p <0.001) and ConvNat (t(796)=12.207, p <0.0001) groups the transcribed text was significantly preferred. However, the ReadNat group significantly preferred the *written* text (t(790)=7.694, p <0.0001) (see Table 4). When imagining text spoken aloud or said in a conversation people find transcriptions of spontaneous speech over textual sources more natural - but when imagining it read aloud people found written text more natural. One-way ANOVA's support the conclusion that instructions affect peoples perceptions. For the transcriptions (F(2, 1196)=41.058, p <0.0001) Bonferroni correction showed the GenNat and ConvNat groups to differ significantly from the ReadNat group (both at p <0.0001) however not in between themselves (p=1). That is, only when rating for reading naturalness are peoples ratings affected by instructions for transcribed speech, and then towards being less natural (see Figure 3). In the written case there was also a significant effect (F(2, 1196)=58.978, p <0.0001) and with Bonferroni correction all differences were significant (p <0.001). So, when rating written text the instructions consistently affected peoples perceptions, people found written text the least natural when rating for ConvNat, more for GenNat and most natural for ReadNat (see Figure 3).

4. General Discussion

The perception of naturalness changes in the context in which it is rated, by simply adding "if you were having a conversation" or "if somebody was reading aloud" the ratings change. When no instructions were given as to what kind of naturalness to rate, participants find spontaneously produced utterances to be more natural - in line with the assumptions of earlier research. In experiment 1 the ReadNat group showed no preference for either mode of speech, when explicitly asked to rate according to naturalness when reading aloud, participants found spontaneously produced utterances *equally* natural. However, when tearing apart audio and text we see a general acoustic preference for spontaneous speech and a preference dependent on instructions for textual stimuli. Thus spontaneously produced utterances are *always* more natural acoustically than read speech - suggesting conversational speech to be the, generally speaking, most natural of the two modes of speech. If this is true it has consequences for how we should be doing speech synthesis. Assuming improved naturalness is the main current challenge in speech synthesis (in particular HMM-based) then it suggests that we should be utilising the preference for conversational speech by basing our models on such speech. This is particularly true if we wish to synthesise conversational speech, but even if we wish to make the most broadly applicable speech synthesis system we should not assume that read speech is a neutral middle ground, that may in fact be conversational speech. This is also supported by the contextual preference for transcribed speech over actual written sources.

From the second experiment, we can see that combining the general preference for spontaneous speech in the audio and the textual results, in which we see a preference for the written sources only for the ReadNat group, yields us the same picture as given in the first experiment. That is, we have successfully managed to tear apart the difference between the acoustics and the meaning content of the sentence by removing the variables in their respective tests. It is important to note that, for the textual case, we have focused on the spoken word, not the written, by instructing participants to rate it according to how natural it would be in various spoken scenarios and not how natural it would be focusing on it as text. In light of the clear effect of instructions on peoples ratings (more below) we would expect instructions geared toward *written* naturalness to yield a differing result. Both the first and the second tests support the hypothesis that naturalness as a metric can be easily influenced by experimental instructions, and that the influence is dependent on the type of data under consideration. This is likely due to the concept of naturalness in general being under-specified, and so by conditioning the experimental setting we can influence our participants toward various interpretations. Knowing this encourages both caution and enables more detail when evaluating synthetic speech. Caution because we must be diligent with the instructions we give participants so as not to bias them in an unwanted direction. More detail as we can condition the metric toward specific aspects of naturalness.

5. Conclusions and Further Work

We have shown that MOS-scale ratings can beneficially be employed to distinguish the conversationality of speech, in fact spontaneous conversational speech is found more natural by listeners than read prompts. We can affect peoples perception of naturalness by simple conditions in the instructions, enabling greater control over the testing scenario while also cautioning its use. Further work includes rigidly defining what is natural in the general case, but also attempting to utilise the apparent advantages of conversational speech. Our results suggests that read prompts may not be the neutral general speech as previously assumed and that this role is more likely attributable to spontaneous conversational speech. The gathering and use of such speech present many challenges which must be met before it is generally applicable, however we intend to attempt the use of such data by gathering an appropriate spontaneous corpora, but also by utilising existing data not recorded specifically for speech synthesis.

6. Acknowledgements

Thanks to Cereproc Ltd. for providing the female data for the first listening test. This work was funded by the JST CREST uDialogue project.

7. References

[1] C. Benoit, M. Grice, and V. Hazan, "The SUS test : A method for the assessment of text-to-speech synthesis intelligibillity using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, pp. 381–392, 1996.

[2] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Blizzard Challenge Workshop*, 2008.

[3] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Blizzard Challenge Workshop*, Bonn, Germany, 2007, pp. 1–6.

[4] S. King and V. Karaiskos, "The Blizzard Challenge 2009," in *Blizzard Challenge Workshop*, 2009.

[5] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Blizzard Challenge Workshop*, 2007, pp. 1–12.

[6] J. Adell, A. Bonafonte, and D. Escudero-mancebo, "Modelling Filled Pauses Prosody to Synthesise Disfluent Speech," in *Speech Prosody*, Chicago, USA, 2010.

[7] S. Andersson, J. Yamagishi, and R. A. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 175–188, Feb. 2012.

[8] N. Campbell, "Towards Conversational Speech Synthesis; Lessons Learned from the Expressive Speech Processing Project," in *SSW6*, Bonn, Germany, 2007, pp. 22–27.

[9] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational Spontaneous Speech Synthesis Using Average Voice Model," in *Interspeech*, no. September, Makuhari, Japan, 2010, pp. 853–856.

[10] E. Blaauw, "Phonetic Characteristics of Spontaneous and Read-Aloud Speech," in *ESCA Workshop on Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication*, no. September, Barcelona, Spain, 1991, pp. 1–5.

[11] S. King and V. Karaiskos, "The Blizzard Challenge 2013," in *Blizzard Challenge Workshop*, Barcelona, Spain, 2013.

[12] J. Adell, D. Escudero, and A. Bonafonte, "Production of filled pauses in concatenative speech synthesis based on the underlying fluent sentence," *Speech Communication*, vol. 54, no. 3, pp. 459–476, Mar. 2012.

[13] J. Kominek and A. W. Black, "CMU ARCTIC databases for speech synthesis," Carnegie Mellon University, Tech. Rep., 2003.

[14] H. M. Marcus-roberts and F. S. Roberts, "Meaningless Statistics," *Journal of Educational Statistics*, vol. 12, no. 4, pp. 383–394, 1987.

[15] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[16] J. J. Goodfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, San Francisco, CA, USA, 1992, pp. 517–520.

[17] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

A Simplified Method of Learning Underlying Articulatory Pitch Target

Hao Liu, Yi Xu

Department of Speech, Hearing and Phonetic Sciences, University College London, UK

{h.liu.12, yi.xu}@ucl.ac.uk

Abstract

Previous research has shown that parameters of the quantitative Target Approximation model (qTA) proposed by Prom-on and Xu can be directly extracted from natural speech with high accuracy through analysis-by-synthesis implemented in PENTAtrainers. While this may raise the possibility that PENTAtrainers actually simulate natural acquisition of prosody production, it is questionable that the human brain actually replicates the full articulatory mechanics represented by qTA in order to learn and control prosody production. In this paper we explore if a much simpler function can be used to extract at least some of the qTA parameters. We first managed to reduce the number of qTA parameters from three to two by evaluating their relative sensitivity. We then tested a pursuit function that learns only pitch target height and slope. Using a corpus of Mandarin utterances varying in lexical tone and focus, we show that parameters learned by the pursuit function can be used in qTA synthesis to generate F0 contours closely resembling those generated with parameters learned with qTA-based analysis-by-synthesis, with the advantage of having a much simpler learning algorithm. These results suggest that it is possible to learn articulatory control parameters for prosody without fully replicating the mechanical process itself.

Index Terms: F0 contour modelling, target approximation, pursuit curve

1. Introduction

It has been recently demonstrated that F0 contours closely resembling those of natural speech can be generated by the PENTA model [1] with a small number of functionally specific pitch targets extracted directly from raw speech data [2, 3]. The F0 contour generation in those studies is done by the quantitative target approximation model (qTA), which simulates a third-order linear system [2]. Two automatic algorithms have been developed, as implemented in PENTAtrainer1 [4] and PENTAtrainer2 [5], to extract the parameters of qTA model from functionally annotated speech data using analysis-by-synthesis controlled by either exhaustive [4] or stochastic [5] optimizations. Such parameter extraction processes could be imagined as analogous to the natural speech acquisition process in which the child presumably learns to speak by discovering, also through analysis-by-synthesis [6], the articulatory control parameters needed to generate adult-like speech patterns. There are two potential problems with this analogy, however. The first is that the number of analysis-by-synthesis cycles is unrealistically large. The second problem is that the analogy assumes that either the child overtly imitates the same adult utterance over and over again, or develops a virtual replica of the qTA model in the brain for both learning and controlling the production of tone and intonation. With these problems in mind, in this paper we explore an alternative learning mechanism that a) uses a simplified model that approximates the core properties of qTA, and b) does not require analysis-by-synthesis searching process.

We will first try to reduce the model complexity from qTA by reducing the number of parameters from three to two by comparing the relative sensitivities of the three model parameters of qTA. We will then test a "pursuit" function [7], using a corpus of Mandarin utterances varying in lexical tone and focus, to show that the pursuit function can learn the two remaining target parameters directly, with the learned values very similar to those found by exhaustive analysis-by-synthesis as implemented in PENTAtrainer1.

2. Pitch modeling

2.1. Target Approximation model

The quantitative target approximation model (qTA) assumes that continuous surface F0 contours are the results of successive, yet non-overlapping underlying articulatory movements, each approaching an underlying target associated with a local host syllable. A target can be either static or dynamic (Figure 1), which can be represented by a simple linear equation:

$$x(t) = mt + b, \tag{1}$$

where b is target height, m is target slope and t is time relative to the onset of the host syllable.

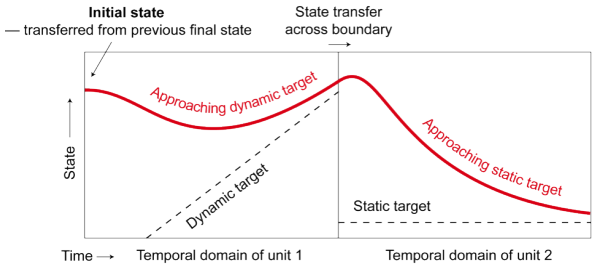


Figure 1: *Target approximation model.*

The qTA model is a third-order critically damped linear system as represented by the following equation

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t}, \tag{2}$$

where $f_0(t)$ is the complete form of the fundamental frequency in semitones, $x(t)$ is the forced response and the polynomial and the exponential are the natural response [2]. λ is the rate of target approximation, i.e., how rapidly the target is approached. The transient coefficients c_1, c_2 and c_3 are jointly determined by the initial F0 dynamic state of the syllable, consisting of F0

level, velocity, and acceleration transferred from the offset of the preceding syllable:

$$c_1 = f_0(0) - b, \tag{3}$$

$$c_2 = f_0'(0) + c_1\lambda - m, \tag{4}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2. \tag{5}$$

At the end of the syllable, the final F0 dynamic state is transferred to the next syllable to become its initial state, which results in a smooth and continuous F0 trajectory across the syllable boundary.

2.2. Control parameter sensitivity assessment

In order to find a simple model that can approximate the qTA model we examined the sensitivity of F0 contours generated by qTA to variations in the three pitch target parameters: m, b, λ.

A six syllable Mandarin phrase, /wó yǒu yí wèi yá yī/, was chosen and recorded by a male native speaker of Mandarin. PENTAtrainer1 [4] was used to find an optimal combination of m, b and λ. Then, three sets of F0 contours were generated by varying one parameter while holding the other two constant at their optimal values. The difference between the generated contour and the optimum contour was then analysed in terms of semitone shifts, as described in the paragraphs below.

Figure 2 displays the error vectors for which the target slope m and TA rate λ are set to be the same as the optimal values but target height b is given five different values. The graph is from the rising tone syllable /wó/. The pattern distributions of error vectors are very regular — all the curves gradually move away from x-axis at the same pace. The common starting point

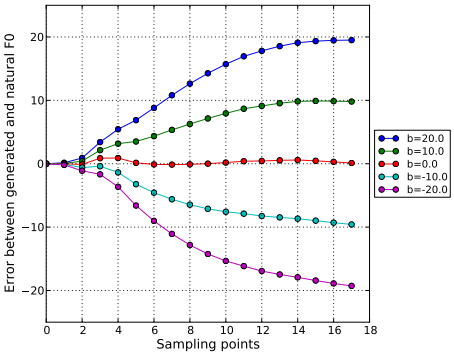


Figure 2: *Error vectors with varying b while m and λ are the same as the optimal values. Measured in semitone.*

is because qTA is a sequential model, and all variations in the current syllable step from the same offset value of the previous syllable. Because /wó/ is the first syllable of the chosen utterance, the starting point is always zero. When the values of b are equidistant from each other, the error vector curves exhibit an even distribution. Note that the middle curve represents the error vector resulting from subtracting the natural F0 contour of the syllable from the contour generated with the "optimal" parameters. The very small deviations from the x-axis indicates that the two contours are very similar to each other.

Figure 3 displays error vectors of contours that vary in m while b and λ are held constant at their optimal values. The fusiform shaped distribution here is due to the fact that in qTA b is defined as the ending point of a target. As a result, all

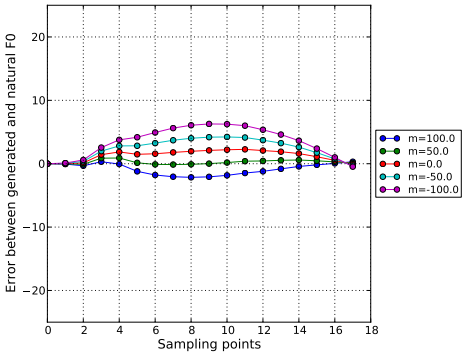


Figure 3: *Error vectors with varying m while b and λ are the same as the optimal. Measured in semitone.*

the generated contours have a fixed tail height by the end of the syllable, i.e., they shared the same offset. This means that, when b is held at its optimal value, the offset F0 of a syllable is virtually guaranteed to be near optimum.

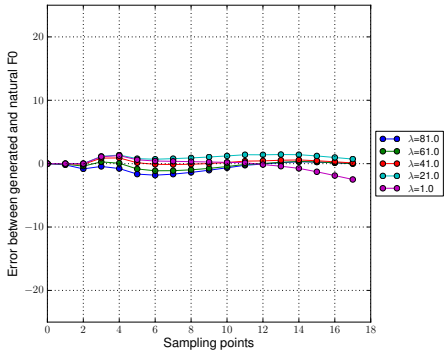


Figure 4: *Error vectors with varying λ while m and b are the same as the optimal. Measured in semitone.*

Figure 4 displays error vectors of cases where only λ is set to vary from the optimal value. The deviations are very small, indicating a much weaker effect than those of b and m. However, this does not mean that λ is unimportant in all cases. It has in fact been demonstrated that, when modelling data contain unstressed syllables and the neutral tone, the role of λ is crucial [3, 8].

The conclusion is that among the three qTA parameters, λ, i.e., target approximation rate, is less important than b, target height and m, target slope, at least for the present data set. It further suggests that a learning procedure that can find close-fitted target heights and target slopes may provide a good approximation to qTA, especially for data where λ does not have important functional significance.

2.3. Pursuit functions for pitch target estimation

The qTA model was designed to generate contours from underlying pitch targets. It is a third-order differential equation which contains non-linear elements. In this form it is not easy for learning, i.e., finding optimum underlying pitch targets from input signal. There is no analytical expression for the inverse of qTA, and so analysis-by-synthesis has to be used to estimate the model parameters from natural speech data. To make mathematical inversion possible, we investigated a method involving "pursuit" functions.

A pursuit curve is the path of an object that seeks to pursue another moving object. Consider a simple case of a hound chasing a fox, where the fox is moving at constant speed and constant direction. The pursuit curve is found under the assumption that the direction of the hound is always towards the current location of the fox, i.e. that the tangent of the pursuit curve at time t is directed towards the location of the hound at time t. Figure 5 shows an example pursuit curve.

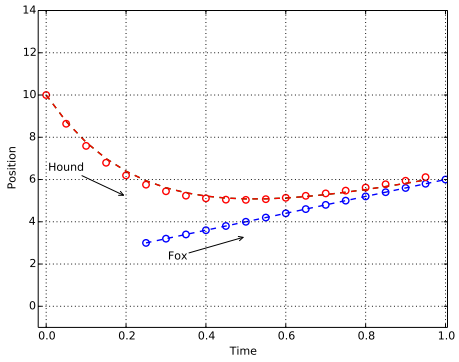


Figure 5: *A pursuit curve. Arrows indicate velocity directions.*

In the situation shown in Figure 5, where the pursued object is moving at constant velocity, the pursuit curve can be shown to have the form of an analytic equation

$$H(t) = F_0 + vt + (H_0 - F_0)e^{-t/l}, \quad (6)$$

where H_0, F_0, v, t, l denotes the initial position of the hound, initial position of the fox, velocity of the fox, time series and the "time lead" of the fox, respectively. We can interpret this as the pursuer attaining the location and velocity of the pursued according to some exponentially decreasing value of time. The rate of attainment is simply related to the time lead of the pursued.

If we use a pursuit curve to simulate the target approximation process, the linear path of the pursued becomes the underlying pitch target, the pursuit curve is the observed F0 contour, and the initial velocity of the pursuer becomes the initial conditions for the F0 at syllable onset.

As a simpler target estimation function, the pursuit function itself does not fit F0 contours as cleanly as qTA since it allows for instantaneous changes in velocity and acceleration at syllable boundaries. This can lead to rather unnatural looking F0 contours as the pursuer changes from one pursued target to another, as can be seen in Figure 6.

Since, however, we only want a simple way of deriving pitch targets represented by b and m from the input contours,

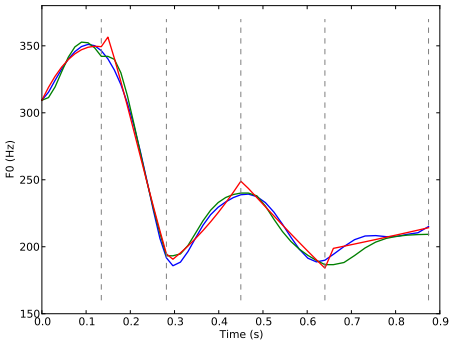


Figure 6: *Clear sharp turning points of pursuit curve at syllable boundaries. (Natural contour in blue, qTA fitted contour in green, pursuit curve in red.)*

these discontinuities may not be important. This is because, once the targets are learned, we will still use qTA to generate the contours from the targets, free of discontinuities at syllable boundaries.

In the next section we will compare the pursuit function with qTA-based analysis-by-synthesis in terms of quality of fit using a small corpus of utterances.

3. Pitch target learning

3.1. Data

480 utterances recorded from a female native Mandarin Chinese speaker were used to explore the fit of the target approximation model. This corpus was originally collected to examine the effects of lexical tones and focus on the formation and alignment of F0 contours [9]. The corpus consists of 24 sentences, each of them was said with four different focus locations and repeated five times. Every sentence consists of three Chinese words, the first and the third are bisyllabic and the second is monosyllabic. So there are five syllables in each sentence (Table 1). Further, the second, third and fourth syllables have varying lexical tones, which were the target syllables for the current experiment. The four focus conditions are: neutral focus (no focus), initial focus (on word 1), medial focus (on word 2) and final focus (on word 3). When a syllable is on-focus, its preceding syllable is pre-focus and its following syllable is post-focus.

Table 1: Tone patterns and corresponding sentences used as recording material. H, R, L, and F represent high, rising, low, and falling tones, respectively.

Word 1	Word 2	Word 3
HH māomī	H mō	HH māomī
HR māomí	R ná	LH mǎdāo
HL māomǐ	F mài	
HF māomì		

3.2. Method

The goal of the experiment was to learn the optimal pitch target slope and height, for each lexical tone in each focus con-

dition, using pitch targets learned by both qTA-based analysis-by-synthesis and the pursuit curve function. Results are then compared in terms of similarity of the discovered targets, overall quality of fit.

Both methods were implemented in Python. In the case of qTA, the exhaustive local search algorithm proposed by [2] was used, with rate of target approximation (λ) held constant at 41.0. The algorithm read the data and parameter constraints, and then iteratively tested all combinations of target values using a set of possible values for target height and target slope for each utterance separately. The parameters that showed the lowest sum square error between the generated and natural F0 contours for each utterance were chosen as the target. The optimal targets for each utterance were then averaged to derive different targets for each tone and focus condition.

For testing the pursuit function, the time lead of the pursued (l) was fixed at 0.075s. Like the qTA approximation rate parameter, the pursued time lead controls the rate at which approximation takes place and might be considered a characteristic of the speaker or speaking style [8, 10].

To fit the pursuit function, a linear least squares method was used over the whole data set. Each observed F0 measurement was expressed in terms of a number of coefficients applied to a vector of 32 unknowns, being the target height and slopes of the 4 tones in the 4 focus conditions. The least squares fit derives the values of the 32 unknowns that minimise the squared error of prediction of the data by the model.

3.3. Results and evaluation

The value of the pursuit function is that it provides a direct means to determine the optimal pitch targets from the measured F0. We would still like those targets to be compatible with qTA, since as mentioned above, the pursuit function has some intrinsic inadequacies for F0 contour generation.

Table 2: Learned functional pitch targets. For focus function, PRE, ON, and POS stand for pre-focus, on-focus, and post-focus regions, respectively.

Focus	Tone	Target slope (st/s)		Target height (st)	
		qTA	pursuit	qTA	pursuit
PRE	H	2.1	7.4	19.0	18.2
	R	30.7	17.7	11.6	12.9
	L	-16.4	-40.9	16.5	20.1
	F	-29.5	-27.3	21.5	21.6
ON	H	2.5	3.3	19.4	19.3
	R	28.7	12.5	10.8	13.0
	L	-79.1	-108.4	21.6	27.0
	F	-56.4	-43.2	28.3	27.1
POS	H	-11.6	-16.8	16.4	17.3
	R	11.9	-1.9	11.3	13.1
	L	-79.7	-119.0	20.4	27.3
	F	-36.8	-35.5	20.1	20.4

As shown in Table 2, the pitch target heights found using the pursuit function were very similar to those found using qTA. There are greater differences between the estimated pitch target slopes, but they are broadly of similar sign and size.

Figure 7 illustrates F0 contours generated from the qTA targets and pursuit targets shown in Table 2 for an utterance containing a "LHL" syllable sequence with focus on the second syllable. The generated contours for qTA and the pursuit function are similar to each other and to the measured F0.

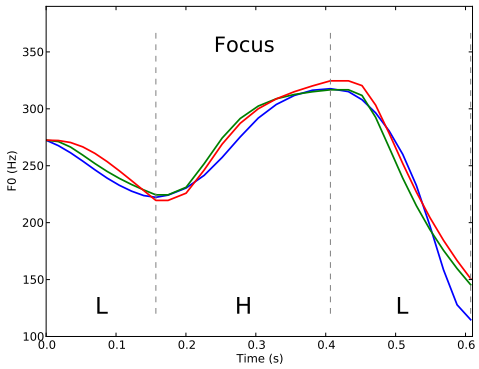


Figure 7: *Contour generated with pursuit targets is very close to the one with qTA targets. (Natural contour in blue, contour generated with qTA targets and pursuit targets are in green and red, respectively.)*

The quality of the targets obtained by the two methods was evaluated by measuring the root-mean-square error (RMSE) of prediction of the data set using the qTA model and the discovered targets. For the targets found by qTA-based exhaustive search, the RMSE of prediction is 1.83 semitones (Pearson $r = 0.8453$). For the targets found by the pursuit function, the RMSE is 1.81 semitones (Pearson $r = 0.8458$). Thus the new method of estimating underlying targets is as least as good as using the qTA model directly for inversion.

In terms of learning efficiency, although the linear least squares with pursuit function can get target parameters in a blink of time, we didn't expect it to be a replacement of simulated annealing as implememnted in PENTAtrainer2. Instead, we hope in future work, it is possible that further improvements can be made, perhaps by seeding the qTA model with pursuit function found targets and then applying some iterative hill-climbing method to find a local minimum error of prediction.

4. Conclusion

In this study, we have shown how a pursuit function can be used in place of the qTA function for the problem of finding underlying pitch targets from measured F0. From the results we can see that the pursuit function enabled a direct means of finding the pitch targets, and more importantly, that the learned targets could be reused by the qTA model for F0 contour production. This finding provides support for our hypothesis that without fully replicating the mechanical process itself, articulatory control parameters for prosody can be learned with simpler learning process which is more conceivably developed in motor control by the human brain.

5. Acknowledgments

We would like to thank Dr. Mark Huckvale (UCL) for originally suggesting the study. The content of the paper has been greatly improved by his helpful comments.

6. References

[1] Xu, Y., "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3, pp. 220–251, 2005.

[2] Prom-on, S., Xu, Y., and Thipakorn, B., "Modeling tone and intonation in Mandarin and English as a process of target approximation," *The Journal of the Acoustical Society of America*, vol. 125, p. 405, 2009.

[3] Xu, Y. and Prom-on, S., "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[4] Xu, Y. and Prom-on, S., "PENTAtrainer1.praat. Available from: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/."

[5] Prom-on, S. and Xu, Y., "PENTAtrainer2: A hypothesis-driven prosody modeling tool," *ExLing 2012*, p. 93, 2012.

[6] Stevens, K. N. and Halle, M., "Remarks on analysis by synthesis and distinctive features," *Models for the Perception of Speech and Visual Form*, pp. 88–102, 1967.

[7] Boole, G., *A treatise on differential equations*. Macmillan & Company, 1859.

[8] Prom-on, S., Liu, F., and Xu, Y., "Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling," *The Journal of the Acoustical Society of America*, vol. 132, p. 421, 2012.

[9] Xu, Y., "Effects of tone and focus on the formation and alignment of F0 contours," *Journal of Phonetics*, vol. 27, no. 1, pp. 55–105, 1999.

[10] Xu, Y., "Fundamental frequency peak delay in Mandarin," *Phonetica*, vol. 58, no. 1-2, pp. 26–52, 2000.

The role of prosody in the encoding of evidentiality

Maria del Mar Vanrell[1], *Meghan Armstrong*[2], *Pilar Prieto*[3]

[1] Institut für Romanische Philologie, Freie Universität Berlin, Berlin, Germany
[2] Languages, Literatures, & Cultures, University of Massachusetts-Amherst, Amherst, USA
[3] Departament de Traducció i Ciències del Llenguatge, ICREA-Universitat Pompeu Fabra, Barcelona, Spain

mariadelmar.vanrell@fu-berlin.de, armstrong@spanport.umass.edu, pilar.prieto@upf.edu

Abstract

The overarching goal of this paper is to advance on the understanding of how evidential meanings are expressed in natural languages. Specifically, we aimed to investigate what type of meaning was encoded in yes-no questions through the combination of the question particle (QP) *que* 'that' and the nuclear intonational pattern L+H* L% in Majorcan Catalan yes-no questions (i.e., *Que és un llibre?*$_{\text{L+H* L\%}}$ *'QP-It's a book?')*, and to understand any temporal information that might be encoded through this construction. Several complementary research methods were used to address our question: the Discourse Completion Task, an acceptability task and a multiple-choice questionnaire. The results show that three types of information are encoded in QP *que*_L+H* L% questions: sentence modality, inference based on direct evidence and immediacy of the evidence.

Index Terms: intonation, question particles, evidentiality, typology, Catalan.

1. Introduction

All languages have some way of marking information source, though different parts of the grammar may be recruited to do so depending on the language. Lexical means for specifying source of information are probably universal. For instance, the lexical item *es veu* is used in Catalan in (1) to indicate that the speaker does not have firsthand knowledge of the proposition ([1]). In Colombian Spanish, the combination of the verb 'say'[1] and the complementizer 'that' can have different functions such as a) reported speech, b) hearsay, c) labelling or d) dubitative ([4], [5]). This is also found in other Romance language such as Galician, Romanian, Sardinian or Sicilian ([6]).

(1) *S'han quedat sense llum a Girona.* ***Es veu*** *que hi ha nevat molt.*
'The power is out in Girona. There must have been a lot of snow.'

(2) *Y eso,* ***dizque*** *es peligroso, no?*
'And that's dangerous (reported), no?'

Some languages have true evidential systems, i.e., inflectional systems with morphemes that have source-marking at the core of their semantics and are obligatory. In other languages, evidentiality specifications are "scattered" throughout the grammar ([4]). While there is some work showing the relevance of prosodic or gestural cues in the expression of epistemicity, i.e., the degree of certainty about a proposition (see [7], [8], [9] for declaratives; [10] for both declaratives and interrogatives, and [11] for interrogatives), research examining prosodic marking of evidential strategies is still scarce. The only investigation we are aware of is [12], who showed that deaccenting is used in Japanese biased polar questions when the speaker expects a positive answer based on public evidence.

Majorcan Catalan is a variety of Catalan that is characterized by exhibiting several strategies for forming polar questions based on bias ([13], [14]). Speakers may choose from different pitch accents as in (3) or may head questions with question particles (QPs) such as *que* (complementizer 'that') or *o* (conjunction 'or'), see (4).

(3) *Teniu mandarines?*$_{\text{¡H+L* L\%}}$ 'Do you have tangerines?' vs *Que hi ha gana?*$_{\text{L+H* L\%}}$ 'Are you hungry?'

(4) *Que encara no ha vengut, s'electricista?* 'The electrician hasn't arrived yet?' vs *O no estàs bo?* 'Aren't you well?'

The main goal of this article is to assess the role intonation plays in encoding information source in questions headed by the QP *que* (i.e., in constructions of the type QP *que*_L+H* L%). If it is the case that information source is being encoded through the use of QP *que*_L+H* L%, we also asked what type of evidential information the speaker might be conveying. Since evidential markers may have temporal restrictions ([15]) we also sought to investigate whether, in cases where evidential information was encoded, temporal information might also be available to listeners. Some languages, such as Sherpa ([16], [17]), encode temporal information through the evidential marker. In the case of Sherpa, *immediate* evidence is encoded. We thus left this as an additional possibility, since often times questions are asked based on information that has just become available in the discourse ([18]).

We tested these questions performing different experiments: a) Experiment 1: the Discourse Completion Task ([19], [20], [21]), b) Experiment 2: the acceptability task and c) Experiment 3: the multiple-choice questionnaire.

2. Experiment 1: the Discourse Completion Task

2.1. Methodology

For this study the Discourse Completion Task involved the creation of a set of situations which contained two evidential conditions (inferred direct evidential and hearsay) plus a non-evidential situation (see examples (5), (6) and (7) respectively). The difference between the two conditions is

[1] The form *diz que* could be a shortened version of the plural form *dicen que* 'they say that' ([2]) which "began as the collocation of a verb introducing indirect speech and its complementizer, and developed into an evidential strategy" ([3]: 16).

that while in direct evidential contexts the speaker infers the proposition from direct evidence (s/he sees it directly), in hearsay contexts inference is based on hearsay or the report of the proposition by another individual. The situations provided in the DCT were read aloud by the first author of this paper to the participants. Then, the participants were asked to respond appropriately to them. The situations were expected to elicit two biased yes-no questions (in the case of the two evidential conditions) and a neutral polar question. The questionnaire was made up of 12 situations (3 conditions x 4 pragmatic contexts). Fifteen speakers (8 females, 7 males) participated in the experiment. The final database was comprised of 180 utterances. Data were coded in Praat ([22]) for the following: a) use of lexical markers (question particles such as *o* 'or', or *que*), b) use of syntactic markers (negative yes-no questions, split questions) and c) prosodic transcription using the latest version of the Cat_ToBI system ([23]). The annotations were collected automatically in .txt format through a Praat script and then transferred to an Excel and a SPSS file for subsequent statistical analysis.

(5) Non-evidential situation: 'You have a bit of a cough and suddenly, while you're talking to a neighbor, you feel a sore throat coming on. Ask her if she has a cough drop'.

(6) Hearsay situation: 'Your cousin visits you. While you are talking, she says that since she's stopped working, she has plenty of free time. You take this to mean that she retired, so you confirm with her'.

(7) Direct evidential situation: 'It's your birthday and your friend gives you a present. You ask him if it's a book, since the package has a book-like shape'.[2]

2.2. Results

Figure 1 shows the percentage of occurrence of the dependent variable COMBINATION OF LEXICAL/SYNTACTIC STRATEGIES AND INTONATION with three levels of lexical/syntactic strategies (no QP, QP and other) and four levels of intonational strategies (¡H+L* L%, H+L* L%, L+H* L% and other)[3] for each of the possible evidential conditions (neutral, hearsay or direct evidence). The results clearly show that whereas neutral and hearsay contexts elicit a high percentage of the ¡H+L* L% intonational pattern without any QP, direct evidential contexts elicit the production of the L+H* L% intonational pattern headed by a QP (*que* or *o*). A Friedman test revealed a significant effect of evidential condition (neutral, hearsay or direct evidence) on the intonation of the y/n questions ($\chi^2(2) = 28.203$, $p < .001$) as well as on the lexical/syntactic marking ($\chi^2(10) = 24.125$, $p < .001$).

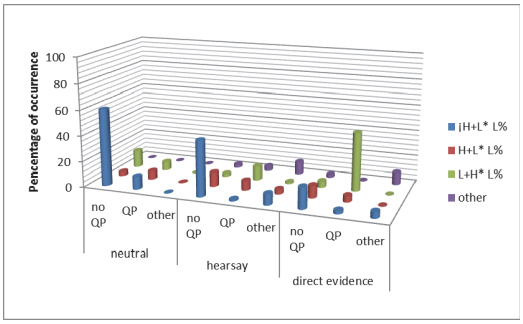


Figure 1: *Percentage of occurrence of the variable* COMBINATION OF LEXICAL/SYNTACTIC STRATEGIES AND INTONATION *for each of the possible pragmatic contexts (neutral, hearsay or direct evidence).*

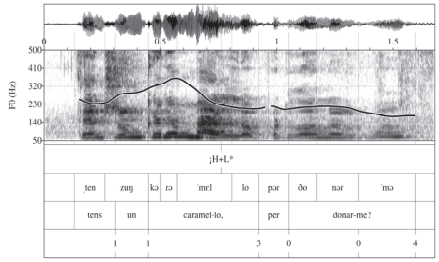


Figure 2: *Waveform and F0 contour of the y/n question Tens un caramel·lo, per donar-me? 'Do you have a candy to give me?' (the most common contour in the neutral pragmatic contexts).*

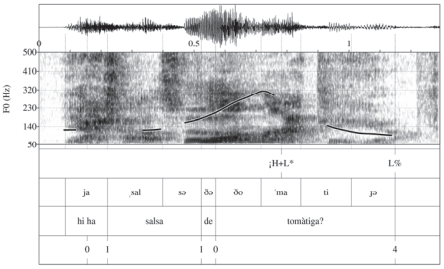
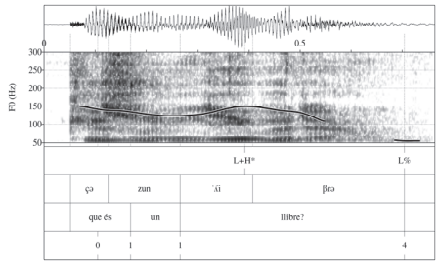


Figure 3: *Waveform and F0 contour of the y/n question Hi ha salsa de tomàtiga? 'There's tomato sauce?' (the most frequent pattern in the hearsay condition).*



[2] All the biased situations used in the experiments had a positive bias, that is, they were situations designed to elicit an affirmative response.

[3] The ¡H+L* L% nuclear configuration is realized as a fall with a leading extra high tone aligned with the end of the preaccentual syllable. The H+L* L% pattern is identical to the previous configuration but with a just high leading tone. The L+H* L% pattern is characterized by a rising movement with the peak aligned with the end of the accented syllable followed by a falling final movement.

Figure 4: *Waveform and F0 contour of the y/n question Que és un llibre? 'QP-It's a book?' (preferred strategy for direct evidential contexts).*

3. Experiment 2: the Acceptability task

3.1. Methodology

Experiment 2 consists of an acceptability task, which allows us to evaluate the degree of perceived appropriateness of target intonational patterns (in this specific case, with the combination of lexical markers) to different pragmatic contexts. Six pragmatic contexts were created with two different evidential conditions (direct evidential for half of the contexts and hearsay for the other half, see (8) and (9) below). Then three different combinations of intonation and QP, which were understood to be semantically appropriate given the pragmatic contexts, were produced, as follows: a) no QP_¡H+L* L% intonational pattern; b) QP *que* _L+H* L% intonational pattern; and c) QP *que* _¡H+L* L% intonational pattern.[4] Thus, we had a total of 36 trials: 2 contexts (direct evidential and hearsay) x 3 combinations of QP and intonational conditions (no QP_ ¡H+L* L%, QP *que*_L+H* L% and QP *que* _¡H+L* L%) x 3 pragmatic contexts + 18 fillers.

Subjects were asked to rate the acceptability of the question produced in a specific context using a 7-point Likert scale (1=totally unacceptable, 7=acceptable). Forty-six Majorcan Catalan listeners (23 female and 23 female) participated in the experiment. The experiment was run by the Survey Gizmo software (online survey software – surveygizmo.com) and was approximately 10 min long.

(8) Hearsay situation: 'Maria arrives at the fruit store and hears the owners talking about how they have the shoes from the shop owner next door, who has recently retired and closed the shop.' Target question: 'Are you going to sell shoes?'

(9) Direct evidential situation: 'Maria arrives at the fruit store and sees that the owners are putting pairs of shoes on the shelves in the entrance.' Target question: 'Are you going to sell shoes?'

3.2. Results

Table 1 shows the mean of the ACCEPTABILITY RATINGS (dependent variable) and the standard deviation of the three possible combinations of QP and intonation (no QP_¡H+L* L%, QP_¡H+L* L% and QP_L+H* L%) for each of the possible evidential conditions (direct evidential and hearsay). As we can observe in Table 1, listeners rate the combination of QP *que* and the L+H* L% intonational pattern as the most natural when it is produced in a direct evidential context (mean of 5.41). However, the same pattern of intonation and QP shows a high rate of acceptability in the hearsay context (mean = 5.12). The results of two Friedman tests showed a statistically significant effect of the combination of QP + intonation on the acceptability mean in hearsay ($\chi^2(2) = 15.510$, $p < .001$) and in direct evidence context ($\chi^2(2) = 32.690$, $p < .001$). The Wilcoxon tests using a Bonferroni correction showed that the QP *que* L+H* L% pattern differs statistically from the rest in both the direct evidential and hearsay conditions ($p < .001$)

Table 1. *Acceptability mean and* ± standard deviation for each combination of QP + intonation and pragmatic context.

	Hearsay	Direct Evidence
no QP_¡H+L* L%	4.56 ±1.69	4.51 ±1.68
QP *que*_¡H+L* L%	4.27 ±1.68	4.34 ±1.95
QP *que*_ L+H* L%	5.12 ±1.73	5.41 ±1.73

4. Experiment 3: Multiple-choice questionnaire

4.1. Methodology

For the multiple-choice questionnaire we created three pragmatically neutral contexts and three different combinations of QP and intonational conditions. Each combination of QP and intonational conditions was inserted in each of the pragmatically neutral contexts. Therefore, we had a total of 14 trials: 3 pragmatically neutral contexts x 3 QP + intonational conditions (no QP_¡H+L* L%, QP *que*_L+H* L% and QP *que*_¡H+L* L%) + 7 fillers. The subjects were asked to answer 2 multiple-choice questions related to the information source and to the time at which the evidence was available. One example of the trials was as follows: —If Maria says 'You're going to sell shoes?' to the owners of the fruit store is because Maria: a) heard that they might sell shoes, b) saw that they might sell shoes, c) heard or saw that they might sell shoes (but I'm not sure), d) I don't know. —When did Maria hear or see it? a) Just now, b) A few hours ago, c) Yesterday, d) I don't know. The 'hearing' response was an intuitive response for the hearsay condition and the 'seeing' response for the direct evidential condition. Experimental presentation was done using the Survey Gizmo software (online survey software – surveygizmo.com) for the forty Majorcan Catalan listeners (25 females, 15 male) that participated in this task and lasted approximately 10 min.

4.2. Results

Figure 6 shows the count of the dependent variable SOURCE OF INFORMATION (four levels: auditory, visual, auditory or visual, DK) for the independent variable COMBINATION OF QP AND INTONATION (three levels: no QP_¡H+L* L%, QP *que*_¡H+L* L% and QP *que*_L+H* L%). The results show that the QP *que*_L+H* % condition elicits the highest rate of visual (54 out of 360 responses) and visual/auditory (48 out of 360) responses and no "I don't know" responses. By contrast, the QP *que*_¡H+L* L% condition elicits the highest rate of auditory responses (43 out of 360) and the second highest rate of visual/auditory responses (41 out of 360). The effect of the combination of QP and intonation was not statistically significant (Friedman test: $\chi^2(2) = 4.516$, $p > .05$). Wilcoxon tests were used to follow up on this finding and they demonstrated that the effect of the QP *que*_L+H* L% significantly differed from the effect of the ¡H+L* L% intonational pattern ($T = 1019.50$, $p < .05$, $r = -22.89$) on the count of the dependent variable SOURCE OF INFORMATION.

[4] The reason why we did not fully cross the lexical and intonational conditions is because the combination of the QP *que* and the L+H* L% pattern is ungrammatical. Thus, we prevented the listeners from rating the degree of acceptability of the target itself, rather than that of the target produced in a specific pragmatic context.

Figure 7 shows the count for the dependent variable TIME OF EVIDENCE (four levels: just now, a few hours ago, yesterday, DK) for the independent variable COMBINATION OF QP AND INTONATION (three levels: no QP_¡H+L* L%, QP *que*_¡H+L* L% and QP *que*_L+H* L%). As observed in the graph, the highest rate of 'right now' responses (70/360) were obtained for the condition QP *que*_L+H* L%. The Friedman analysis showed that the combination of lexical and intonational strategies had a statistically significant effect on the dependent variable TIME OF EVIDENCE (Friedman test: $\chi^2(2) = 13.225$, $p = .001$). The subsequent Wilcoxon tests revealed statistically significant differences between the combination of QP_L+H* L% and the rest of conditions ($p < .01$).

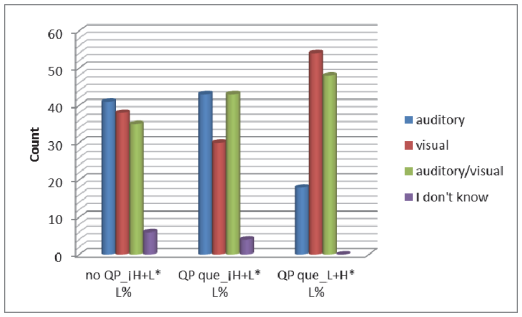


Figure 6: *Count of the dependent variable SOURCE OF INFORMATION for each combination of QP + intonation.*

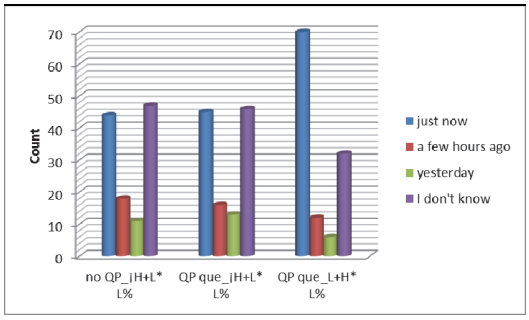


Figure 7: *Count of the dependent variable TIME OF EVIDENCE for each combination of QP + intonation.*

5. Discussion and conclusions

The aim of this study was to investigate what type of meaning was encoded through the combination of the QP *que* and the nuclear pattern L+H* L% in Majorcan Catalan y/n questions, and also to understand any related temporal information being conveyed through this construction. To this end, a set of three experimental tasks were carried out: a production experiment using the Discourse Completion Task methodology, a perception test (acceptability task) and another perception test based on a multiple-choice questionnaire. The results obtained from these different experimental approaches allow us to conclude that three types of information are encoded in QP *que*_L+H* L% questions. First, sentence modality, specifically interrogative modality, is encoded by the combination of the question particle and the rising-falling intonational pattern. Second, this particular construction informs us that the speaker has inferred the proposition through direct evidence (visual, in the specific case of this work) and now wants to ask for confirmation of this inference. Finally, listeners that hear the QP *que*_L+H* L% construction infer that the propositional content became available to the speaker just prior to the time of the utterance. Therefore, listeners are able to infer information about time reference directly through the QP *que*_L+H* L%. and do not need to rely on pragmatic context for this information. The construction we describe, then, also conveys immediate evidence like the Sherpa case mentioned above.

In this paper, only visual direct information was tested for direct evidential contexts. For instance, [15] showed that sentences with the particle *–te* in Korean signal sensory evidence, regardless of the tense that they occur with. We suspect that this is also the case for QP *que*_L+H* L% questions, in Majorcan Catalan. Unlike the *–te* particle, however, our data show a temporal restriction such that this construction can only be used just after the proposition was activated, and as a result listeners infer this temporal information when they hear the construction. Further research will be necessary to test whether the results obtained for the sense of sight are also applicable to other senses.

In sum, we conclude that the QP *que*_L+H* L% questions work as a construction, that is, a learned pairing of form and meaning ([24], [25]) which conveys sentence modality, source-marking and temporal information. In her paper on Korean evidentials, [15] notes the various strategies for encoding evidential information reported by [26] in his survey of 418 languages: verbal affix/clitic (131 languages) or particles (65 languages). [15] calls for research on languages that do not express evidentiality through distinct evidential morphemes, proposing that this will help us to understand how evidential meanings are expressed and how they make specific evidential distinctions. Our results show that without recruiting a specific part of the grammar (intonation), source-marking and temporal information are no longer available to the listener. Thus, we confirm that intonation is indeed a part of the grammar available for conveying evidential meaning, in this case working in tandem with the particle *que*. Our findings therefore have important typological implications for the study of evidential meaning.

6. Acknowledgments

A preliminary version of this paper was presented at the conference *Phonetics and Phonology in Iberia* (June 2013, Lisboa, Portugal) and the conference *Hispanic Linguistics Symposium* (October 2013, Ottawa, Canada). We are grateful to the participants at these conferences for their helpful comments and suggestions. We also thank all the subjects that participated in the different experiments. This research was funded by projects FFI2011-23829/FILO, BFU2012-31995 (awarded by the Spanish Ministry of Science and Innovation and the Spanish Ministry of Economy and Competitiveness) and 2009 SGR 701 (awarded by the Generalitat de Catalunya).

7. References

[1] González, M., "Indirect evidence in Catalan: A case study", in Ll. Payrató and Cots, J.M. [Eds.], The Pragmatics of Catalan, 146-172, Mouton de Gruyter, 2012.

[2] López-Izquierdo, M., "L'émergence de dizque comme stratégie médiative en espagnol médiéval", Cahiers d'études hispàniques médiévales, 29: 483-493, 2006.

[3] Miglio, V.G., "Online databases and language change: the case of Spanish dizque", St.Th. Gries, St. Wulff and Davies, M. [Eds.], Corpus-linguistic Applications. Current studies, new directions, 7-28, Rodopi, 2009.

[4] Aikhenvald, A.Y., Evidentiality, Oxford University Press, 2004.

[5] Travis, C., "Dizque: a Colombian evidentiality strategy", Linguistics, 44(6): 1269-1297, 2006.

[6] Cruschina, S. and Remberger, E., "Hearsay and reported speech: Evidentiality in Romance", Rivista di Grammatica Generativa, 33: 95-116, 2008.

[7] Swerts, M. and Krahmer, E., "Audiovisual prosody and feeling of knowing", Journal of Memory and Language, 53: 81-94, 2005.

[8] Dijkstra, C., Krahmer, E. and Swerts, M., "Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence, in R. Hoffmann and Mixdorff, H. [Eds.], Proceedings of the Third International Conference on Speech Prosody, 1-4, Dresden, 2006.

[9] Borràs-Comes, J., Roseano, P., Vanrell, M.M., Chen, A. and Prieto, P., "Perceiving uncertainty: facial gestures, intonation, and lexical choice", in C. Kirchhof, Z. Malisz and Wagner, P. [Eds.], *Proceedings of the 2n Conference on Gesture and Speech in Interaction*, 2011.

[10] Gravano, A., Benus, S., Hirschberg, J., German, E.S. and Ward, G., "The effect of prosody and semantic modality on the assessment of speaker certainty", in P. Barbosa, S. Madureira and Reiss, C. [Eds.], Proceedings of the fourth Speech Prosody 2008 Conference, 401-404, 2008.

[11] Vanrell, M.M., Mascaró, I., Torres-Tamarit, F. and Prieto, P., Intonation as an Encoder of Speaker Certainty: Information and Confirmation Yes-No Questions in Catalan, Language and Speech, 56(2), 163-190, 2013.

[12] Hara, Y., Kawahara, S., "The prosody of public evidence in Japanese: A rating study, in Ch. Jaehoon, A. Hogue, J. Punske, D. Tat, J. Schertz and Trueman, A. [Eds.], Proceedings of the 29th West Coast Conference on Formal Linguistics, 353-361, Cascadilla Press, 2012.

[13] Prieto, P. and Cabré, C. [Ed.], Atles interactiu de l'entonació del català. Online: http://prosodia.upf.edu/atlesentonacio/, accessed on October 2013, 2007-2012.

[14] Vanrell, M.M. and Mascaró, I., "Balear", in Prieto, P. and Cabré, T. [Ed.], L'entonació dels dialectes catalans, 75-100, Publicacions de l'Abadia de Montserrat, 2013.

[15] Lee, J., "Temporal constraints on the meaning of evidentiality", Nat Lang Semantics, 21: 1-41, 2013.

[16] Woodbury, T., "Interactions of tense and evidentiality: A study of Sherpa and English, in W. Chafe and Nichols, J., Evidentiality: The linguistic encoding of epistemology, 188-202, Ablex, 1986

[17] Kelly, B., A grammar and glossary of the Sherpa language, in Tibeto-Burman languages of Nepal: Manange and Sherpa, 197-324, The Australian National University, 2004.

[18] Büring, D. and Gunlogson, C. "Aren't positive and negative polar questions the same?", paper presented at the LSA Annual Meeting, Chicago, Ms., University of California, Los Angeles and University of California, Santa Cruz. Online: http://hdl.handle.net/1802/1432, accessed on January 2011, 2000.

[19] Blum-Kulka, S., House, J. and Kasper, G., "Investigating cross-cultural pragmatics: An introductory overview", in S. Blum-Kulka, J. House and Kasper, G. [Eds.], Cross-cultural pragmatics: Requests and apologies, 13-14, Norwood, NJ: Ablex, 1989.

[20] Billmyer, K. and Varghese, M., "Investigating instrument-based pragmatic variability: Effects of enhancing discourse completion tests", Applied Linguistics, 21(4):517-552, 2000.

[21] Félix-Brasdefer, C., "Data collection methods in speech act performance: DCTs, role plays, and verbal reports", in A. Martínez-Flor and Usó-Juan, E. [Eds.], Speech act performance: Theoretical, empirical, and methodological issues, 41-56, John Benjamins, 2010.

[22] Boersma, P. and Weenink, D., Praat: Doing phonetics by computer [Computer program], Version 5.3.55. Online: http://www.praat.org, accessed on 2 Sep 2013.

[23] Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R. and Vanrell, M.M., "Intonational phonology of Catalan and its dialectal varieties", in S. Frota and Prieto, P. [Eds.], Intonational variation in Romance, OUP, in press, to appear in 2014.

[24] Goldberg, A., Constructions, University of Chicago Press, 1995.

[25] Goldberg, A., Constructions at work, Cambridge University Press, 2006.

[26] De Haan, F., "Coding of evidentiality", in M. Haspelmath, M. Dryer, D. Gil and Comrie, B. [Eds.], The world atlas of language structures, 318-323, Oxford University Press, 2005.

Prosody in Swiss French Accents: Investigation using Analysis by Synthesis

Pierre-Edouard Honnet[1], *Alexandros Lazaridis*[1], *Jean-Philippe Goldman*[2], *Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Geneva, Switzerland
{pierre-edouard.honnet, alaza, phil.garner}@idiap.ch, jean-philippe.goldman@unige.ch

Abstract

It is very common for a language to have different dialects or accents. The different pronunciations of the same words is one of the reasons for the different accents, in the same language. Swiss French accents have similar pronunciation to standard French, but noticeable differences in prosody. In this paper we investigate the use of standard French synthetic acoustic parameters combined with Swiss French prosody in order to evaluate the importance of prosody in modelling Swiss French accents. We use speech synthesis techniques to produce standard French pronunciation with Swiss French duration and intonation. Subjective evaluation to rate the degree of Swiss accent was conducted and showed that prosody modification alone reduces perceived difference between original Swiss accented speech and standard French coupled with original duration and intonation by 29%.

Index Terms: French accents, Swiss prosody, duration, intonation, speech synthesis

1. Introduction

The perception of different regional accents in a language can result from several factors. In French, the accents can vary because of different factors according to the regions. Eventhough there are noticeable differences at the pronunciation level of some phones between "Français de Référence" (FR) defined by Morin [1] as standard and Canadian French (or Quebec French) [2], between FR and Swiss French the differences in prosody have an important role in accent discrimination.

In automatic speech recognition (ASR), regional or foreign accents and dialects of a language bring variations that decrease performance of systems. It was shown by Huang [3] that the two main sources of inter-speaker variation were gender and accent.

Kat [4] proposes two solutions to overcome the variability introduced by accent: using a wide training database which includes accented data, or building accent-specific systems which will be used according to the accent of the speech to be recognised. In the literature, there are many other attempts to tackle the accent issue (mainly for non-native accented speech) in ASR by using adaptation techniques [5, 6, 7]. More generally, ASR systems are often confronted with non-native accents, and need to counteract effects of accent component.

Conversely, in text-to-speech synthesis (TTS), producing accented speech is desirable for some applications like speech-to-speech translation (S2ST), foreign language learning and dialect synthesis. Synthesising accented speech is still a quite new and challenging area. In most cases, different accents are modelled separately using different training data. There is only limited recent work on regional accent adaptation in TTS. Astrinaki [8] proposed interpolation of TTS models using closest speakers to a chosen geographical position. In this way, the English voice has average characteristics of these speakers representing the specific regional accent. Another work by Gutierrez [9] consists of generating intermediary accent transformations between native and foreign speakers, to evaluate pronunciation of learners (computer assisted pronunciation training). This research topic can be seen as part of TTS for under-resourced languages and cross-lingual speaker adaptation for TTS.

In the SIWIS [1] project, we are aiming at personalised S2ST, i.e. being able to recognise, translate to another language and synthesise speech with an output voice sounding like the input speaker's. One of the goals is to improve prosody rendering in the TTS part of the system. The final goal is a system enabling communication between speakers of different languages. As people are generally more comfortable when speaking to someone with the same accent as theirs, synthesising speech with the same accent is more convenient for the user.

Considering the differences in prosody between French and Swiss French speakers, we believe that using robust acoustic French TTS models combined with Swiss French prosody models will allow us to synthesise Swiss accented speech. Adapting acoustic level features from a language to an accent can be done using standard speaker adaptation techniques [10, 11]. On the other hand, adapting prosody is a much more challenging issue.

In this paper, a preliminary step towards this idea is to investigate whether it is feasible to produce speech with Swiss accent using standard French acoustic models. In this direction an attempt is made to explore the importance of prosody in Swiss French accent by evaluating the degree of accent of partially synthetic speech. Our hypothesis is that using only Swiss prosody modification allows to identify Swiss accents, even with standard French pronunciation. For that we combine standard French synthetic speech and Swiss duration and intonation. Native French listeners evaluated the accent of the resulting speech.

The rest of the paper is organised as follows. We first give an overview of the differences in prosody between FR and Swiss French. Then, we describe how to exploit French TTS models to synthesise Swiss accented French. A description of the data we used is made. Experiments are detailled in the following section. Finally we conclude and propose future directions.

2. Swiss French accent

It is important to underline that the differences between FR and Swiss French are limited, since the speakers are geographically close and furthermore, linguistically, Romandie (the French

[1]Spoken Interaction with Interpretation in Switzerland, http://www.idiap.ch/project/siwis/front-page

speaking part of Switzerland) would not be distinguished from Eastern and Southeastern France according to Knecht [12].

Consequently, in this paper, the focus is only on the acoustic aspects and prosody of Swiss accent and not on lexical differences nor on the grammatical or semantic structure. It is difficult to define a description of Swiss accent in a global way, mainly because there are different granularities of accent distinction within Switzerland. Most of the French native Swiss people can distinguish the accents by canton (administrative Swiss regions). Moreover, within a canton, people are even able to distinguish accents among cities. We attempt to give a quick overview of the shared peculiarities in Swiss accents.

Some differences in pronunciation between Swiss and French speakers exist, but according to Swiss regions, these are not equally strong. Metral [13] gives more details on segmental aspects of Swiss accents.

There are some divergences – as often in the area of prosody – on the rhythm topic, i.e. Swiss speakers are known to speak slower than French. Miller [14] showed that on read speech samples, speaking rate was the same for French and Swiss (from Vaud canton) speakers, but the articulatory rate (excluding pauses) was slower for Swiss speakers. French speakers use more pauses, which decreases their speaking rate. Schwab [15] recently lead an empirical study to verify whether Swiss people indeed speak slower than French people or not. The findings showed that pause frequency and duration were not different among some French, Belgian and Swiss speakers. However, articulation rate was found to be slower for Swiss speakers.

Schwab [16] compared two Swiss regional accents with French accent, regarding penultimate accentuation, shows that Swiss speakers are more likely to accentuate penultimate syllables than French speakers. Variations were also observed among Swiss regions with different strategies in expressing prominence on these syllables.

Swiss speakers are often said to produce more variations in their intonation, however it is hard to study the phenomenon due to the variety of intonation patterns. By accentuating different syllables, they generate different intonation patterns that may sound more "lively" to French listeners.

3. Combining standard French average spectral parameters with Swiss French prosodic features

Our hypothesis is that Swiss prosody plays a major role in Swiss accent discrimination and hence that adding prosodic information will help perceiving the Swiss accent. If this is true, the degree of accent should get closer to the original Swiss accented speech when Swiss prosody is added to non accented speech.

Using analysis by synthesis, we can modify some parts of a speech signal without altering other parts. Based on this principle, we propose to generate synthetic spectrum using French TTS models (corresponding to vocal tract in the source-filter model) and combine it with original duration and intonation of Swiss French.

3.1. Acoustic and prosodic features

The acoustic features commonly used in HMM-based speech synthesis were used: mel cepstral coefficients plus energy coefficient, pitch and band aperiodicity and the first and second derivatives for each feature. The duration information was estimated with forced alignment based on these features and the existing model parameters.

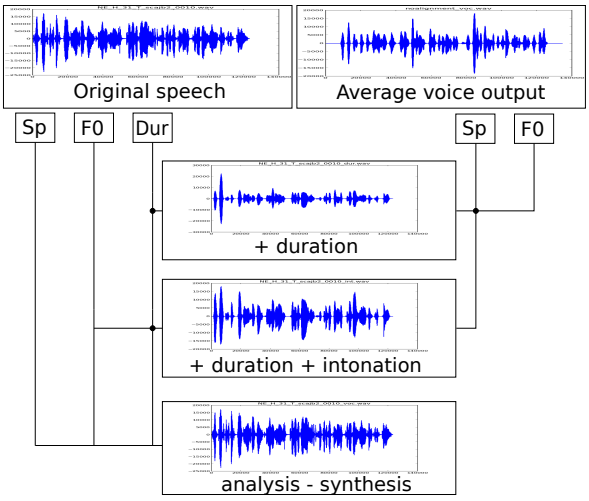


Figure 1: *Experimental setup. Features are Sp for spectrum, F0 for fundamental frequency, Dur for duration.*

3.2. Average French TTS models

We used statistical parametric speech synthesis models (hidden Markov models) to generate speech parameters that were used to "standardise" the pronunciation of all the speakers.

For that, we trained an average French HMM-based speech synthesis system using the HTS framework [17] and speaker adaptive training [18]. Average voice models are more robust than single speaker models and do not require extensive amounts of data from a single speaker.

Then, we could synthesise speech parameters and incorporate original duration and intonation information according to the prosodic features we wanted to evaluate using a vocoder.

3.3. Resynthesis

In parametric speech synthesis, the speech signal is represented by some parameters, spectrum and excitation in our case, and needs to be reconstructed to produce a waveform.

3.3.1. Vocoding

To homogenise the quality of the samples we want to produce, analysis synthesis was performed on the original waveforms, which consists of feature extraction and reconstructing the signal using a vocoder.

A completely synthetic waveform was also produced, using the French average models presented earlier. The parameters generated were finally put through the same vocoder.

3.3.2. Use of duration information

We used duration as a first prosodic feature to be added to the average synthetic speech. For that, we first extracted the duration information from the original waveforms using forced alignment: given the speech features and the corresponding transcription (full-context phonetic labels in our case), the Viterbi algorithm is used to estimate phone and state boundaries.

Using the state duration, a forced-aligned synthesis was performed, i.e. parameter generation given the known state sequence. The resulting speech was composed of synthetic parameters, but aligned in time with the original speech. The result was synthetic speech with original phoneme level duration.

3.3.3. Use of duration and intonation information

In this case, we also performed time alignment, and we replaced the synthetic intonation (log F_0) with the original one. After vocoding, the output was a speech signal composed of synthetic spectrum and aperiodicity coupled with original duration and intonation.

The reason for using both original intonation *and* duration is that it is not possible to use only original intonation, because the other parameters (spectral information) have to be aligned with the excitation part to reconstruct the speech signal.

Figure 1 gives an overview of the experimental setup described in this section.

4. Databases

The data used in this work comes from two databases: the BREF database [19] and a part of the PFC database [20] with additional content [21].

BREF is a French read-speech corpus designed for speech recognition model training and testing. The sentences to be read by the speakers were chosen from the French newspaper *Le Monde*. It consists of recordings from 120 selected speakers (55 males and 65 females), recorded in a sound-proof room. The complete database represents more than 100 hours of speech. The speech from 10 male speakers was used for this work.

The dataset taken from the PFC database consists of read speech by Swiss French speakers and French speakers from Paris. These data have been recorded in 5 cities: Paris (France) and 4 cities in 4 different Swiss cantons, i.e. Martigny (Valais), Nyon (Vaud), Neuchâtel (Neuchâtel), and Geneva (Geneva). For each location, 4 male and 4 female speakers born and raised in the city were recorded. In this work, 12 speakers were selected among the 20 male speakers available.

5. Experiments

The experiment conducted consisted of the generation of partially synthetic speech combined with Swiss French prosodic information, and was evaluated with a subjective listening test.

5.1. French TTS model

The TTS models were trained on a subset of the BREF database composed of 6857 sentences (about 12 hours of speech) from 10 male speakers. We used 39 mel cepstral coefficient with energy coefficient, log F_0, 21 band aperiodicities extracted every 5 milliseconds with STRAIGHT [22] and their first and second derivatives. 5 state left-to-right HSMMs were trained with full-context labels using version HTS 2.1 and speaker adaptive training [18].

5.2. Subjective evaluation

One common sentence was selected for the 10 Swiss and 2 French male speakers from our PFC dataset. Only male speakers were used to match with our existing TTS average French male models. This sentence was used in previous Swiss accent related studies evaluations [23, 24].

> *"La côte escarpée du mont St Pierre connaît des barrages chaque fois que les opposants de tous les bords manifestent leur colère."*

It was segmented manually and the orthographic transcriptions were corrected manually before full-context label creation (adding pauses and hesitations). Features were extracted from Swiss French data the same way as for training data. The trained TTS models were then used to estimate the duration of Swiss speech data.

A listening test was conducted in order to evaluate the degree of accent of the file generated as described in section 3.3. For this purpose, a webpage was built enabling subjects to listen to 1 completely synthetic file and 1 file with original duration, 1 file with original duration and intonation and 1 vocoded file for each of the 12 speakers, which sums up to 37 files in total. The vocoded version is perceptually very close to the original recorded speech as it is only analysis synthesis. It enables us to allow for the vocoder effect in the perception. For each file, the listeners had to give a degree of Swiss accent between 1 and 5, 1 being "no accent" and 5 "strong accent" (in the instructions, "no accent" was defined as *standard accent* and close to Paris accent). The test took approximately 10 minutes, and the listeners could listen to the files as many times as they wanted.

28 subjects did the test. Among them, there were 17 males and 11 females, 23 were French and 5 were Swiss (2 from Vaud, 1 from Valais, one from Neuchâtel and one from St Gallen). Their age was distributed in four age ranges (8 were 19-25 years old, 10 were 26-35, 7 were 36-55 and 3 were 56 - 75).

5.3. Results

5.3.1. Degree of accent

Figure 2 shows the mean and standard deviation of the three versions of the file for each speaker – the fourth version displayed in black, which is identical for each speaker, corresponds to the average voice output. The means and variances show that when adding intonation and duration the values get closer to the vocoded version than just adding duration, and modifying only duration gives closer values than the average voice output, as we expected. For the speakers with highest degree of accent (based on the vocoded version), $NE75$, $NY31$, $NY32$, $NY59$ and $NY70$ ($PA86$ has different behaviour), the means of the *intonation + duration* version is still much lower than the vocoded one. $PA86$ is a 86 year old Parisian and although he does not have a Swiss accent, his accent was perceived as strong. The average voice being based on French accent and pronunciation, he has the same pronunciation as the average voice. Adding the prosody resulted in a degree of accent close to the original, explained by both correct prosody and pronunciation.

These results are confirmed by a Wilcoxon signed rank test which was perform for each speaker among the four versions presented (3, plus the baseline average voice), as the data is ordinal [25]. In the case of average version against vocoded version, 9 out of the 12 speakers have significantly different scores; $GE24$, $GE27$ and $PA33$, corresponding to the least accented speakers, are not significantly different. In the case of the version with duration information against the vocoded version, 7 still have significantly different scores: $MA24$ and $NE31$ are not significantly different. Finally, when adding original intonation, the 5 speakers mentioned before as *very different* from the vocoded version are significantly different. $GE55$ and $PA86$ are not significantly different for these versions.

5.3.2. Prosody effect on accent perception

Table 1 shows the means of absolute differences between scores per speaker. For each speaker, a comparison was made between 2 versions of the file among the average voice output (ave), the version including duration (dur), the version including duration and intonation (int) and the vocoded version which is the reference (voc). In 8 cases out of 12, the combination of duration

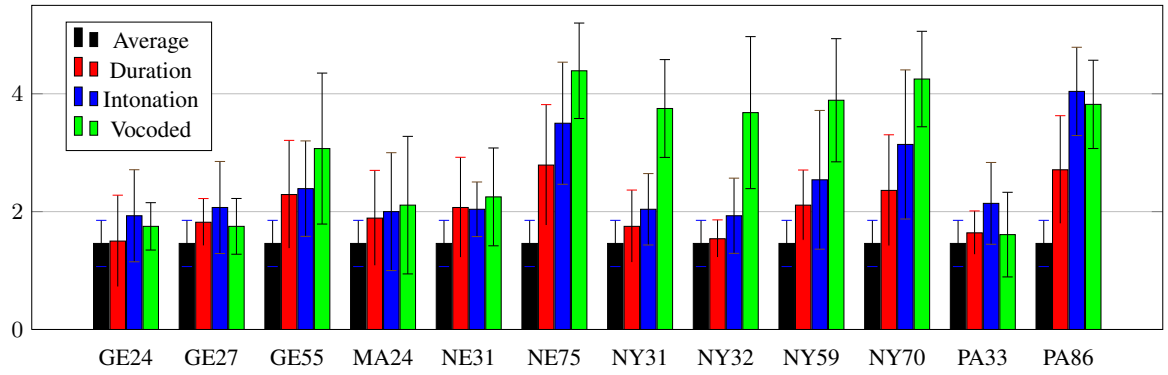


Figure 2: *Mean degree of accent for each version for the 12 speakers - Output of average TTS, with duration information, with duration and intonation, vocoded version*

Table 1: *Mean differences between configurations per speaker*

Speaker	$GE24$	$GE27$	$GE55$	$MA24$	$NE31$	$NE75$	$NY31$	$NY32$	$NY59$	$NY70$	$PA33$	$PA86$
ave_dur	1.04	0.82	0.96	0.96	0.96	0.86	0.89	0.96	0.93	0.79	1.29	0.64
ave_int	1.29	1.32	1.25	1.18	1.21	1.04	1.29	1.14	1.61	1.39	1.50	1.29
dur_int	0.61	0.79	1.07	1.14	0.75	0.68	0.61	0.96	0.89	1.11	1.00	0.93
ave_voc	1.68	1.93	1.54	1.25	1.57	1.54	1.75	1.89	1.60	1.60	2.21	1.79
dur_voc	**0.93**	1.54	**1.36**	1.64	**0.96**	1.04	**1.29**	1.57	1.39	1.39	1.64	1.57
int_voc	0.96	**1.25**	1.57	**1.07**	1.00	**1.00**	1.39	**1.46**	**1.21**	**1.14**	**1.21**	**1.21**

Table 2: *Mean differences between configurations*

	average	duration	intonation	vocoded
average	0	0.93	1.29	1.70
duration	X	0	0.88	1.36
intonation	X	X	0	1.21
vocoded	X	X	X	0

and intonation is closer to the vocoded version (values in bold). The 4 other cases give the advantage to the version including only duration information.

Table 2 gives the global absolute difference between each system. The last column gives the distance between the vocoded speech and the other versions. We can see that between the average voice output and the version with duration information we reduce the distance to the vocoded version by 20%, between the version with duration and the version including duration and intonation, the reduction is 11% and the overall improvement from average to duration and intonation version gives 29% improvement. A Wilcoxon signed rank test confirmed that the differences between score absolute differences were significant ($p - value < 0.01$ in the 3 cases).

It demonstrates that prosody plays an important role in Swiss accent perception. However, for the most accented speakers, prosody alone is not enough to obtain the same degree of accent. In these cases, adequate pronunciation is required to perceive the Swiss accent. This is backed up by the fact that accented Parisian speech can be produced with standard French pronunciation and specific prosody.

The low number of Swiss subjects did not allow us to evaluate the difference in accent perception between French and Swiss listeners, but the numbers showed similar trends for both groups.

6. Conclusions

In this paper we investigated the use of standard French pronunciation with Swiss prosody. This preliminary work was done with a view to adaptating French speech synthesis to Swiss accents. We hypothesised that Swiss accent was mainly characterised by its prosodic aspects. Analysis synthesis method and HMM-based speech synthesis were used to produce synthetic average French speech parameters which were then combined with natural speech prosodic features.

A subjective evaluation was conducted through a listening test to determine whether the degree of Swiss accent can be approached by modifying only the prosody of synthetic speech. The results showed that for 7 male speakers out of 12, using original duration and intonation with synthetic spectral parameters was not distinguished significantly from the original speech by the listeners. The difference of the scores between original speech and unaccented synthetic speech was significantly reduced by 20% by adding original duration and by 29% when adding original duration and intonation. This showed that prosody is important in the perception of Swiss accent. We also found that in the case of strong accents, prosody is not enough to model Swiss accent with standard French pronunciation. We did not use intensity information in this experiment, which would probably give further improvement.

Our future work will be to investigate the use of adaptation techniques for the pronunciation of the synthesis system to Swiss accents, and evaluate the impact of prosody in accent perception with Swiss pronunciation. The intonation contour could be investigated at a finer level to understand differences between regional accents. It would also be interesting to evaluate the accent identification rather than the degree of accent eventhough it is a difficult task for listeners.

7. Acknowledgements

This research is funded by the Swiss National Science Foundation under the SIWIS project – FNS Grant CRSII2_141903.

8. References

[1] Y. C. Morin, "Le français de référence et les normes de prononciation," *Cahiers de l'Institut de linguistique de Louvain*, vol. 26, no. 1, pp. 91–135, 2000.

[2] M.-H. Côté, "Laurentian French (québec): extra vowels, missing schwas and surprising liaison consonants," in *Phonological variation in French: illustrations from three continents.*, R. Gess, C. Lyche, and T. Meisenburg, Eds. Amsterdam: John Benjamins, 2012.

[3] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 1377–1380.

[4] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1999, pp. 221–224.

[5] S. Aalburg and H. Hoege, "Foreign-accented speaker-independent speech recognition." in *Proceedings of the International Conference on Spoken Language Processing*, 2004, pp. 1465–1468.

[6] X. He and Y. Zhao, "Fast model selection based speaker adaptation for nonnative speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 298–307, 2003.

[7] W. K. Liu and P. N. Fung, "MLLR-based accent model adaptation without accented data," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, 2000, pp. 738–741.

[8] M. Astrinaki, J. Yamagishi, S. King, N. d'Alessandro, and T. Dutoit, "Reactive accent interpolation through an interactive map application," in *Proceedings of the 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, August 2013, p. 265.

[9] R. Gutierrez-Osuna and D. Felps, "Foreign accent conversion through voice morphing," Department of Computer Science and Engineering, Texas A&M University, Tech. Rep., 2010.

[10] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[11] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[12] P. Knecht, "Le français en Suisse romande: aspects linguistiques et sociolinguistiques." in *Le français hors de France*. Paris: Valdman, A., 1979, pp. 249–258.

[13] J.-P. Métral, "Le vocalisme du français en Suisse romande. considérations phonologiques," *Cahiers Ferdinand de Saussure*, no. 31, pp. 145–176, 1977.

[14] J. Sertling Miller, "Swiss French prosody: intonation, rate, and speaking style in the Vaud canton," Ph.D. dissertation, Graduate College of the University of Illinois, Urbana-Champaign, 2007.

[15] S. Schwab and I. Racine, "Le débit lent des suisses romands: mythe ou réalité?" *Journal of French Language Studies*, pp. 281–295, 2013.

[16] S. Schwab, M. Avanzi, J.-P. Goldman, P. Montchaud, I. Racine *et al.*, "An acoustic study of penultimate accentuation in three varieties of French," in *Proceedings of Speech Prosody*, 2012.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[18] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[19] L. F. Lamel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for french," in *Proceedings of EUROSPEECH*, 1991, pp. 505–508.

[20] J. Durand, B. Laks, and C. Lyche, *Phonologie, variation et accents du français*. Paris, Hermès, 2009.

[21] M. Avanzi, "A corpus-based approach to french regional prosodic variation," in *The third Swiss Workshop on Prosody*, Geneva, 2014.

[22] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[23] I. Racine, S. Schwab, and S. Detey, "Accent(s) suisse(s) ou standard(s) suisse(s) ? Approche perceptive dans quatre régions de Suisse romande." in *La perception des accents du français hors de France.*, A. Falkert, Ed., 2013, pp. 41–59.

[24] M. Avanzi, G. Christodoulides, S. S., B. A., and G. J.-Ph., "La variation prosodique régionale et stylistique en français – analyse de neuf points d'enquête PFC," in *Journées PFC*, Paris, 2013.

[25] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," in *Proceedings of Blizzard Challenge Workshop*, 2007.

Hierarchical stress generation with Fujisaki model in expressive speech synthesis

Ya Li [1], *Jianhua Tao* [1], *Keikichi Hirose* [2], *Wei Lai*[1, 3], *Xiaoying Xu*[1, 3]

[1] National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Department of Information and Communication Engineering, University of Tokyo, Japan
[3] Department of Chinese Language and Literature, Beijing Normal University, Beijing, China
yli@nlpr.ia.ac.cn, jhtao@nlpr.ia.ac.cn, hirose@gavo.t.u-tokyo.ac.jp,
laiwei_0508@126.com, xuxiaoying2000@bnu.edu.cn

Abstract

This paper introduces a hierarchical stress generation for expressive speech synthesis. In the previous study, we proposed a novel hierarchical Mandarin stress modeling method, and the text-based stress prediction experiments demonstrates a reliable stress assignment can be obtained from textual features. However, the stress model should be further verified to be an effective and efficient prosody model in a Text-to-Speech system. In this work, Fujisaki model known as an ideal global representation of prosody is adopted to construct the pitch contours. To illustrate the effect of stress model, the Fujisaki model parameters are automatically predicted by the textural feature with and without stress information. The synthetic speech sounds more natural than that without stress modeling. The RMSE of the pitch contour and the feature importance analysis also show stress information can improve the pitch modeling. This work offers a promising method to accurate pitch modeling for Mandarin expressive speech synthesis.

Index Terms: speech synthesis, Fujisaki model, stress, hierarchical modeling, pitch accent

1. Introduction

Expressive speech synthesis has gained a lot of attention recently because people are no longer satisfied with the flat synthetic speech in navigators, automatic call-center, and information broadcasting system etc. Therefore, the accurate modeling of prosody which can express the para-linguistic information, such as emotion, attitude, intentions, speaker characteristics and making the speech sound more vivid becomes particularly important. Stress is the perceptual prominence within words or utterances, and it constitutes the peaks and valleys of the pitch contours, which is an important factor of prosody.

Although previous work on stress realization which based on concatenation system [1, 2] can produce high quality speech, the expressiveness of the synthetic speech still relies on the audio corpus they used. Recently, HMM-based speech synthesis (HTS) draws growing attention for its flexibility in expressive speech synthesis. HTS-based stress generation can be categorized as direct modeling [3] and indirect modeling [4, 6], which mainly refers to prosodic parameter transformation.

The direct modeling is introducing the stress related question into the question set which is used in the HMM models clustering in HTS. However, the speech generated by this approach cannot convey stress clearly in HTS due to the weakness of emphasis/stress cues and statistical averaging effect of HTS [4]. Badino *et.al,* argue that more sophisticate context features should be designed to obtain a clear emphasis/prominence realization [5].

Regarding the indirect modeling, Yamafishi *et. al.*, [6] use speaking style interpolation and adaptation for HMM-based expressive speech synthesis. Maximum Likelihood linear Regression (MLLR) model is adopted in the style adaptation. Yu, *et.al.*, [4] utilize two-pass decision tree model and factorized decision tree model to extract word-level emphasis patterns from natural English speech, and then embed the emphasis model in the HTS framework. Although the speech generated by prosodic parameter transformation can convey stress effectively, it happened sometimes that a few syllables turned out too strong compared with the adjacent syllables, which makes the whole utterance sound unnatural. This indicates that the tradeoff between prominence and naturalness is hard to balance.

The ultimate goal of our work is synthesizing human-like expressive speech with stress. In the previous study, we proposed a novel hierarchical Mandarin stress modeling method [7]. The top level of this model emphasizes stressed syllables, while the bottom level focuses on unstressed syllables for the first time due to its importance in both naturalness and expressiveness of synthetic speech. The text-based stress prediction experiment demonstrates we can get a reliable stress assignment from textual features. However, the stress model should be further verified to be an effective and efficient prosody model in a Text-to-Speech system.

Therefore, generation process model of fundamental frequency contours known as Fujisaki model is adopted to generate pitch contours in this work. The reason of adopting Fujisaki model lies in two aspects. First, Fujisaki model is also a superpositional quantitative model for representing F0 contour of speech, which is perfect match the two-level hierarchical stress model we proposed, and thus we can directly control the hierarchical stress. Although some hierarchical pitch modeling methods [8-10] have already been proposed recently, the hierarchy is implicit modeled. Second, Fujisaki model is not only a parametric F0 contour stylization, but also has physiological and physical basis [11], and can well represent the long-term features than the commonly used frame-by-frame analysis method.

Some work on stress/prominence/focus realization has already been conducted with Fujisaki model. Kiriyama *et al.*, [12] and Chen *et. al.*, [13] implement a rule-base focus control with Fujisaki model. Ochi, *et.al.*,[14] also use Fujisaki model to control focus, but they predict the differences of the Fujisaki model commands between with and without focus utterances. The method is confirmed by the experiments. But, there are some constrains and limitations in the prosodic difference modeling.

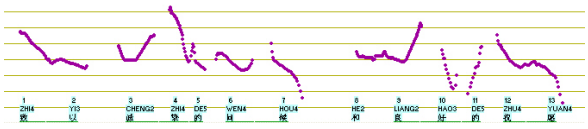
In this work, we want to testify whether the hierarchical stress model could improve the pitch modeling and how much improvement it will achieved. Therefore, we semi-automatically built a 500 sentence corpus with Fujisaki model parameters annotation as the first step. Then we constructed two Fujisaki model parameter prediction models using decision trees, among which, one utilizes the common features used in prosodic model parameter prediction, and the other introduces the stress information compared with the first model. Afterwards, the continuous pitch contours are generated by the two models. Listening test, objective experiment and features importance analysis are carried out. The results show stress model can improve the pitch modeling.

The rest of the paper is organized as follows. Section 2 introduces the hierarchical modeling method, including the hierarchical stress modeling and the superpositional modeling of F0 contour by Fujisaki model. Section 3 shows the details of text-based Fujisaki model parameter prediction with and without the hierarchical stress features. Experimental results and discussions are given in Section 4, and followed by the conclusion and future research in Section 5.

2. Hierarchical modeling

2.1. Hierarchical Mandarin stress model

Mandarin stress can be categorized as sentence stress and word stress from the range of their influence. Considering the importance of unstressed syllable in the naturalness and intelligibility of speech, a novel two-level Mandarin stress modeling method was proposed, in which, word level unstressed syllable investigation are emphasized for the first time, and in sentence level stressed syllables are studied as traditional methods do [7]. Fig. 1 shows a hierarchical stress assignment of a speech sample. First, the sentence stress is assigned onto words, and then each syllable's stress level is assigned within the word. The sentential stressed (denote as *3* in sentence stress) and unstressed syllable (denote as *1* in word stress) are the research focus of this model.



(a) Natural pitch contours

	zhi4yi3	cheng2zhi4de5	wen4hou4	he2	liang2hao3de5	zhu4yuan4
Sentence Stress	2	3	2	1	2	2
Word Stress	3 2	3 2 1	2 3	2	3 2 1	3 2

(b) Hierarchical stress assignment of speech sample (a)

Figure 1: Hierarchical Mandarin stress modeling.

2.2. Hierarchical pitch modeling with Fujisaki model

Fujisaki model is a command-response model that describe F0 contour as the superposition of the outputs of phrase and tone(/accent) commands [11].

Unlike most non-tone languages, which only have positive tone commands, Mandarin has both positive and negative tone commands. The tone command configuration for Mandarin can be found in [11]. The model (for tonal language) can be formulated by the following equations:

$$\begin{aligned}\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} Ap_i Gp(t-T_{0i}) \\ + \sum_{j=1}^{J}\{A_{1j}[Ga(t-T_{1j}) - Ga(t-T_{2j})] \\ + A_{2j}[Ga(t-T_{2j}) - Ga(t-T_{3j})]\}\end{aligned} \quad (1)$$

$$Gp(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & for\ t \geq 0 \\ 0, & for\ t < 0 \end{cases} \quad (2)$$

$$Ga(t) = \begin{cases} \min[1-(1+\beta t)\exp(-\beta t), \gamma], & for\ t \geq 0 \\ 0, & for\ t < 0 \end{cases} \quad (3)$$

where F_b , $G_p(t)$ and $G_a(t)$ are base fundamental frequency level, phrase commands and tone commands respectively, the detailed symbolic representations can be found in [11].

2.3. Integrating the two hierarchical models

Figure 2 illustrates the integration and relationship between Fujisaki model and the hierarchical stress model. The sentence level stress is corresponding to the phrase command in Fujisaki model, and the word stress is corresponding to the tone command. Inspired by this relationship, we utilize the sentence stress information to improve the phrase command prediction, and the word stress feature is introduced in the tone command prediction. We expect that the Fujisaki model commands can be estimated more accurately through this manner. Then we can evaluate the prosody of the synthetic speech to verify the hierarchical stress generation method.

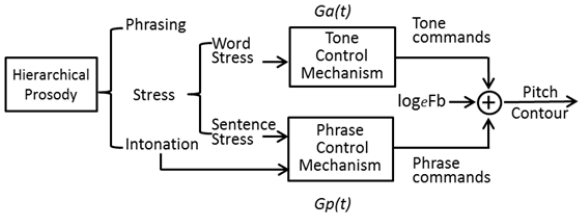


Figure 2: Integration and relationship between hierarchical stress model and Fujisaki model.

3. Fujisaki model parameter prediction

3.1. Corpus construction

In this work, we built a 500 sentences corpus, selected from the stress annotated corpus introduced in [7]. The Fujisaki model parameters are extracted automatically at first and then manually corrected. Fig. 3 is a sample of the Fujisaki model parameter labeling result through the FujiParaEditor [15].

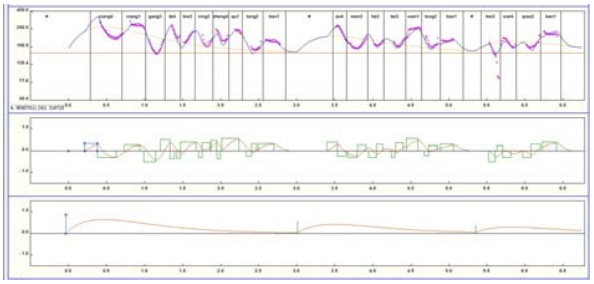


Figure 3: *Fujisaki model parameter annotation for Mandarin.*

The assignment of the phrase command usually coincides with the intonation phrase, but this is not a strict rule in the corpus labeling. The value of α, β, γ are the same as [11]. Ten percent of the data is reserved for testing.

3.2. Feature extraction

To verify the stress generation method, we designed two groups of textural features which are used in Fujisaki model parameter prediction. The first group includes the baseline features without stress information, and the second group includes the stress information as well as the baseline features. The sliding window is used for feature extraction in these two groups and the window size is five. For tone command prediction, the feature extraction unit is syllable, and for phrase command prediction, the unit is word.

As for the baseline feature set, different features are selected for the tone command and phrase command predictions. For the tone command prediction, the features are tone (indicate as *t*, hereinafter), syllable boundary (*bk*), the distance from the current syllable to the beginning/end of the sentence (*db/de*), the length of the word (*len*) to which the syllable belongs and its position in the word (*position*).

For the phrase command prediction, the long range context features are selected, namely, the Part-of-Speech of the word (*p*), the length of the word (*l*), the distance from the current word to the beginning/end of the sentence (*db/de*), the phrase length count by syllable and word respectively (*dis2sbsyl*, *dis2sbw*), and the index of the word in the intonation phrase to which it belongs (*posinphrase*).

In the second feature group, whether the syllable is unstressed or not is introduced in the tone command prediction and whether the word is sentential stressed is introduced into phrase command prediction. The windows for the stress related feature selection is also five. To verify the stress generation method, the hierarchical stress information is extracted from the annotated corpus rather than the automatic prediction from textual features.

3.3. Decision trees

According to Eq(1)-Eq(3), three parameters, with or without a phrase command, phrase command amplitude, and the command starting time, should be predicted from textual features. For tone command, five parameters, namely, two amplitudes and three timing values are selected as the targets. It should be noted that for tone 1, tone 3 and tone 5 syllables, one amplitude and one time value are set to zero. All the timing values are relative time, and are the offsets from the syllable(/word) starts time.

In the Fujisaki model parameters prediction, with or without a phrase command for each word is a binary classification problem, while the other commands, such as amplitude and command starting time, are continuous. J48 is adopted in the first task, and M5P is utilized for others which are both decision trees and implemented in Weka [16]. M5P combines a conventional decision tree with the possibility of linear regression functions at the nodes. In the preliminary experiments, we have tried several statistical models, M5P is slightly better than others in this study, such as C4.5 tree, J48 and linear regression.

4. Experiments and discussion

4.1. Pitch contours generation

The pitch contours are then generated by superimposing the phrase command and tone command responses. To simplify the work, the syllable duration and the base fundamental frequency level, F_b, are assumed to be the same with those in the training corpus. Then all the predicted time values can be converted to absolute values, and constitute a time sequence. Finally the synthetic speech can be generated by PSOLA algorithm which is implemented in Praat.

4.2. Experimental results

The average classification result for with or without a phrase command at a word's boundary is 73.69%. By introducing the hierarchical stress information, the average classification accuracy increases to 78.4%.

Table 1 shows the prediction results for the rest Fujisaki model parameters with continuous values, which only utilize the textual features. The first two rows are the results for phrase command predictions, including the phrase command amplitude A_{p1} and starting time T_0. The rest rows represent the tone commands prediction. It shows that the stress information can enhance the Fujisaki model parameter prediction, however, the improvement is small.

Table 1. *Text-based phrase command prediction using M5P decision tree.*

Model	Baseline (without stress)		With hierarchical stress feature	
	Corr.	RMSE	Corr.	RMSE
A_{p1}	0.85	0.18	0.85	0.17
T_0	0.81	106 ms	0.81	105 ms
A_{t1}	0.91	0.19	0.92	0.19
A_{t2}	0.94	0.11	0.94	0.11
T_1	0.58	43 ms	0.58	42 ms
T_2	0.80	34 ms	0.81	32 ms
T_3	0.57	54 ms	0.60	52 ms

Table 2. *Average RMSEs of utterance pitch contour predicted by models with stress and without stress.*

Experiment	RMSE (Hz)
without stress	46
with hierarchical stress	45

To further check the hierarchal stress generation, we align all the Fujisaki model parameters obtained by the automatic prediction, and combine the base frequency (F_b) to generate a continuous pitch contour. Table 2 shows the RMSEs between the natural speech and the F0 predicted by models with and without hierarchical stress. It also indicates that with stress information, pitch contour can be more accurately modeled. Fig. 4 shows pitch contours of two synthetic utterances generated by Fujisaki model. In this figure, the 8th syllable "*shang4*" is unstressed; the pitch is lower in the proposed pitch contours. On contrary, the 9th and 10th syllables"*fei1 chang2*" (In fact, they constitute a word, which means *very*.) are stressed, and "*fei1*" gets the final stress assignment through the

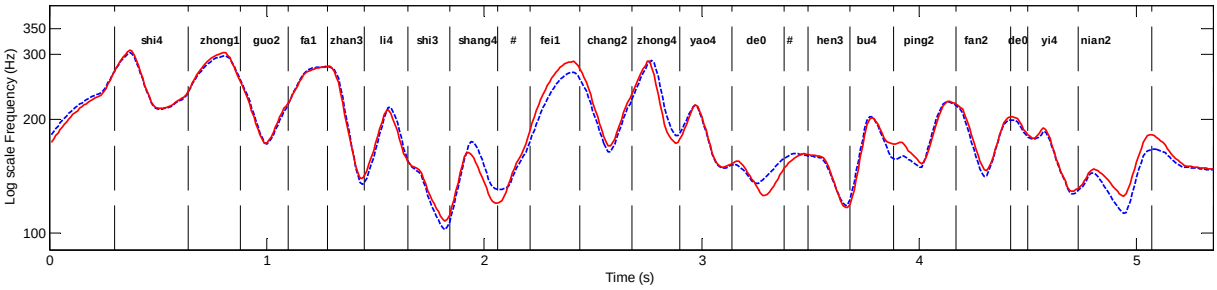


Figure 4: *Pitch contours generated by Fujisaki model with stress and without stress (line : with stress, dot: without stress).*

hierarchical mechanism, therefore, its pitch is higher in the proposed pitch contour. The informal listening test also shows the synthetic speech with stress model which using PSOLA algorithm sounds more natural. The over-salience syllables are hardly found in the synthetic speech. However, it should note that not all Mandarin tones are fully realized in connected speech, thus the variation of tone commands is complex and difficult to model using a decision tree which combines linear regression model in the nodes. Because of the tone command prediction error, some syllables sound unnatural, which makes the whole utterance sounds weird if there are too much tone command prediction errors. In such cases, the naturalness of the synthetic speech is worse than that generated by HTS. Nevertheless, the prominence and naturalness is well balanced in this method because the nature of command-response mechanism.

To evaluate the effect caused by hierarchical stress information in Fujisaki model parameter prediction, we also conducted a feature importance analysis by correlation feature selection [17]. Table 3 shows the feature selection results for each Fujisaki model parameter prediction. In this table, symbols, such as *p, l, t, bk,* represent the features introduced in Subsection 3.2. *s* represents the stress information. The number at the end of each feature denotes the offset in feature extraction. For example, *p_1* represents the Part-of-Speech of the previous word, and *s1* represents the stress information of the next syllable/word. Table 3 clearly illustrates stress information is indeed important in almost every parameter prediction.

Table 3. Feature importance in each Fujisaki model parameter prediction.(Y/N means with or without phrase command)

Prediction target	Feature importance (descending order)
Y/N	l_1, dis2sbw
A_{p1}	p_1, p0, l_2, l_1,dis2sbw,s_2,s_1,s0
T_0	p_2, p_1, p0, l_2, l_1, dis2sbw, s_2, s_1, s0
A_{t1}	t0, s0
A_{t2}	t0, bk0, db
T_1	t_2, t_1, t0, t1, bk2, bk0, s0
T_1	t0, s_1, s0, s1
T_3	bk2, s_2, s0

5. Conclusions

This paper introduces an attempt in Mandarin expressive speech synthesis by manipulating the stress generation with pitch contour generation process model (Fujisaki model). The Fujisaki model parameters are automatically predicted by the textural features with and without stress information. The hierarchical Mandarin stress modeling method is adopted to control phrase command and tone command correspondingly. And then the continuous pitch contours are generated and further evaluated. The experiments show hierarchical stress information can improve the pitch modeling both in global and local range. The advantage of the proposed method is it can make a good balance between prominence and naturalness compared with the previous direct and indirect stress/prominence modeling in HTS. This work offers a promising method to accurate pitch modeling for Mandarin expressive speech synthesis.

However, the accuracy of automatic Mandarin Fujisaki model parameter perdition needs further improvement, especially for the tone command. As reviewer suggests, a more comprehensive consideration of tones and how they interact with default focus, contrastive stress and different types/degree of stress in a sentence would be necessary. We believe that once this problem is solved, the MOS of synthetic speech can be greatly improved. Moreover the syllable duration variation in stress generation should be taken into consideration too. We will put more effort in these two fields in the future.

6. Acknowledgements

The author would like to thank anonymous reviewers for their valuable comments.

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, and No.61375027), and partly supported the Major Program for the National Social Science Fund of China (13&ZD189) and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

7. References

[1] W. Zhu, "A Chinese Speech Synthesis System with Capability of Accent Realizing," Journal of Chinese Information Processing, vol. 21, pp. 122-128, 2007.

[2] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, et al., "Modelling prominence and emphasis improves unit-selection synthesis," presented at the INTERSPEECH, Antwerp, Belgium, 2007.

[3] Y. Wu, "Research on HMM-based speech synthesis," Doctoral dissertation, University of Science and Technology of China, 2006.

[4] K. Yu, F. Mairesse, and S. Young, "Word-level emphasis modelling in HMM-based speech synthesis," in Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010, pp. 4238-4241.
[5] L. Badino, J. S. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in HMM based speech synthesis," in INTERSPEECH, Brighton, UK, 2009, pp. 520-523.
[6] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis–towards TTS with arbitrary speaking styles and emotions," in Proc. of Special Workshop in Maui (SWIM), 2004.
[7] Y. Li;, J. Tao;, and X. Xu, "Hierarchical Stress Modeling in Mandarin Text-to-Speech," in INTERSPEECH, 2011, pp. 2013-2016.
[8] Y. Qian, H. Liang, and F. K. Soong, "Generating natural F0 trajectory with additive trees," in INTERSPEECH, 2008, pp. 2126-2129.
[9] H. Zen and N. Braunschweiler, "Context-Dependent Additive log F0 Model for HMM-Based Speech Synthesis," in INTERSPEECH 2009, 2009, pp. 2091-2094.
[10] M. Lei, Y. Wu, F. K. Soong, Z. H. Ling, and L. Dai, "A Hierarchical F0 Modeling Method for HMM-Based Speech Synthesis," in INTERSPEECH, Makuhari, Chiba, Japan, 2010, pp. 2170-2173.
[11] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," in Speech Prosody 2004, International Conference, 2004.
[12] S. Kiriyama, K. Hirose, and N. Minematsu, "Prosodic focus control in reply speech generation for a spoken dialogue system of information retrieval," in Speech Synthesis, 2002. Proceedings of 2002 IEEE Workshop on, 2002, pp. 139-142.
[13] G. P. Chen, Y. Hu, R. H. Wang, and H. Mixdorff, "Quantitative analysis and synthesis of focus in Mandarin," in International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, 2004, pp. 25-28.
[14] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," in Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009, pp. 4257-4260.
[15] H. Mixdorff, H. Fujisaki, G. P. Chen, and Y. Hu, "Towards the automatic extraction of Fujisaki model parameters for Mandarin," in Eighth European Conference on Speech Communication and Technology, 2003.
[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, 2009.
[17] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, 1999.

Too cautious to vary more? A comparison of pitch variation in native and non-native productions of French and German speakers

Frank Zimmerer, Jeanin Jügler, Bistra Andreeva, Bernd Möbius, Jürgen Trouvain

Department of Computational Linguistics and Phonetics, Saarland University, Germany
{zimmerer|juegler|andreeva|moebius|trouvain}@coli.uni-saarland.de

Abstract

This article presents preliminary results indicating that speakers have a different pitch range when they speak a foreign language compared to the pitch variation that occurs when they speak their native language. To this end, a learner corpus with French and German speakers was analyzed. Results suggest that speakers indeed produce a smaller pitch range in the respective L2. This is true for both groups of native speakers. A possible explanation for this finding is that speakers are less confident in their productions, therefore, they concentrate more on segments and words and subsequently refrain from realizing pitch range more native-like. For language teaching, the results suggest that learners should be trained extensively on the more pronounced use of pitch in the foreign language.

Index Terms: pitch variation, L1, L2, language learning

1. Introduction

When learning a foreign language, especially as adults, it is extremely hard to reach native-like skills in phonetics and phonology of this language. One of the reasons for this hardship is that the phonetic and phonological knowledge of the native language (L1) can interfere with the phonetic and phonological system of the foreign language (L2) (e.g., among many more, [6], [14], [18]). For instance, in German, voiceless plosives are produced with a long Voice Onset Time (VOT), whereas in French, VOT for voiceless plosives is rather short. When producing stops in L2, German speakers usually do not adapt their production, and their voiceless French stops do not sound like a native French production would.

However, apart from segmental differences that are hard to be learned perfectly, a foreign accent might also occur due to prosodic interference from the native language in L2 (e.g. [2], [26], [27]). Languages have been shown to differ with respect to the pitch range they use, their exact pitch contours and the exact placement of pitch changes (e.g. [3], [4], [5], [7], [10], [11], [13], [17], [19], [21], [22]). A study by Mennen and colleagues [22] suggests, for example, that there are differences both in level as well as range for English and German speakers in the respective L1. This finding is supported by data presented by Andreeva and colleagues [3] who found differences in level and range for Bulgarian, English, German, and Polish speakers. Keating and Kuo [17] found several differences between English and Mandarin speakers in pitch level and range, also depending, for instance, on the task the speakers were engaged in. Some of the differences that occur between languages can also be partly explained by (socio-) cultural factors (e.g. [17], [30]).

The difficulty to reach native-like performance in the prosodic realization of an L2 is arguably aggravated by the fact that when speaking a foreign language pitch variation is apparently compressed compared to the pitch range that is standard for native speakers (e.g. [9], [10], [16], [20], [29]), which can result in a foreign accent. Furthermore, the lack of correct pitch variation can lead to be perceived as speaking in a monotonous way (e.g. [15], [16]). One possible explanation for this compression is that L2 learners are less confident about speaking the foreign language, or that they focus on getting the segmental pronunciation and the placement of stress correctly before expanding the pitch range as native speakers do. For instance, Mennen [20] showed that Dutch (L1) speakers of Modern Greek failed to produce the same pitch range as native Greek speakers. Also, in a study investigating pitch range of Finnish (L1) speakers of Russian (L2), Ullakonoja [29] found smaller pitch ranges in the L2 production for these speakers compared to native Russian speakers. Furthermore, the results show that the pitch range was also different compared to the Finish L1 productions, and that extended stays in Russia led to a larger pitch range, implying the learnability of this intonation feature. Similarly, Busà and colleagues [10] found (non significant) long-term distributional (LTD) differences in the production of Italian speakers of English compared to native English productions. In a study of Arabic native speakers a comparison of their native productions with productions in English (L2) did not show a language effect [1]. In this study, however, the speaker group was quite advanced and living in the L2 environment. Thus, the results maybe due to the higher L2 proficiency level of the speakers.

Other studies have assumed a compressed pitch range *a priori*, and focused on the improvement of pitch range suggesting that the compression of pitch range can be overcome with enhanced training methods (e.g. [7], [15], [16]).

This short overview suggests a clear trend for a compressed pitch range in L2 speech production. Moreover, it seems that training helps to decrease the degree of pitch range compression. However, it is not clear whether the finding of reduced pitch range in the L2 is an universal tendency or whether pitch range compression is dependent on the language pair under investigation. This paper contributes to the research on L2 pitch range by investigating whether learners of French and German compress their pitch range when speaking their non-native language (German and French), compared to their L1 productions, to find out whether there is a general trend for pitch compression in L2 production, irrespective of L1. The construction of the corpus allows for a direct comparison for each speaker in each of the languages, because the same speakers were recorded in both languages (see also section 2.2).

A second question that will be touched upon is the extent to which language learners are able to learn to suppress pitch compression, that is, we investigate whether advanced learners compress pitch range less than beginners. The results reported in this paper are preliminary: the number of speakers (7 per native language) is rather small, the number of male and

female speakers is not evenly distributed across the two native languages and the number of advanced learners and beginners is also not equal. Even so, stable effects found in the present study will provide a strong motivation for an investigation based on a larger number of speakers.

2. Methods and materials

Different LTD measures have been used to quantify pitch range differences in the past (e.g. [15], [17], [22]). The quantifications that have been analyzed range from linguistically defined tonal structures to different measures of F_0 (e.g. in Hz or semitones) or a combination thereof. Furthermore, there is the overall level of pitch (usually calculated as mean value over time) and the range of variation within a given speech sample. In this paper we concentrate on the latter. It is outside the scope of this paper to discuss the (dis)advantages of one measure over another. We focus on the so-called Pitch Dynamism Quotient (PDQ), which allows for a normalized evaluation of pitch variation [15] (see also [25]) where it is called frequency modulation factor – and below, section 2.3.).

2.1. General corpus description

A bilingual learner corpus served as the basis for the analysis reported here. The corpus was created with French (L1) learners of German (L2) and German (L1) learners of French (L2) at the LORIA institute in Nancy, France and the institute of phonetics at Saarland University in Saarbrücken, Germany [28].

For the corpus, 7 speakers of each language were recorded. They were recorded both in the respective L2 and in their native language. This design allows for a within-subject comparison of the productions in the two languages. Furthermore, recording settings and the text data are comparable across the languages (as L1 or L2) and identical within a language (i.e. French texts read by both French and German participants and *vice versa*). Therefore, comparisons between individual speakers with different L1 can be drawn as well as differences between a group of speakers that differ with respect to what language they speak as native language. In each group (French and German native speakers) there were 5 beginners (A1-A2 level according to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)). Additionally, 2 advanced learners (B2-C1 Level) of the respective language were recorded. In the group of French native speakers, 6 male speakers and 1 female speaker participated (15-22 years, M: 20.1 years, SD: 2.4 years), whereas for the German native speaker group, 5 female speakers and 2 male speakers were recorded (15-26 years old, M: 20.6 years, SD: 4.2 years). The 3 teenage speakers (all 15 years of age, all had completed the change of the voice, 1 French and 2 German highschool students) being part of the corpus were all male. The data of all 14 speakers were used for the analysis in this study.

Recordings were made in quiet office rooms with head-mounted microphones, which were amplified and digitized (16khz, 16 bit) in a M-AUDIO Fast Track USB device. Recordings were saved on Windows Laptop computers with a custom-made software that was developed at LORIA ("Corpusrecorder", [12]). Each sentence (and each story) was saved as a separate audiofile.

2.2. Text materials for the analysis

The corpus data analyzed for this paper consisted of read sentences and read short stories [28]. Both groups of speakers read a set of 25 sentences (*sentence-condition*) in their respective L2, before reading two short stories (an advertising text about ecological economic development and the story of the three little pigs), also in the respective L2 (*story-condition*). Then, the same tasks (reading a set of 25 sentences and a translated version of the two stories, slightly longer in the German than in French version) were recorded in the respective L1 of the participants. The sentences were different in the two languages, but similar in content and length. Due to a technical defect, only 24 sentences were recorded for 1 native speaker of French in the L2. Otherwise, all audiofiles could be used for further analysis.

2.3. Pitch analysis

Pitch measurements were done in three steps. First, F_0 was extracted automatically by means of the ESPS algorithm ("get_f0" [24]) from all files. For male speakers time steps of 0.01 seconds were used, whereas the computation of F_0 for female speakers was done in time steps of 0.005 seconds.

Then, manual correction was performed with PRAAT [8] to exclude cases of octave jumps, wrong measurements (e.g. when the algorithm found voicing in silent intervals or mistakes due to creaky voice, or other artifacts produced by the algorithm that were not based on changes in F_0, but possibly distorting the results). Other microprosodic variation (e.g. due to stop realization) was retained. In a final step, the first two F_0 values after an unvoiced segment and the last F_0 value before an unvoiced segment were excluded. This was done because the vocal folds need some time to achieve their intended vibration rate (i.e. start or stop of vibration) and this kind of irregular vibration could also exaggerate the F_0 variation. The resulting data was analyzed by means of the JMP software [23].

The number of female and male speakers differed strongly between the two groups of native speakers. This was one of the main reasons to investigate pitch variation with the PDQ as defined as the standard deviation divided by the F_0 mean [15]. This measure normalizes F_0 variation data and allowed us to concentrate on language differences while minimizing group differences at the same time. The lower the PDQ, the smaller is the variation.

Note, however, that due to the uneven distribution of male and female speakers across languages, comparisons between the native speaker groups (e.g. a claim like "French is generally more variable than German") cannot be made despite the use of PDQ (and thus, the normalization). First of all, the preliminary analysis of seven speakers per native language is not enough for general, far reaching claims concerning two languages. But more importantly, pitch *range* has been found to differ between male and female speakers (e.g. [25]). Therefore, possible differences in the results between the speakers of the two languages cannot be attributed with certainty to language differences, the source could as well be a gender difference. However, on a general level, differences between German and French pitch range may be expected on the basis of prior research (e.g. [5], [19]). Koreman and colleagues [19] presented evidence that both French and German speakers use F_0 to indicate accentuation, but they do so differently.

Therefore, what is possible, and this is done in this paper, is to investigate the pitch variation of the German group in the native and non-native language, as well as the same comparison for the French native speakers. The results will be reported in the next sections.

3. Results

Mean F_0 and the standard deviations were calculated for every audiofile (henceforth "item"). Subsequently, the PDQ was calculated for these items. These PDQ values were the basis for all subsequent analyses.

Overall mean PDQ for French speakers in their L1 is 0.134 when reading sentences, and 0.142 when reading the stories. In their L2, French speakers produce a mean PDQ of 0.119 for the sentences and 0.13 for the stories. German speakers have a mean PDQ in their L1 of 0.142 for the sentences and 0.146 for the stories. In their French productions, the mean PDQ drops to 0.118 for the sentences and 0.133 for the stories. Figure 1 depicts the overall mean PDQ results for the two speaker groups in the respective task languages.

Subsequently, the PDQ values were entered into a linear mixed model with *PDQ* as dependent factor, *speaker* and *item* as random factors and *native language* (French/German), *task language* (French/German), and *task* (sentence/story) as independent factors, as well as all their possible interactions.

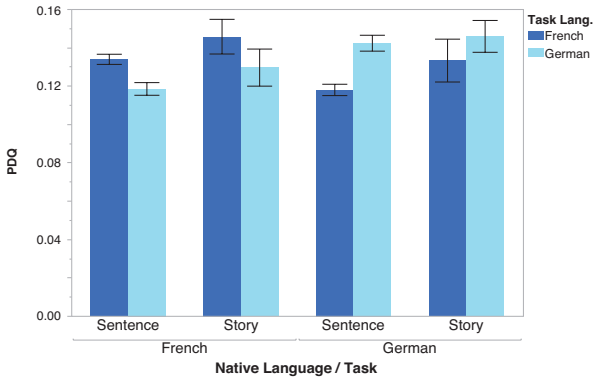


Figure 1: Mean PDQ for French and German speakers depending on task language and task.

The results of the statistical analysis indicate that there was a main effect of *task* ($F(1,735)=5.52$ $p<0.05$). Overall, speakers produced higher pitch variation when reading the stories compared to their sentence productions (mean PDQ for stories: 0.139; mean PDQ for sentences: 0.128). Also, the interaction of *native language* and *task language* was significant ($F(1,735)=14.85$, $p<0.0001$). A post-hoc planned comparison showed that for both language groups, the native productions were significantly higher in PDQ than the non-native productions (French L1: $t(2)=-2.51$, $p<0.05$; German L1: $t(2)=-2.94$, $p<0.01$).

No other factors or interactions reached significance. This first analysis suggests that irrespective of the L1, F_0 variation is compressed in L2 productions. Also, the F_0 range depends on the task in which participants were engaged. Reading stories leads to a greater variation in pitch than reading short sentences, irrespective of native language or task language.

A second model was calculated to tap into the effect of proficiency. To this end, the first model was extended by the factor *proficiency* (Beginner/Advanced) and all its interactions. The results indicate that proficiency level was not a significant factor, nor were any of its interactions. Again, *task* ($F(1,729)=5.66$, $p<0.05$) and the interaction of *native language* and *task language* ($F(1,729)=9.11$, $p<0.01$) were significant. Figure 2 shows the PDQ for the different fluency levels. As can be seen, the overall results remain very similar. However, for the French advanced learners of German, the *story condition* shows a small trend into a different direction. Here, the PDQ for the stories in German is slightly higher than the PDQ for the stories read in French. Note that there were only two advanced speakers in both native speaker groups that read the two stories.

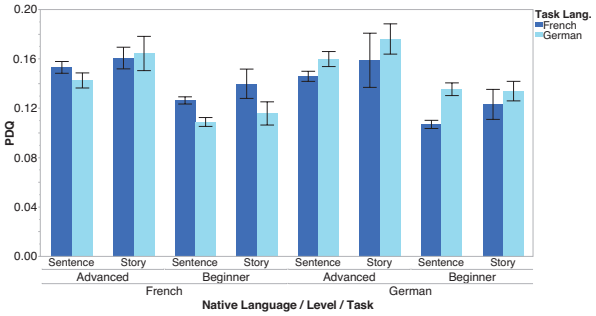


Figure 2: Mean PDQ for French and German speakers depending on task language, task and proficiency level of the learners.

Finally, we looked at individual speaker differences (Figure 3 for French native speakers and Figure 4 for German native speakers). The mean PDQ for French speakers speaking French shows a range from 0.102 to 0.162, whereas for the same speakers, the mean PDQ ranges in German from 0.078 to 0.171. For the German speakers, PDQ in German ranges from 0.083 to 0.194, whereas in the French production, the range lies between 0.069 and 0.164. This indicates that there is some individual variation, but that most speakers have a compressed F_0 in their L2 productions.

However, it becomes also apparent that three speakers show a somewhat different behavior. In the French group, there were two speakers (503 and 501) where the L1 productions showed a smaller PDQ than the L2 productions. One of them (503) is an advanced learner, the other a beginner (501). In the German group, one speaker (006) had almost identical PDQ values for L1 and L2 productions. This speaker was an advanced learner of French. Due to the small number of speakers overall, further interpretation of these results is not possible.

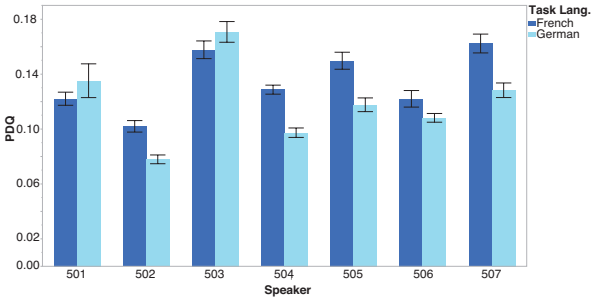


Figure 3: Mean PDQ of individual French speakers depending on task language. Speakers 503 and 505 are advanced learners.

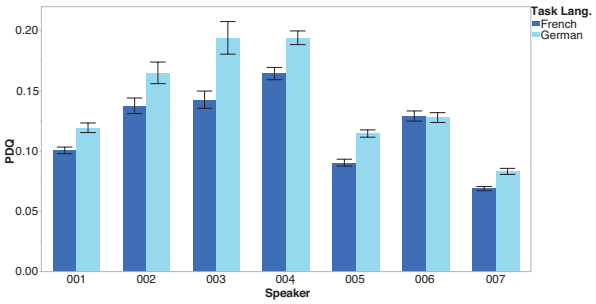


Figure 4: Mean PDQ of the individual German speakers depending on task language. Speakers 004 and 006 are advanced learners.

4. Discussion & conclusions

The first question this article was set out to investigate is whether there is a general trend that French learners of German and German learners of French compress their pitch range when speaking L2. The results indicate that the pitch range (measured in PDQ) is indeed smaller in the L2 for both language groups, that is, irrespective of the native language. This means that the actual pitch range that is produced by native speakers when they speak their native language is larger than the pitch range in their L2. This result is compatible with results presented by other researchers who also found a compressed pitch range in L2 productions (e.g. [9], [10], [20], [29]).

At this point, we cannot give clear answers to the question why this is the case. However, we can give a speculative explanation. This explanation would entail a factor of insecurity, because learners are not completely confident how and where exactly the correct F_0 targets have to be achieved. Furthermore, because they concentrate on other aspects, such as the correct pronunciation of segments, or the correct timing of stress, they could have a tendency to disregard the exact pitch range that is used by native speakers of a language. Future research needs to investigate the reasons for pitch range compression in more detail.

Furthermore, the results also replicate findings that there are differences in pitch range depending on the task (e.g. [1], [21]). In this study, speakers showed a higher PDQ in the story condition compared to the sentence condition. This finding is not surprising: the single sentences that were read by participants, occurred without any context and arguably do not have a communicative function. Therefore, contrastiveness or givenness, for instance, do not play a role in contrast to the story condition. Especially for the story of the three little pigs, a lively, narrative production is more likely than for single sentences.

As for the second question whether the L2 proficiency level of L2 has an influence on the pitch range compression, the results presented here are not conclusive. They do not exclude the possibility of an overall trend for advanced speakers to produce greater pitch variation in L2 compared to L1 and are closer to native speakers (as results by [29] would suggest). Of the three speakers that did not compress pitch range in the L2 production, two are advanced learners. However, such interpretations have to be made with caution, because there were only two advanced learners for each native speaker group. Additionally, there was also one French beginner with a higher PDQ in the non-native productions.

For language teaching, the results suggest that students should be made aware of the lack in pitch variation in order to sound more native-like. Results indicate that a special teaching of pitch variation indeed reduces the foreign accent (e.g. [7], [15], [16]). Pitch range differences that occur across languages otherwise might be aggravated (see also [4] for a cross-linguistic comparison of such differences).

As indicated in the introduction, the results reported here are preliminary. For instance, the two language groups differed with respect to the number of male and female speakers. Furthermore, there were only 2 advanced learners, compared to 5 beginners. More generally, the number of 7 speakers is not very high. However, despite the rather small and somewhat unbalanced number of speakers, we were able to obtain results that are promising for further research. The results reported in this article are in need of replication with larger speaker groups before far-reaching claims can be made. But the preliminary results are a starting point for further research, with a corpus that has more speakers. Such a corpus is planned to include many more speakers (see [28]) balanced for gender and proficiency level. Future research could also concentrate on the effect of different speaking styles on pitch range in a L2, such as reading alone, versus speaking in front of an audience, or engaging in conversations. In such scenarios, it would also be interesting to investigate whether convergence of pitch range can be observed, in that one speaker changes his or her behavior in response to an interlocutor.

5. Acknowledgements

This research was funded by ANR and DFG via the IFCASL project. We would like to thank our colleagues in Saarbrücken and Nancy who helped us with the acquisition of the data, and whose discussions improved this article tremendously as well as the three anonymous reviewers.

6. References

[1] Abu-Al-Makarem, A., & Petrosino, L., "Reading and spontaneous speaking fundamental frequency of young Arabic men for Arabic and English languages: A comparative study", Perceptual and Motor Skills, 105:572-580, 2007.

[2] Anderson-Hsieh, J., Johnson, R., & Koehler, K., "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure", Language Learning, 42(2):529-555, 1992.

[3] Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., & Trouvain, J., "Comparison of pitch range and pitch variation in Slavic and German languages", Proc. Speech Prosody, Dublin, Ireland, 2014.

[4] Aoyama, K., & Guion, S. G., "Prosody in second language acquisition: Acoutic analyses of duration and F0 range", in O.-S. Bohn & M. J. Munro [Eds.], The role of language experience in second-language speech learning - In honor of James Emil Flege, 281-297, Amsterdam: John Benjamins, 2007.

[5] Barry, W., Andreeva, B., & Steiner, I., "The phonetic exponency of phrasal accentuation in French and German", Proc. Interspeech 2007, Antwerp, Belgium, 1010-1013, 2007.

[6] Best, C. T., "A direct realist view of cross-language speech perception". In W. Strange (Ed.), Cross-language studies of speech perception: A historical review, 171-206, York: Timonium, 1995.

[7] Bissiri, M. P., & Pfitzinger, H. R., "Italian speakers learn lexical stress of German morphologically complex words", Speech Communication, 51(10):933-947, 2009.

[8] Boersma, P., & Weenink, D., "PRAAT: Doing phonetics by computer", (Version Version 5.3.59), retrieved from http://www.praat.org/, 2013.

[9] Busà, M. G., & Stella, A., "Intonational variations in focus marking in the English spoken by North-East Italian speakers", In M. G. Busà & A. Stella [Eds.], Methodological perspectives on second language prosody - Papers from ML2P 2012, 31-35, 2012.

[10] Busà, M. G., & Urbani, M., "A cross linguistic analysis of pitch range in English L1 and L2", Proc. 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong, 380-383, 2011.

[11] Caspers, J., & Kepinska, O., "The influence of word-level prosodic structure of the mother tongue on production of word stress in Dutch as a second language", Proc. 17th International Congress of Phonetic Sciences (ICPhS XVII), Hong Kong. 420-423, 2011.

[12] Colotte, V., "Corpus Recorder", Nancy: LORIA, 2013.

[13] Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S., "Persistent stress 'deafness': The case of French learners of Spanish", Cognition, 106:682-706, 2008.

[14] Flege, J. E., Munro, M. J., & Fox, R. A., "Auditory and categorical effects on cross-language vowel perception", Journal of the Acoustical Society of America, 95(6):3623-3641, 1994.

[15] Hincks, R., "Processing the prosody of oral presentations" Proc. InSTIL/ICALL2004 NLP and Speech Technologies in Advanced Language Learning, Venice (Italy), 63-66, 2004.

[16] Hincks, R., & Edlund, J., "Promoting increased pitch variation in oral presentations with transient visual feedback", Language Learning & Technology, 13(3):32-50, 2009.

[17] Keating, P., & Kuo, G., "Comparison of speaking fundamental frequency in English and Mandarin", The Journal of the Acoustical Society of America, 132(2):1050-1060, 2012.

[18] Kingston, J., "Learning foreign vowels", Language and Speech, 46(2-3):295-349, 2003.

[19] Koreman, J., Andreeva, B., & Barry, W., "Accentuation cues in French and German", Proc. Speech Prosody 2008, Campinas, Brazil. 613-616, 2008.

[20] Mennen, I., "Can language learners ever acquire the intonation of a second language?", Proc. STiLL, Marholmen (Sweden), 17-20, 1998.

[21] Mennen, I., Schaeffler, F., & Docherty, G. " Pitching it differently: A comparison of the pitch ranges of German and English speakers", Proc. 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken, 1769-1972, 2007.

[22] Mennen, I., Schaeffler, F., & Docherty, G., "Cross-language differences in fundamental frequency range: a comparison of English and German", The Journal of the Acoustical Society of America, 131(3):2249-2260, 2012.

[23] SAS., "JMP" (Version 10), Cary (NC): SAS Institute, 2012.

[24] Talkin, D., "A Robust Algorithm for Pitch Tracking (RAPT)", In Kleijn, W. B. and Paliwal, K. K. [Eds.], Speech Coding and Synthesis. New York: Elsevier, 1995.

[25] Traunmüller, H., & Eriksson, A., "The frequency range of the voice fundamental in the speech of male and female adults", (manuscript http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf. last access: December 11, 2013), 1995.

[26] Trofimovich, P., & Baker, W., "Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech", Studies in Second Language Acquisistion, 28:1-30, 2006.

[27] Trouvain, J., & Gut, U. [Eds.], "Non-native prosody: Phonetic description and teaching practice", Berlin/New York: Mouton de Gruyter, 2007.

[28] Trouvain, J., Laprie, Y., Möbius, B., Andreeva, B., Bonneau, A., Colotte, V., Fauth, C., Fohr, D., Jouvet, D., Mella, O., Jügler, J., & Zimmerer, F., "Designing a bilingual speech corpus for French and German language learners", Proc. Corpus et Outils en Linguistique, Langues et Parole: Statuts, Usages et Mésuages, Strasbourg, 32-34, 2013.

[29] Ullakonoja, R., "Comparison of pitch range in Finnish (L1) and Russian (L2)", Proc. 16th International Congress of Phonetic Sciences (ICPhS XVI), Saarbrücken. 1701-1704, 2007.

[30] van Bezooijen, R., "Sociocultural aspects of pitch differences between Japanese and Dutch women", Language and Speech, 38(3):253-265, 1995.

Selection of Training Data for HMM-based Speech Synthesis from Prosodic Features - Use of Generation Process Model of Fundamental Frequency Contours -

Tomoyuki Mizukami[1], Hiroya Hashimoto[2], Keikichi Hirose[1], Daisuke Saito[1], and Nobuaki Minematsu[2]

[1] Graduate School of Information Science and Technology,
[2] Graduate School of Engineering,
University of Tokyo
{mizukami, hiroya, hirose, dsk_saito, mine}@gavo.t.u-tokyo.ac.jp

Abstract

Generation process model of fundamental frequency (F_0) contours is ideal to represent global movements of F_0's keeping a clear relation with back-grounding linguistic information of utterances. Using the model, improvements of HMM-based speech synthesis are expected. A new method is developed to cope with erroneous F_0's of utterances included in HMM training corpus. F_0 extraction errors not only cause wrong F_0's, but also degrade segmental features of synthetic speech, since they affect the over-all accuracy of speech analysis. The method is to exclude speech segments from HMM training, where extracted F_0's are largely different from those generated by the generation process model. Experiments on speech synthesis showed a clear improvement in synthetic speech quality when phoneme-based exclusion is conducted with a properly selected threshold.

Index Terms: F_0 contour, generation process model, HMM-based speech synthesis, training segment selection

1. Introduction

HMM-based speech synthesis attains a special attention from researchers in speech communication field, since it can generate a good quality of speech from a rather limited size of speech corpus with flexible controls on voice qualities and utterance styles [1]. It can handle fundamental frequencies (F_0's) of speech in the same frame-by-frame manner with other acoustic features such as mel-cepstral coefficients. It is easy to prepare a training corps, since F_0 values can be directly used for HMM training without assuming any models of F_0 contours. However, this in turn causes demerits; it generally produces over-smoothed F_0 contours with occasional F_0 undulations not observable in human speech. Automatic extraction of F_0's occasionally outputs erroneous results; wrong F_0's and voiced/unvoiced decision errors. These errors not only degrade synthetic speech quality from prosodic feature aspect but also affect extraction of spectral envelope features, resulting in degradation also from segmental feature aspect.

The generation process model of F_0 contours (F_0 model) developed by Fujisaki and his co-workers can solve the problems of HMM-based speech synthesis [2]. The model represents a sentence F_0 contour in logarithmic scale as super-position of accent components on phrase components. These components are known to have clear correspondences with linguistic and para-/non- linguistic information, which is conveyed by prosody. Thus, using this model, a better control is possible in F_0 contour generation than the frame-by-frame control. Because of clear relationship between generated F_0 contours and linguistic and para-/non- linguistic information of utterances, manipulation of generated F_0 contours is possible, leading to a flexible control of prosody. We already have developed several methods, which use the F_0 model-generated F_0's in HMM-based speech synthesis, and showed their advantages in F_0 controls [3-7]. Recently, a method has been proposed, which uses F_0's approximated by the F_0 model instead of observed F_0's of training corpus for HMM training [8]. The method can partly solve the problem, which arises from F_0 extraction errors, but cannot avoid spectral envelope features affecting the synthetic speech quality.

In the current paper, the F_0 model is used in a different way to cope with the issue of erroneous F_0's; to exclude from HMM training samples which have large differences in F_0's from those approximated by the F_0 model. The performance of the method may rely on "how accurately the F_0 model parameters can be extracted from observed F_0 contours." For this purpose, the paper uses the method of automatic extraction of model parameters recently developed by the authors [8]. The method is motivated on how experts do when finding the F_0 model parameters, and utilizes linguistic information of utterances and our knowledge on Japanese prosody.

The rest of the paper is organized as follows: following to the explanation on the F_0 model, a new method of automatic extraction of F_0 model commands is briefly introduced in section 2. Section 3 shows the proposed method of datum selection for HMM training with the results of speech synthesis experiments. Some discussions on the method are given in Section 4. Section 5 concludes the paper.

2. Automatic extraction of F_0 model commands

2.1. F_0 model

Movements of F_0 along time axis are well represented by the F_0 model, which is a command-response model that describes F_0 contours in logarithmic scale as the superposition of phrase and accent components [3]. The *i*-th phrase component $G_{pi}(t)$ of the F_0 model is generated by a second-order, critically-damped linear filter in response to an impulse-like phrase command, while the *j*-th accent component $G_{aj}(t)$ is generated by another second-order, critically-damped linear filter in response to a stepwise accent command:

$$G_{pi}(t) = \begin{cases} \alpha_i^{\,2} t\, e^{-\alpha_i t} & t \geq 0 \\ 0 & t < 0 \end{cases}, \quad (1)$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1+\beta_j\, t)\, e^{-\beta_j t}, \gamma\right] & t \geq 0 \\ 0 & t < 0 \end{cases}. \quad (2)$$

Based on the analysis of Japanese utterances, time constants α_i and β_j are known to be fixed to values around 3.0 s^{-1} and 20.0 s^{-1}, respectively. The parameter γ, which thresholds accent components, can also be set to a fixed value around 0.9. An F_0 contour is then given by the following equation (assuming natural logarithm):

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \{ G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j}) \} , \quad (3)$$

where, F_b is the bias level, I is the number of phrase components, J is number of accent components, A_{pi} is the magnitude of the i^{th} phrase command, A_{aj} is the amplitude of the j^{th} accent command, T_{0i} is the time of the i^{th} phrase command, T_{1j} is the onset time of the j^{th} accent command, and T_{2j} is the reset time of the j^{th} accent command.

2.2. Method

When searching F_0 model parameters, time constants α_i and β_j, threshold γ, and bias level F_b are usually fixed and are put out of the search process. Therefore, the search is done for F_0 model parameters related to phrase and accent commands, and is called F_0 model command extraction. Several methods have already been developed for automatically extracting F_0 model commands from given F_0 contours. Their basic idea is: smoothing to avoid micro-prosodic and erroneous F_0 movements, interpolating to obtain continuous F_0 contours, and taking derivatives of F_0 contours to extract accent command locations and amplitudes [9, 10]. Phrase commands are extracted from the residual F_0 contours (F_0 contour minus extracted accent components) [9], or from low frequency components of F_0 contours (assuming F_0 contours as waveforms) [10-12]. Extracted phrase and accent commands are optimized by a successive process. These methods, however, are not robust for pitch extraction errors, and produce commands not corresponding to the linguistic information of the utterances to be analyzed. Although attempts have been conducted to reduce extraction errors by introducing constraints (on command locations) induced from the linguistic information, their performances are still not satisfactory [13].

Interpolation of F_0 contours has a drawback since it relies on F_0's around voiced/unvoiced boundaries, where F_0 extraction is not always precise. This situation leads to extraction of false commands. Micro-prosodic F_0 movements at voiced consonants may also degrade the command extraction performance, since they are not counted in the F_0 model. To avoid false extractions, a new method is developed which accounts F_0 contours only of vowel segments [8]. In turn, since no F_0 is available between vowels, it comes difficult to extract accent commands from F_0 contour derivatives. Therefore, the new method takes features of Japanese prosody into account. In Japanese, F_0's of a syllable take either High or Low values corresponding to accent types. The method extracts phrase commands first viewing "Low" parts, and then find accent command amplitudes from "High" parts. The method can extract minor phrase commands, which are difficult to be found from the residual F_0 contours. We can say that the new method is motivated from the human process of command extraction.

The method consists of three steps; pre-processing, parameter extraction, and optimization. As for accent phrase boundaries, and accent types, which are necessary for the following processes, those used in HMM-based speech synthesis are used; the method is in good match with HMM-based speech synthesis. Phoneme boundaries are detected by the forced alignment using Julius as the recognizer [14]. A significant improvement in extraction performance as compared to conventional methods is observed. Since it is developed taking Japanese prosody into account, further investigations are necessary to make it applicable to other languages. The detail of the method is given in [8].

3. Speech synthesis with selected training data

3.1. Selection of training data

Figure 1 shows an example of F_0's extracted by STRAIGHT [15], and their F_0 model approximation. Here, their absolute difference in (natural) logarithmic values in each frame is denoted as F_0 difference:

$$F_{0diff}(t) = \left| \ln F_{0obs}(t) - \ln F_{0\,model}(t) \right| . \quad (4)$$

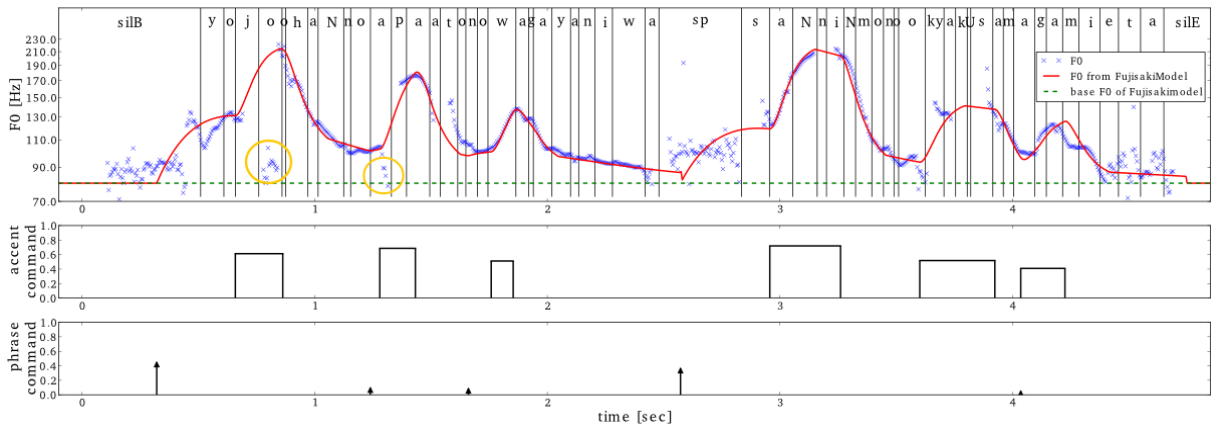


Figure 1: *An example for extracted F_0's (×) and their F_0 model approximation (in red solid line). F_0 model parameters (accent and phrase commands) are also shown. ("yojoohaNno apaatono wagayaniwa, saNniNmono okyakuga mieta": Three guests visited to my house, which was an apartment of only four and half "tatami" space.)*

$F_{0obs}(t)$ and $F_{0\,\mathrm{mod}\,el}(t)$ denote extracted F_0 and F_0 model F_0 at frame t, respectively. Half pitch extraction errors are observable around the second /o/. Also pitch extraction errors occur around the latter half of the second /a/. These segments may affect badly for the HMM training. It is possible to correct F_0's and to use them for HMM training, but the erroneous F_0's also cause errors in mel-cepstral coefficients (through STRAIGHT analysis). Here, speech segments with large F_0 differences are excluded from the data for HMM training. (Although F_0's are "wrongly" extracted at some speechless parts/pauses, they are ignored through F_0 model approximation and selection processes.) Two schemes are checked; one to exclude whole sentences which include many frames with large F_0 differences, and the other to exclude phoneme segments with large F_0 differences.

3.2. Experiments

Speech synthesis experiments are conducted using ATR continuous speech corpus of 503 sentences by male speaker MMI [16]. Out of 503 sentences, 450 sentences are used for HMM training, keeping 53 sentences for evaluation. Utterances for HMM training are analyzed using STRAIGHT with 5 msec frame shift. Fundamental frequencies are searched between 80 Hz to 250 Hz. Mel-cepstral coefficients and aperiodicity indices are calculated from analysis results by STRAIGHT using SPTK [17]. Feature parameters used for HMM training and speech synthesis processes have 138 dimensions: mel-frequency coefficients (0^{th}–39^{th}), aperiodicity indices (0-1 kHz, 1-2 kHz, 2-4 kHz, 4-6 kHz, 6-8 kHz), logarithmic F_0, and their Δ and Δ^2 features. Five-state left-to-right hidden semi-Markov model with single Gaussian distribution for each state, provided in HTS-2.1 [18], are used. A Gaussian distribution is represented by a diagonal conversion matrix. Decision tree-based context clustering is conducted with MDL stop criterion.

F_0 model parameters are extracted automatically from F_0 contours of training sentences using the method explained in section 2.2. Then, F_{0diff} is calculated for each voiced frame. When a frame has F_{0diff} larger than 1.0, and none of F_{0diff}'s of neighboring 10 frames (preceding 5 and following 5 frames) exceeds 1.0, F_0 model parameter extraction is re-conducted by neglecting the frame, and F_{0diff}'s are re-calculated.

As for sentence-based datum selection, utterances are excluded from HMM training, when they include "more than 2 frames with F_{0diff}'s larger than 1.0" or "more than 10 frames with F_{0diff}'s larger than 0.8." Forty three sentences are excluded from the 450 sentences by the process. These parameters for datum selection are determined through a preliminary experiment. Speech synthesis is conducted for two cases; one is when all 450 sentences are used for training (conventional method), and the other is when 43 sentences are excluded from the training (proposed method). Synthetic speeches obtained by these two methods are compared through a listening test with 3 scale scoring (1: better quality by the proposed method, 0: similar quality, -1: better quality by the conventional method). Two native speakers of Japanese conducted the listening test. Each speaker evaluated 10 sentences, which are selected randomly from 53 sentences for testing. The averaged result with 95 % confidence interval is 0.250 ± 1.337, indicating no significant difference between two methods.

Sentence-based datum selection may have a shortcoming that parts without F_0 extraction errors (and thus useful for HMM training) are thrown away together with other parts with erroneous F_0's. In order to cope with this situation, phoneme-based datum selection is conducted. Figure 2 shows the process. When second /a/, which is sand-witched by phonemes /r/ and /y/, includes a frame (frames) with large F_{0diff}, it is deleted from the training data. However, other phonemes in the same sentence without large F_{0diff}'s are left for HMM training. Three thresholds for "large F_{0diff}" are set so that 5 %, 10 % or 30 % of total voiced phoneme segments of the training corpus are excluded from the training. (No selection process is conducted for segments judged as voiceless.) Two versions of synthetic speeches, one by the conventional method and the other by the proposed method are compared through the listening test with 5 scale scores (+2: clearly better quality by the proposed method, +1: slightly better quality by the proposed method, 0: similar quality, -1: slightly better quality by the conventional method, -2: clearly better quality by the conventional method.). Nine native speakers evaluated the synthetic speeches by randomly selecting 20 sentences out of 53 sentences for testing in one turn. Every speaker conducted tree turns. The averaged results with 95 % confidence intervals are summarized in Table 1. Advantage of the proposed method over the conventional method is clear for the 5 % exclusion level.

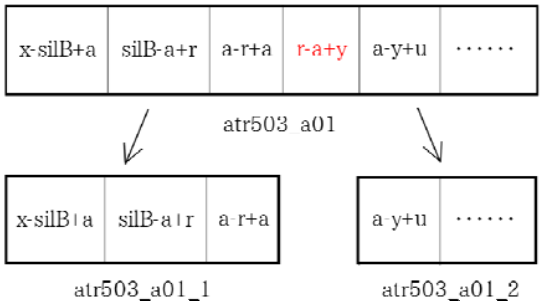


Figure 2: *Process of phoneme-based datum selection. Tri-phone HMM's, which are adopted for Japanese speech synthesis, are assumed. If tri-phone "r-a+y" has erroneous F_0's, it is excluded from the HMM training data.*

Table 1. *Scores of listening test with 95% confidence intervals for three levels of phoneme-based exclusion. Positive values indicate better quality by the proposed method.*

Level of exclusion	Score with 95 % confidence interval
5 %	0.300±0.148
10 %	0.139±0.147
30 %	-0.378±0.162

Figure 3 compares F_0 contours generated by the conventional method and the proposed method. Around sentence initial ("hyogeN"), a stable F_0 contour is obtained by the proposed method. F_0 contours around "nooryoku" and

"tsukeru" show better matches with accent types of the parts by the proposed method.

4. Discussion

Experiments are also conducted similarly to speech samples uttered by another speaker (female speaker FTY). Advantage of the proposed method, however, is not shown. The reason for the result will be due to the fact that F_{0diff}'s for FTY utterances are smaller than those for MMI. Selection of training data decreased the training datum size, and might affect negatively to the synthesized speech quality. Although, in the current experiments, the phoneme selection is conducted so that a fixed percentage of the training corpus remains, a scheme needs to be developed to use an absolute F_{0diff} threshold for selection. Also different thresholds can be used for different phonemes. A further study is necessary on the issue.

Reasons for large F_{0diff} can be divided into two cases; pitch extraction error and micro-prosody. Pitch extraction error can be further categorized several cases; double/half pitch errors which can be corrected through a post processing, pitch errors with large aperiodicity, and voiced/unvoiced decision errors. These errors should be treated differently. It should be noted that the method only counts voiced segments and no selection process works for voiceless frames. However, as is clear from Figure 1, F_0's are extracted for some unvoiced phonemes. When they have large F_{0diff}'s, they can be excluded from the HMM training. It is necessary to check how these exclusions affect the synthetic speech quality.

5. Conclusions

A method is developed to exclude speech segments from HMM training where their extracted F_0's are largely different from F_0's generated by the generation process model. Results on listening test for synthetic speech by the proposed method and the original HMM-based speech synthesis showed a clear improvement in synthetic speech quality when exclusion is done in phoneme-basis. Using F_0 values generated by the F_0 model without excluding segments with large F_{0diff} is another possible way to cope with erroneous F_0's [8]. A combined method is in our future research scope.

6. Acknowledgement

This work is partly supported by Grant-in-Aid for Scientific Research (B) #24300068, JSPS, and the Major Program for the National Social Science Fund of China (13&ZD189).

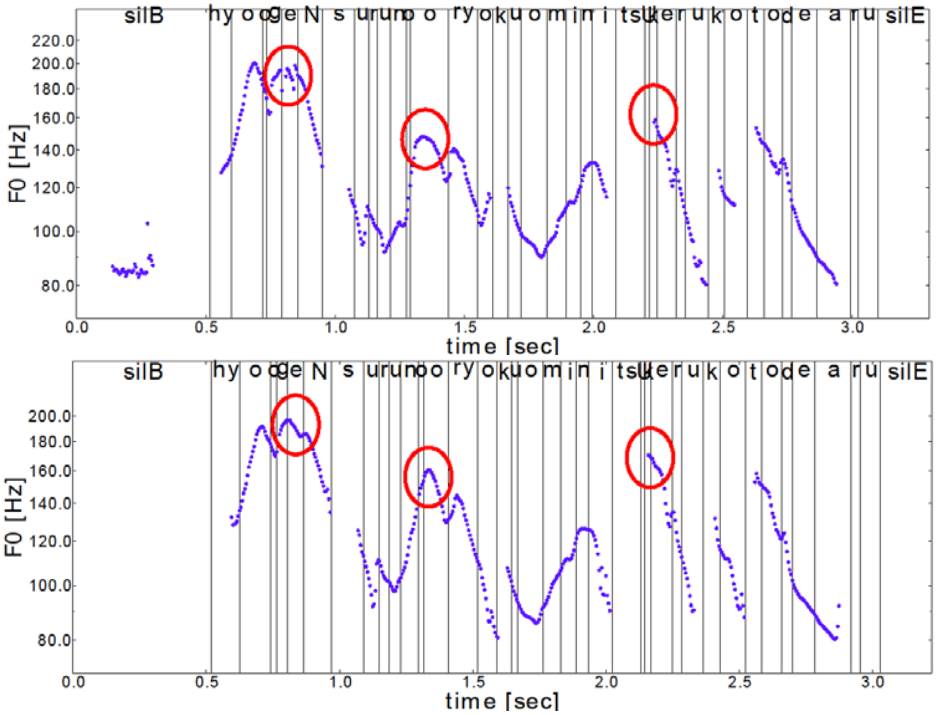


Figure 3: *F_0 contours obtained by the conventional method and the proposed method (phoneme-based). Improvements by the proposed method are observable at parts with red circles. ("hyogeNsuru nooryokuo minitsukeru kotodearu": It is to obtain a skill for presentation.)*

7. References

[1] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. IEEE ICASSP*, pp.1315-1318, 2000.

[2] H. Fujisaki, and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242, 1984.

[3] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, Nos.3-4, pp.385-404, 2005.

[4] K. Ochi, K. Hirose, and N. Minematsu, "Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model," *Proc. IEEE ICASSP*, pp.4485-4488, 2009.

[5] K. Hirose, K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu, "Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency," *Proc. INTERSPEECH*, pp.2793-2796, 2011.

[6] T. Matsuda, K. Hirose, and N. Minematsu, "Applying generation process model constraint to fundamental frequency contours generated by hidden-Markov-model-based speech synthesis," *Acoustical Science and Technology, Acoustical Society of Japan*, Vol.33, No.4, pp.221-228, 2012. .

[7] K. Hirose, H. Hashimoto, J. Ikeshima, and N. Minematsu, "Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model," *Proc. International Conf. on Speech Prosody*, pp.171-174, 2012.

[8] H. Hashimoto, K. Hirose, and N. Minematsu, "Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis," *Proc. INTERSPEECH*, 4 pages, 2012.

[9] S. Narusawa, N. Minematsu, K. Hirose, and H. Fujiaski, "A method for automatic extraction of model parameters from fundamental frequency contours of speech," *Proc. IEEE ICASSP*, pp.509-512, 2002.

[10] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. IEEE ICASSP*, vol.3, pp.1281-1284, 2000.

[11] V. Strom, "Detection of accents, phrase boundaries and sentence modality in German with prosodic features," *Proc. EUROSPEECH, V*ol. 3, pp. 2039.2041, 1995.

[12] A. Sakurai and K. Hirose, "Detection of phrase boundaries in Japanese by low-pass filtering of fundamental frequency contours," *Proc. International Conference on Spoken Language Processing*, Vol.2, pp.817-820, 1996.

[13] K. Hirose, Y. Furuyama, and N. Minematsu, "Corpus-based extraction of F_0 contour generation process model parameters," *Proc. INTERSPEECH*, pp. 3257-3260, 2005.

[14] http://julius.sourceforge.jp/

[15] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0, and aperiodicity estimation," *Proc. IEEE ICASSP*, pp.3933-3936, 2008.

[16] A. Kurematsu, K.Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, Vol. 9, pp.357–363, 1990.

[17] http://sp-tk.sourceforge.net/

[18] http://hts.sp.nitech.ac.jp/

SVR vs MLP for Phone Duration Modelling in HMM-based Speech Synthesis

Alexandros Lazaridis, Pierre-Edouard Honnet, Philip N. Garner

Idiap Research Institute, Martigny, Switzerland

{alaza, pierre-edouard.honnet, phil.garner}@idiap.ch

Abstract

In this paper we investigate external phone duration models (PDMs) for improving the quality of synthetic speech in hidden Markov model (HMM)-based speech synthesis. Support Vector Regression (SVR) and Multilayer Perceptron (MLP) were used for this task. SVR and MLP PDMs were compared with the explicit duration modelling of hidden semi-Markov models (HSMMs). Experiments done on an American English database showed the SVR outperforming the MLP and HSMM duration modelling on objective and subjective evaluation. In the objective test, SVR managed to outperform MLP and HSMM models achieving 15.3% and 25.09% relative improvement in terms of root mean square error (RMSE) respectively. Moreover, in the subjective evaluation test, on synthesized speech, the SVR model was preferred over the MLP and HSMM models, achieving a preference score of 35.93% and 56.30%, respectively.

Index Terms: phone duration modelling, Support Vector Regression, Multilayer Perceptron, HSMM explicit duration modelling, HMM-based speech synthesis

1. Introduction

Prosody plays a very important role in verbal communication. Changing prosody can completely change the meaning of the message which is conveyed through speech [1]. There are three main aspects of prosody: duration, pitch and intensity [2]. Duration is a prosodic factor controlling the speaking rate, the rhythm of the speech [3]. Controlling this factor gives the ability to the speaker to emphasize more on some parts of a sentence and less on others, helping the listener to perceive the proper message. In the same way, duration and prosody in general, are essential factors in the field of speech synthesis.

Statistical parametric speech synthesis techniques, and hidden Markov models (HMMs) in particular, provide a framework for the task of speech synthesis, achieving on one hand high quality synthetic speech and on the other hand giving a high degree of flexibility in modelling and transforming various aspects of the speech, such as speaker identity, age, gender, emotions and prosody [4, 5, 6, 7]. Over the last years, many improvements have been introduced in HMM-based speech synthesis, one of them being the use of hidden semi-Markov models (HSMMs) [8]. The advantage of HSMMs is the explicit modelling of state duration using Gaussian distributions instead of the implicit modelling of HMMs by the transition probabilities of the states. In the training phase of HSMM-based speech synthesis, a decision tree is built according to some phonetic and linguistic features and a set of fixed binary (yes/no) questions controlling and even sometimes limiting [9, 10] the structure of the tree. The minimum description length (MDL) is used as a splitting and stopping criterion [11]. In the synthesis phase, the decision tree is traversed for each target unit until reaching a leaf node. The mean value of this leaf node is used as the predicted duration for the target unit. The drawback of this approach is that these trees cannot represent properly all the target units in speech synthesis [12].

To improve synthetic speech and alleviate monotonous prosody and specifically monotonous durations, a lot of research has been done over the last years. Various approaches and techniques, such as modelling duration combining models of multiple levels, e.g. state and phone levels [13], state, phone and syllable levels [14], using full covariance Gaussian distribution [15] or implementing Gamma distribution instead of Gaussian [16], have been introduced for this task. Furthermore, a lot of focus has been given on using external duration models, forcing their predicted durations on HMM-based speech synthesis [17, 18]. A variety of machine learning algorithms have been used for state, phone or syllable duration modelling, such as decision trees [19, 20], Bayesian Networks [21], Linear Regression [22], Instance-based learning [22], Support Vector Regression [23, 24], Multilayer Perceptron [25, 26], or even fusion of these algorithms [27, 28], to improve the accuracy.

The motivation behind this work is to investigate how Support Vector Regression (SVR) and Multilayer Perceptron (MLP) algorithms, which have been used successfully in various tasks, could improve, as external phone duration models, the prediction accuracy in order to improve the quality of synthesized speech. To our best knowledge, these two algorithms have never been compared in the same experimental conditions. We believe that the limitations caused by the use of HSMMs for the duration modelling, e.g. the use of a specific set of questions for the decision tree, or the difficulty of the decision trees to model complex context dependencies [9], could be overcome with the use of the external duration models. Furthermore, we consider that the ability of SVR in coping well with high-dimensional space in respect to the training data will result in a more robust duration model in comparison to a model build using the MLP. An American English male database is used (CMU-ARCTIC-RMS) for these experiments [29].

The rest of the paper is organized as follows. The HSMM explicit duration model and the two external phone duration modelling approaches, MLP, SVR are presented in Section 2 and 3 respectively. The experimental setup and results are presented in Section 4. In Section 5, the conclusions are given.

2. HSMM-based Duration Modelling

In HMM-based speech synthesis duration modelling is done at the state level, through the state sequence modelling. Consequently, the phone duration modelling is performed through the state modelling. The reasoning behind this approach is the fact that the state sequence modelling is not only responsible for the duration of the phones, but also is the basic structural element of the HMMs for spectrum and f0 modelling and generation.

In HMM-based speech synthesis duration modelling is done implicitly through the transition probabilities of the HMM

states i.e. an exponential distribution, making this structure unsuitable for modelling properly the timing in synthetic speech. The advantage of HSMMs in respect to HMMs, is the explicit modelling of state duration using Gaussian distributions instead of the implicit modelling by the transition probabilities of the states. Although the Gaussian distribution is clearly wrong (it implies negative durations are possible), it is a suitable approximation to the true distribution.

In the training phase, using the state durations and the phonetic and prosodic context-dependent features of the training data, a decision tree is built. This decision tree is constructed based on some predetermined binary questions concerning the content of the features (e.g. is the current syllable accented, is the previous phone fricative, etc.). For controlling the growth of the tree and the splitting of the nodes, the minimum description length (MDL) criterion is used [11]. Finally, the leaf nodes of the tree correspond to different clusters of the training data, sharing the same distributions (i.e. mean and variances). In the synthesis phase, according to the unseen data, the tree is traversed from the root node until a leaf node is reached. In this way, the Gaussian distributions of the leaf nodes are used to determine the duration of the synthesized speech.

Using HMMs/HSMMs for modelling duration in speech synthesis leads to some drawbacks. First, it is inefficient to express complex context dependencies such as XOR, parity or multiplex problems by decision trees [9]. In order to be able to cope with such cases, decision trees must become very large. Furthermore, the mean values of the Gaussian distributions of the leaf nodes are inadequate, due to over-generalization, to deal properly, in respect to duration modelling, with all the cases of the unseen data during the synthesis phase.

For overcoming these problems and improving the duration modelling accuracy, two external models are implemented and evaluated in this work, using the SVR and MLP algorithms.

3. External Phone Duration Modelling

In this section the two external phone duration models, the MLP and SVR, are described. When an external phone duration model is used in HMM-based speech synthesis, the predicted duration of the phone during the synthesis phase, has to be forced upon the HMMs [30]. Consequently the HMMs for each phone, having the predefined phone duration, are used only to distribute the predicted phone duration to the states.

3.1. Multilayer Perceptron

The Multilayer Perceptron is a feed-forward neural network having one or more hidden layers between the input and output layers [31]. Having a feed-forward architecture means that the connections between all the units and layers follow only one direction, from input units to the output units. Apart from the input units, each unit is modelled using a non-linear activation function. Furthermore, each unit of a layer is connected with a specific weight to every unit of the next layer. Consequently, the input layer is connected to the output layer through a weighted linear combination of non-linear functions. In this way the input data are transformed into another space, where can be linearly separable. In our experiments the MLP was implemented using one hidden layer.

3.2. Support Vector Regression

A Support Vector Machine (SVM) constructs a hyperplane in a high-dimensional space, which can be used for classification (SVM) and regression (SVR) tasks [32]. The basic idea governing the SVR is the production of a model that can be expressed through support vectors which define the hyperplane. A linear regression function is used to approximate the training instances by minimising the prediction error. A parameter ε defines a tube around the regression function. In this tube the errors are ignored. The parameter ε controls how closely the function will fit the training data. The parameter C is the penalty for exceeding the allowed deviation defined by ε. The larger the C, the closer the linear regression function can fit the data [33].

For our experiments the support vector regression (SVR) model [34], which employs the sequential minimal optimization (SMO) algorithm for training a support vector classifier [32], was used. Many kernel functions have been used in SVR such as the polynomial, the radial basis function (RBF) and the Gaussian functions [35], etc. In this paper, after some preliminary experiments, the RBF kernel was selected [35].

4. Experiments

As mention earlier, there are two hypotheses we are interested in verifying with the following experiments. Firstly, whether external models could build more robust phone duration models than the explicit modelling of HSMMs. Secondly, whether the SVR model, since SVMs cope in a better way with the high-dimensionality of the feature space than MLP, would outperform the MLP external PDM.

4.1. Experimental Setup

In this section the database along with the feature set used in the experiments are presented. Furthermore, the setup of the HSMM explicit duration model and the external SVR and MLP phone duration models are described. The same HSMM framework, used for the explicit duration modelling, was also used for the HMM-based speech synthesis models used for synthesizing speech (using the predicted by each model durations) for the subjective evaluation test. The implementation of the external SVR and MLP PDMs was done with the WEKA software [36].

4.1.1. Database and Feature Set

For the experiments, the RMS voice of the CMU-ARCTIC database was used [29] which is a database of standard size for speaker-dependent HMM-based speech synthesis. The RMS voice is an American English male containing 1132 sentences of reading style speech. The data were divided into three sets, a training set containing 900 sentences for training the three models, a development set of 100 sentences for fine tuning the external phone duration models and a test set of 132 sentences for evaluating the three models with objective and subjective evaluation tests. Throughout the entire database, all starting and ending silences in each sentence were removed. Only the internal silences (silences between words in each sentence - phone "pau") were kept and modelled as the rest of the phones. This concluded to a phone set of 41 phones.

Concerning the features used for training the HSMM system, a standard in HMM-based speech synthesis set of features was used, composed of phonetic and prosodic features such as phone identity, identity of the two previous and next phones, number of syllables in a word, accented/stressed syllable, etc., concluding to a 53 feature set. For the external phone duration models, the same feature set was used expanded with some additional articulatory features. These articulatory features correspond to information such as the category of the phone (e.g. vowel, approximant, nasal, etc.), vowel length (e.g. short, long, etc.), height (e.g. high, middle, low), frontness (e.g. front, mid-

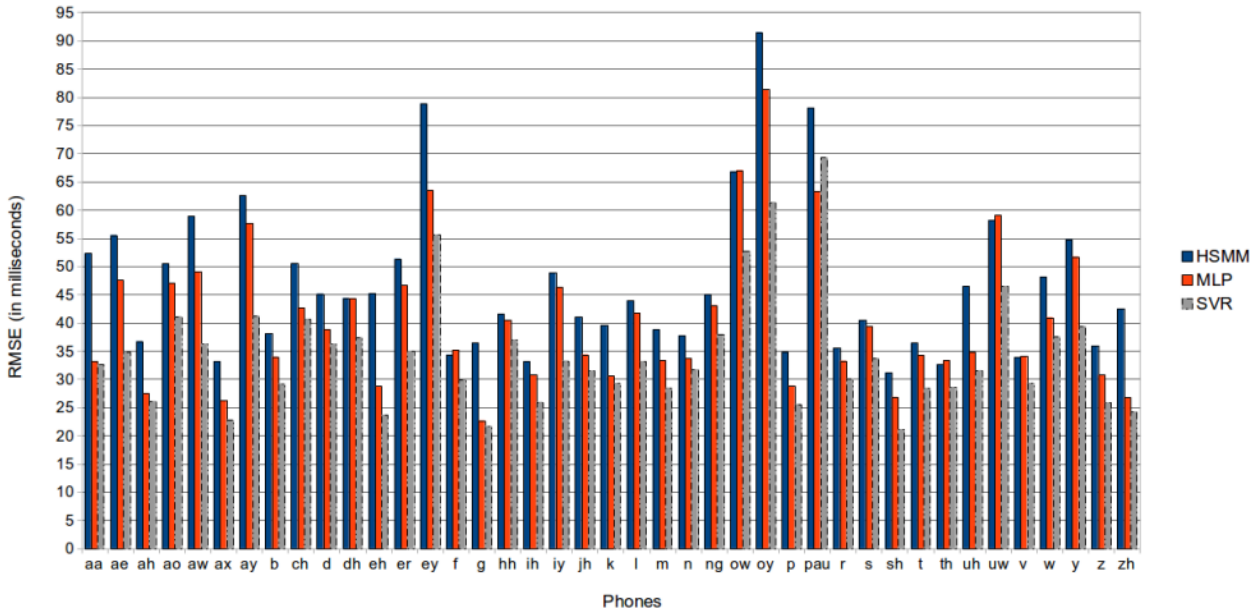


Figure 1: *This figure shows the RMSE for the three PDMs (HSMM, MLP and SVR) per phone*

dle, back), place of articulation (e.g. labial, alveolar, palatal, etc.), etc. The Relief [37] feature selection algorithm was used in some preliminary experiments for selecting among 37 binary articulatory features and their temporal (one previous and one following phones) information, the most appropriate ones. The final feature set consisted of 100 features.

4.1.2. HSMM model

For the implementation of the HSMM model, the version 2.2 of the HTS toolkit [38] was used. The speech data which were used had 16kHz sampling frequency. Five-state, left-to-right, no-skip HSMMs were used. The speech parameters which were used for training the HSMMs were 24th order mel-cepstral coefficients [39], log-f0 and 21-band aperiodicities [38], along with their delta and delta-delta features, extracted every 5 milliseconds (ms). The number of the used questions and the number of the leaf nodes of the decision tree were 304 and 547 respectively. STRAIGHT [40] was used for the analysis and synthesis phase of the HSMM-based speech synthesis.

4.1.3. MLP model

For the MLP model, a backpropagation based approach was used. The 100 input units (features) were converted to 570 units since all the nominal (categorical) features were converted to binary ones - an attribute with k nominal values is transformed into k binary attributes. The MLP model consisted of the input layer with 570 units, a hidden layer (H) with 10 units and an output layer with one unit (phone duration). The learning rate (L) of MLP, to ensure that the weights converge to a response fast enough without producing oscillations [41], was set equal to 0.05. The momentum term (M), which determines the degree to which each weight change will depend on the previous weight change, was set equal to 0.05. The epoch of the MLP (N), which determines the maximum number of iterations in which the full training set is presented in the model, was set equal to 500. These values were selected after a grid search (H={5:1:90}, L={0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 2.0}, M={0.01, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0, 1.5, 2.0}, N={50, 500, 1000, 5000, 10000, 50000}) of the model on the development set in respect to the RMSE of the model.

4.1.4. SVR model

For training the SVR model, as in the case of MLP, 570 binary features were used. In our experiments the RBF kernel was used as mapping function for the SVR. The ε and C parameters, where $\varepsilon \geq 0$ is the maximum deviation allowed during training and $C > 0$ is the penalty parameter for exceeding the allowed deviation, were set equal to 0.005 and 0.5 respectively. The gamma (G) parameter of the RBF function, determining the RBF width, was set equal to 0.05. These values were selected after a grid search fine tuning (ε={0.001, 0.003, 0.005}, C={0.5, 1.0, 1.5, 10, 100}, G={0.01, 0.03, 0.05}) of the model on the development set in respect to the RMSE of the model.

4.2. Experimental Results

For the evaluation of the models both objective and subjective tests were done for evaluating the accuracy of the models and the overall quality of the synthesized speech respectively.

4.2.1. Objective Evaluation

In the objective evaluation, the root mean square error (RMSE) in terms of milliseconds (ms) between the predicted and the reference (the original phone boundaries of the database) phone durations was used. To determine the phone duration prediction in ms using the HSMM model, the sum of frames of each of the five states on each phone was calculated and multiplied by the frame shift of the model. In Table 1, the overall performance accuracy of the three models (HSMM, MLP, SVR) on the development and test sets is presented. The MLP and the SVR models managed to outperform the HSMM one with a relative improvement in terms of RMSE of 11.56% and 25.09%

Table 1: *This table reports the accuracy in terms of RMSE (ms) for the three PDMs for the developement and test sets.*

Set	Phones	HSMM	MLP	SVR
Dev	All	43.89	37.11	**33.07**
Test	All	43.97	38.89	**32.94**
Test	Vowels	48.96	42.13	**33.95**
Test	Cons	39.91	36.31	**31.87**

Table 2: *This table shows the subjective evaluation (ABX test) for the three pairs (HSMM vs MLP, HSMM vs SVR and MLP vs SVR).*

ABX test	***HSMM vs MLP***			***HSMM vs SVR***			***MLP vs SVR***		
Set	**HSMM**	**Eq.**	**MLP**	**HSMM**	**Eq.**	**SVR**	**MLP**	**Eq.**	**SVR**
Set1	26.43%	28.57%	45.00%	15.00%	26.43%	58.57%	16.43%	51.43%	32.14%
Set2	24.62%	32.31%	43.08%	11.54%	34.62%	53.85%	18.46%	41.54%	40.00%
Both	25.56%	30.37%	44.07%	13.33%	30.37%	56.30%	17.41%	46.67%	35.93%

respectively, verifying our first hypothesis, i.e. these external models are able to build more robust models than the HSMM explicit duration modelling. Furthermore, the SVR model in comparison to the MLP model achieved a relative improvement of 15.3%, showing the superiority of the SVR over the MLP model, verifying our second hypothesis, i.e. the SVR could model better the phone durations in comparison to MLP. As it was expected, the SVR model managed to cope in a better way with the high-dimensionality of the feature space in comparison to the MLP model.

Moreover, as it was expected, since the development and test sets are not involved in the training procedure of the HSMM model, the RMSE for these sets are almost identical. In the case of the MLP model on the test set, a 4.79% relative decrease in terms of RMSE in respect to the development set (used for the fine tuning of the model) is achieved, showing some degree of overfitting of the model to the development set. On the other hand, in the case of the SVR model, even though the model is fine tuned using the development set, the RMSE for the development and test sets are almost identical, showing an additional advantage of the SVR over the MLP, i.e. the ability of SVR to make robust model without overfitting to the development set.

In Table 1, the overall accuracy in terms of RMSE on the test set separately for vowels and consonants is presented. It can be seen that these results follow the overall results described above. For all models, the RMSE calculated on the vowels is higher than the one on the consonants, which can be attributed to the fact that the mean of the standard deviation of the vowels (on the reference-original durations of the phones) is significantly higher than the respective one for the consonants.

In Figure 1, the RMSE in milliseconds for each phone for the three models on the test set is presented. It is shown that the SVR model managed to outperform the MLP and the HSMM models for all phones apart from the silence ("pau"), where MLP model achieved the best performance followed by the SVR model. The biggest difference between the SVR and the HSMM models is shown in phone "ey", where SVR model achieved a 37.74% relative improvement over the HSMM model in terms of RMSE. On the other hand the smallest difference for the respective models is shown in phone "hh", where SVR model achieved a 10.9% relative improvement over the HSMM model in terms of RMSE. In the comparison between SVR and MLP models, the biggest difference is shown in phone "ay", where SVR model achieved a 28.36% relative improvement over the MLP model in terms of RMSE. The smallest difference is shown in phone "aa", where SVR model achieved a 1.64% relative improvement over the MLP model in terms of RMSE. Comparing the MLP and HSMM models, it can be noticed that in several cases (e.g. "f", "th", "uw" phones), the MLP model was outperformed by the HSMM model. On the other hand the biggest difference for the respective models is shown in phone "g", where MLP model achieved a 37.76% relative improvement over the HSMM model in terms of RMSE.

4.2.2. Subjective Evaluation

In order to investigate whether the objective performance among the three models is reflected to the overall quality of the synthesized speech, a subjective evaluation test was done. Using the HSMM-based speech synthesis framework described earlier, the phone durations predicted by the SVR and MLP models were forced on the speech synthesis system, in order to synthesize speech using the predicted durations and determining internally the state sequence of the model.

The subjective evaluation was composed by three ABX tests, comparing each of the model to the other two. Two sets of ten sentences where randomly chosen from the test set and were evaluated by 14 and 13 subjects respectively. For every sentence, the subjects were presented with three pairs of samples (HSMM vs MLP, HSMM vs SVR and MLP vs SVR) in random order, without any knowledge about the three systems and a reference sample. In each case the subjects had to choose between the two samples of the pair in terms of sounding closer to the reference one (synthesized with forced alignment using the reference durations) or they could choose "equal" if they had no preference over them.

In Table 2, the results of the ABX tests are presented. As can be seen the SVR model was preferred over the MLP and HSMM models with a score of 35.93% over 17.41% and 56.30% over 13.33% respectively. Furthermore, the MLP model was preferred over the HSMM model with a score of 44.07% over 25.56%. These results follow the trend of the objective evaluation, showing that the SVR managed to build a robust model capable of outperforming the explicit HSMM model and moreover the other external model using MLP, in the overall quality of the synthesized speech.

5. Conclusions

In this paper we compared the HSMM-based explicit duration modelling with two external phone duration models using Support Vector Regression (SVR) and Multilayer Perceptron (MLP). The goal was to investigate whether and in which degree the external duration models outperform the HSMM model and if this is perceived in a subjective evaluation test on the quality of the synthesized speech. The experiments were done on an American English male speaker database. In both objective and subjective tests, it was shown clearly that the external duration models are more appropriate in the task of phone duration modelling in comparison to the explicit duration modelling of HSMMs, verifying our initial hypothesis.

Additionally, between the two external models, the SVR one outperformed the MLP model verifying our second hypothesis that the SVR would managed to tackle this task more efficiently. In the objective test, SVR model managed to outperform the MLP and HSMM ones showing a relative improvement in terms of root mean square error of 15.3% and 25.09% respectively. Finally, the subjective evaluation test showed that the superiority of the SVR model over the MLP and HSMM models is reflected also on the quality of synthetic speech, achieving a preference score of 35.93% and 56.30% over them, respectively. As future work it would be interesting to investigate how these three models perform using larger databases.

6. Acknowledgements

This work has received funding from the Swiss National Science Foundation under the SIWIS project.

7. References

[1] J. Laver, *Principles of Phonetics*. Cambridge: Cambridge University Press, 1994.

[2] T. Dutoit, *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers, 1997.

[3] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Communication*, vol. 50, no. 5, pp. 405 – 415, 2008.

[4] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *EUROSPEECH*, 1999.

[5] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *IEEE ICASSP*, vol. 2, 2001, pp. 805–808.

[6] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," in *EUROSPEECH*, 1997, pp. 2523 – 2526.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 9, pp. 1406 – 1413, 2007.

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. of ICSLP*, 2004.

[9] S. Esmeir, S. Markovitch, and C. Sammut, "Anytime learning of decision trees," *Journal of Machine Learning Research*, vol. 8, pp. 891 – 933, 2007.

[10] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using Deep Neural Networks," in *IEEE ICASSP*, 2013, pp. 7962–7966.

[11] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *The Journal of The Acoustical Society of Japan (e)*, vol. 21, pp. 79 – 86, 2000.

[12] Y. Q. Zhi-Jie Yan and F. K. Soong, "Rich context modeling for high quality HMM-based TTS," in *INTERSPEECH*, 2009, pp. 1755–1758.

[13] Y.-J. Wu and R.-H. Wang, "HMM-based trainable speech synthesis for chinese," *J. Chinese Inf. Process.*, vol. 20, pp. 75 – 81, 2006.

[14] B. Gao, Y. Qian, Z. Wu, and F. K. Soong, "Duration refinement by jointly optimizing state and longer unit likelihood." in *INTERSPEECH*, 2008, pp. 2266–2269.

[15] H. Lu, Y.-J. Wu, K. Tokuda, L.-R. Dai, and R.-H. Wang, "Full covariance state duration modeling for HMM-based speech synthesis," in *IEEE ICASSP*, 2009, pp. 4033–4036.

[16] Y. Ishimatsu, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Investigation of state duration model based on gamma distribution for HMM-based speech synthesis," IEICE Tech. Rep., Tech. Rep. 352, 2001.

[17] H. Siln, E. Hel, J. Nurminen, and M. Gabbouj, "Analysis of duration prediction accuracy in HMM-based speech synthesis," in *Proc. of Speech Prosody*, 2010.

[18] J. Latorre, S. Buchholz, and M. Akamine, "Usages of an external duration model for HMM-based speech synthesis," in *Proc. of Speech Prosody*, 2010.

[19] M. D. Riley, "Tree-based modeling for speech synthesis," in *Talking Machines: Theories, Models, and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, Eds. Amsterdam: Elsevier, 1992, pp. 265 – 273.

[20] Q. Guo, N. Katae, H. Yu, and H. Iwamida, "Decision tree based duration prediction in Mandarin TTS system." *Journal of Chinese Language and Computing*, vol. 17, no. 1, pp. 97–106, 2007.

[21] O. Goubanova and S. King, "Bayesian Networks for phone duration prediction," *Speech Communication*, vol. 50, no. 4, pp. 301 – 311, 2008.

[22] A. Lazaridis, T. Ganchev, T. Kostoulas, I. Mporas, and N. Fakotakis, "Phone duration modeling: overview of techniques andperformance optimization via feature selection in the context of emotional speech," *International Journal of Speech Technology*, vol. 13, no. 3, pp. 175–188, 2010.

[23] A. Lazaridis, I. Mporas, T. Ganchev, and N. Fakotakis, "Support Vector Regression fusion scheme in phone duration modeling," in *IEEE ICASSP*, 2011, pp. 4732–4735.

[24] K. K. Sreenivasa Rao and B. Yegnanarayana, "Modeling syllable duration in indian languages using Support Vector Machines," in *Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on*, 2005, pp. 258–263.

[25] U. Ogbureke, J. Cabral, and J. Berndsen, "Explicit duration modelling in HMM-based speech synthesis using a hybrid hidden Markov model-Multilayer Perceptron," in *SAPA - SCALE Conference*, 2012.

[26] K. S. Rao and B. Yegnanarayana, "Modeling duration of syllables using Neural Networks," *Computer Speech & Language*, vol. 21, no. 2, pp. 282 – 295, 2007.

[27] A. Lazaridis, I. Mporas, T. Ganchev, G. Kokkinakis, and N. Fakotakis, "Improving phone duration modelling using Support Vector Regression fusion," *Speech Communication*, vol. 53, no. 1, pp. 85 – 97, 2011.

[28] A. Lazaridis, T. Ganchev, I. Mporas, E. Dermatas, and N. Fakotakis, "Two-stage phone duration modelling with feature construction and feature vector extension for the needs of speech synthesis," *Computer Speech & Language*, vol. 26, no. 4, pp. 274 – 292, 2012.

[29] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223 – 224.

[30] T. Masuko, "HMM-based speech synthesis and its applications," Ph.D. dissertation, Tokyo Institute of Technology, 2002.

[31] S. Haykin, *Neural Networks: A Comprehensive Foundation, 2nd ed.* New York: Macmillan College Publishing, 1998.

[32] A. Smola and B. Scholkopf, "A tutorial on Support Vector Regression," Royal Holloway College, London, U.K., Tech. Rep. NeuroCOLT Tech. Rep. TR 1998-030, 1998.

[33] H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, second ed.* Morgan Kauffman Publishing, 2005.

[34] J. Platt, "Fast training of Support Vector Machines using sequential minimal optimization," in *Advances in Kernel Methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 185 – 208.

[35] B. Scholkopfand and A. Smola, *Learning with Kernels*. MIT Press, Cambridge, 2002.

[36] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[37] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the Tenth National Conference on Artificial Intelligence*, ser. AAAI'92. AAAI Press, 1992, pp. 129–134.

[38] "HMM-based speech synthesis system version 2.2," 2011. [Online]. Available: http://hts.sp.nitech.ac.jp

[39] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis — a unified approach to speech spectral estimation," in *Proceedings of the International Conference on Spoken Language Processing*, Yokohama, Japan, 1994, pp. 1043–1046.

[40] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *MAVEBA*, 2001.

[41] "Multilayer Perceptron in Wikipedia." [Online]. Available: http://en.wikipedia.org/wiki/Multilayer_perceptron

Variation in Prosodic Boundary Strength: a study on dislocated XPs in French

Elisabeth Delais-Roussarie [1], *Ingo Feldhausen* [2]

[1] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France
[2] Goethe-Universität Frankfurt, Germany /UMR 7018-LPP, Université Paris 3, France

Elisabeth.roussarie@wanadoo.fr, ingo.feldhausen@gmx.de

Abstract

Three independently motivated types of information are usually assumed to influence prosodic boundary placement and to play a role in their relative strength: the morpho-syntactic structure, the information structure and the metrical complexity. The phonetic realization associated with the different boundary types (in particular IP and ip) is also assumed to vary.

Based on data of clitic left-dislocations in French, we argue here that differences in the relative strength of the prosodic boundary occurring at the end of the dislocated XP (i.e. an intermediate (ip) or an intonational phrase (IP) boundary) cannot be derived in a straightforward manner from these three types of information. In a production experiment, where the syntactic and information structure were controlled, while the metrical complexity was varied, the analysis of the data achieved with a semi-automatic tool, ANALOR, showed that the strength of the boundary occurring at the right edge of the dislocated object NP displayed a high degree of variability. In addition, the results indicate a lack of correlation between metrical complexity and boundary strength. The results lead us to argue that a sort of phonological neutralization occurs in certain textual contexts. This neutralization does not allow for distinguishing between intermediate and intonational phrase boundaries in all cases.

Index Terms: prosodic phrasing in French, boundary strength, phonetic realization, clitic left-dislocation (CLLD)

1. Introduction

In most studies dedicated to prosodic phrasing and intonation, an utterance is considered to be segmented into hierarchically organized prosodic constituents (see, among others, [1], [2], [3], [4] und [5]). Three independently motivated types of information are usually assumed to influence prosodic boundary placement and to play a role in the evaluation of the relative strength of the boundary:

- The morpho-syntactic structure, as prosodic phrase boundaries align to designated edges of various syntactic phrases (right or left edge of heads of maximal projections, etc.). See [5], [6], [7] and [8], in which syntax-prosody mapping is expressed in terms of alignment constraints, as well as [3], [4] and [10] for different ways of describing prosody-syntax mapping.
- The information structure, as the topic or the informational focus of an utterance may call for the realization of specific prosodic boundary. See, for instance, the various constraints that account for the alignment of the prosodic phrases to the topic (as in [8],) or focus constituent (as in [11], [12], [13] and [14]).
- The metrical structure, since the size or the metrical structure of a syntactic unit, may influence prosodic phrase placement. In French, for instance, it has been shown that the size of the accentual phrase (or prosodic word) is usually limited to six or seven syllables (see, among others, [7], [15] and [16]). As for Spanish, Catalan and Portuguese, [17] showed that a constituent length of five syllables has an important effect on prosodic phrasing, in contrast to syntactic branchingness.

Even if these various types of information may influence boundary placement, we argue here that they cannot always account for the relative strength of the boundary. A sort of neutralization occurs in some contexts and prevents one from distinguishing between differences in boundary strength, in particular the difference between ip and IP boundaries (see [18] and [19] for a description of these two distinct prosodic constituents in French).

Our proposal is based on the investigation of the prosodic phrasing of clitic left-dislocations (CLLDs) in French, while using a semi-automatic procedure to assign boundary strength. CLLD is an optimal phenomenon for several reasons. First, in terms of syntactic structure, the canonical word order is changed according to the dislocation of a constituent. This yields a specific prosodic pattern, in which the right edge of the dislocated XP is aligned with a prosodic phrase boundary: many (prosodic and syntactic) studies claim that the CLLD constituent is typically demarcated by an IP boundary (e.g. [10], [20]), and some others show that the right edge of CLLDs coincide with either an IP boundary, or an ip boundary (see [8], [13], or [21]). Note however that some studies argued that CLLDs may be prosodically unmarked, or aligned with a lower level boundary such as prosodic word or accentual phrase (see [22] and [23]). Second, with respect to the information structural status, the CLLD constituent is usually considered to be a topic and to be given (in a general sense; see, among others, [23] for a discussion of the interpretative function of CLLD). Third, in the context of metrical structure, the branchingness or the size of the CLLD constituent can vary. We apply this option in the present study. [23] showed that the length of the dislocated XP may play a role in the strength of the prosodic boundary occurring at its right edge (IP, ip, or no real marking which leads to phrase the dislocated XP with the subsequent sentence material). Catalan CLLDs, in contrast, do not show any restructuring effects, according to [8].

Thus, by using CLLD structures, we control for the syntactic structure and the information packaging of the sentence (neither changes), while modifying the metrical structure (expressed in terms of branchingness or length). If a one-to-one correlation exists between the syntax of CLLD and its prosody, no variation in boundary strength (ip or IP) should occur. The same holds for the information status: since the interpretation of the CLLD constituent remains the same, the boundary strength should not differ. In addition, if metrical weight plays a role, boundary strength should be clearly correlated with the (non-) branchingness of the dislocation.

The results, however, show that there is great variation in both boundary strength and phonetic realization in each condition.

The paper is organized as follows. In section 2, the major theoretical issues concerning the representation of prosodic structure and the definition of prosodic phrases in French are given. Section 3 presents the methodology and the corpus used to carry out this research. In section 4, we give a description of the results obtained by analyzing prosodic phrasing in our data. Section 5 consists of a discussion of our findings and presents some perspectives for future research.

2. Background and problematic

In the last few decades, many theoretical studies in prosody have been dedicated to prosodic structure and its organization. Focus has been given to two different issues: the internal organization of the prosodic hierarchy and the criteria involved in the definition of the different prosodic units.

2.1 Prosodic hierarchy and levels of structuring

In most prosodic descriptions, an utterance is considered to be segmented into units that are hierarchically organized. The number of levels above the word level is usually assumed to be either two (e.g. the accentual phrase and the intonational phrase or the phonological/intermediate phrase and the intonational phrase as in [4] and [25]) or three (the accentual phrase, the intermediate phrase and the Intonational Phrase, see, [18] and [26] for French).

Even in studies that are not overtly based in the metrical-autosegmental framework, two or three distinct levels of phrasing above the word level are usually assumed for French (see, among others, [23] and [27]). It is worth mentioning that two authors working on French intonation argue for a different approach: Mertens proposes only a single unit called the *groupe intonatif* ([9], [10]), while Martin proposes that the number of levels of phrasing critically depends on the morpho-syntactic structure of the sentence, [16], [28], [29].

The proposal made here is done within the metrical-autosegmental framework (see [1] and [2]) and relies on the idea that three levels of phrasing are distinguished in French (e.g. the AP, the ip and the IP as in [18], and [26]). We will show, however, that the distinction between ip and IP phrases may be neutralized in some contexts.

2.2 Criteria for the definition of the prosodic constituents

Among the works focusing on the definition of different types of prosodic phrases, a distinction can be made between roughly two categories of work:

- In some works, realizational differences in the prosodic events occurring at phrase boundaries are crucial in distinguishing boundary strength (e.g. [18], [19]). In a certain sense, phonetic and phonological criteria are thus given priority compared to syntax-prosody mapping in the definition of the prosodic phrases.
- In others studies, the definition of prosodic phrases is considered to be constrained by the mapping between the morpho-syntactic, the informational and the metrical structure (see [4], [5], [6] and [8]).

As far as we are concerned, the difference between these two approaches should be limited to a unique perspective: one would hope that both approaches will lead to the same results.

In this work, we assume that differences in boundary strength (here: ip vs. IP) are associated with various prosodic realizations, but we also assume that some categorical differences exist between the different levels of phrasing which can be accounted for by the linguistic criteria explaining boundary placement. Our aim here is thus to determine which criteria come into play in the placement of the two prosodic phrase boundary. This will be done by calculating boundary strength through an automatic procedure in comparable structures.

3. Methodology

To carry out this research, data were gathered by means of a production test. The obtained utterances were then analyzed by using a semi-automatic procedure. The data collection protocol and the analysis procedure are explained in the next sub-sections.

3.1 Corpus and data collection protocol

The experiment was conducted in Paris (France). Ten native speakers of Standard French were recorded. For the present study, six speakers were analyzed, ranging in age from 22 to 29. All subjects were post-graduate students and remained naive as to the purpose of our investigation. The data were recorded as WAV files (16bits, 44.1 kHz) with the Roland UA-55 Quad Capture USB audio interface and the AKG C520 headworn condenser microphone. The speakers were asked to read sentences, for which a context was given.

The utterances to be produced were designed to contain clitic left-dislocations and consisted of simple assertions and questions as shown in (1). While the left-dislocated constituent in the assertions consisted of either one (1a) or two (or three) lexical words (1b), in the questions it consisted of only one lexical word (1c). In order to guarantee the givenness of the left-dislocated constituent (in bold in (1)) and the newness of the core clause, each utterance was preceded by a corresponding context, in which the dislocated element was mentioned, as exemplified in (2) for (1a,b) and in (3) for (1c); cf. the underlined element. The corresponding contexts allow for the classification of the left-dislocated elements as *active topics*, since the latter are active in the speaker's mind and are textually given, i.e. they have just been mentioned in the discourse.

(1a) **La bouteille** Jean-Marie l'a donnée au voisin.
'The bottle, Jean-Marie gave it to the neighbor.'

(1b) **La bouteille de Bordeaux** J.-M. l'a donnée au voisin.
'The bottle from Bordeaux, J.-M. gave it to the neighbor.'

(1c) **Et ce roman** tu l'as déjà lu ?
'And this novel, did you already read it?'

(2) Que s'est-il passé avec <u>la bouteille de Bordeaux</u>? Elle est où?
'What happened to the bottle from Bordeaux? Where is it?'

(3) Je t'ai apporté <u>un roman policier</u>. Je l'ai mis sur la table.
'I brought you a detective novel. I put it on the table.'

In our data, the left-dislocated element was always chosen as to fulfill the function of the object, there being an ongoing discussion of whether sentence-initial subjects in French are actually clear instances of left-dislocation (see [22], and [31]).

There were a total of 144 target sentences (four sentences of type (1a), five sentences of type (1b), three sentences of type (1c), multiplied by two repetitions and six speakers). Additional filler clauses were added.

The subjects were recorded in a quiet room at the Linguistics Department of the University of Paris 7. The stimuli were presented in a pseudo-randomized order on sheets of paper with roughly six target and filler sentences per page. The subjects were told to read the stimuli out loud at a normal rate of speech. Since no sentence-internal punctuation was used, the subjects were told to first read the sentences silently before uttering them aloud. Each recording started with a short practice session.

3.2 Prosodic analysis of the utterances

The data were analyzed in three steps. First, the data were transcribed by the two authors using *praat* [32]. Second, the utterances were automatically segmented into phones, syllables, and graphemic words by means of the speech processing script *EasyAlign* [33]. The obtained segmentations were controlled and corrected when necessary by the authors. Third, the strength of the prosodic breaks associated with the right edge of the dislocated elements was determined by means of the semi-automatic annotation software ANALOR (see [23], [34] and [35] for more details). The software automatically measures the four acoustic parameters (i) relative syllable duration, (ii) relative F0 average, (iii) slope contour amplitude, and (iv) presence of an adjacent silent pause, detects the prominent syllables, and assigns a degree of prominence to each syllable, be it prominent or not. The degree of prominence is calculated on the basis of the parameters that are considered as fundamental for marking prominence in French (see, among others, [23], [34] and [35]). ANALOR's calculations of the degree of prominence rely on two elementary principles, the *quantity principle* and the *compensation principle* (see [22]). The first states that the larger the number of acoustic parameters involved in the detection of a prominence, the stronger the prominence is perceived. The second balances the parameters with one another: if one parameter shows a high prominence score and the other a low one, they are perceived as presenting a medium score together. The results of the calculation are presented in grades ranging from 0 (no prominence) to 10 (high prominence). Since accentuation plays a crucial function in the demarcation of the prosodic units in French (see, among other,[9], [25], [26] and [36]) , we have inferred the boundary strength from the degree of prominence assigned to the last metrical syllable of each word: the higher the degree of prominence, the stronger the break.

Thanks to the software, the detection of the level of prominence for each syllable is robust and we could thus avoid the variation which typically occurs between experts in manual prosodic annotation (see, for instance, [1], p. 288ff). Since accentuation and phrasing are closely related in French (see [9], [23], [25] and [36] among others), we associate the degree of prominence with specific prosodic constituents:

- Level > 2 and < 3 = AP boundary
- Level 3 and 4 = ip boundary
- Level 5 and higher = IP boundary

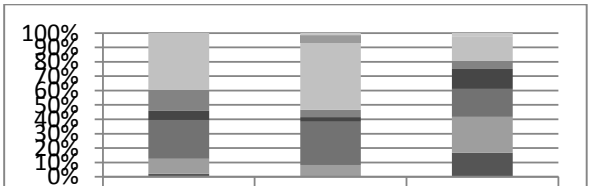
In describing the results in the next section, we rely mostly on the prominence levels. It is only in section 5 that the association between degree of prominence and prosodic constituents (or boundary strength) will be discussed in greater detail.

4. Results

The results indicate that there is generally a prosodic boundary at the right edge of the left dislocated constituent, and its strength forms a gradient ranging from 2 (as > 2) to 9. In addition, this boundary corresponds to an ip or higher boundary in more than 97% of the cases in assertions. In questions, the prosodic boundary is weaker in more than 15% of the cases. In 16,7 % of the cases, it corresponds to an AP boundary.

Table 1 gives the scores of the boundaries at the right edge of the dislocated constituents as detected by ANALOR. Column 1 represents pattern (1a), column 2 (1b), and column 3 (1c). The bars at the top of the table illustrate the percentages given in the table for each score (from 2+ to 9; scores 1 and 10 are not attested in our data).

Table 1. *Prominence of final syllable in CLLD constituent (ANALOR degrees).*



	Decl. 1w	Decl. 2w	Interr. 1w
9	0	1,7	2,7
8	0	5	0
7	39,6	46,7	16,7
6	14,6	5	5,6
5	6,2	3,3	13,9
4	27,1	30	19,4
3	10,4	8,3	25
2	2,1	0	16,7

4.1 Dislocated constituent as a single word in assertions

The most prominent score in the assertive sentences with one lexical word (Decl. 1w) was 7 (39.6%; 19 out of 48 instances). The second highest score for this condition is 4, with 27.1% (13/48). In addition, scores 3, 5, and 6 were also attested more than one time (score 3: 10.4% = 5/48; score 5: 6.2% = 3/48; score 6: 14.6% = 7/48). Score between 2 and 3 is attested only once (2.1%). The percentages of scores ranging from 4 to 7 add up to 87.5% (42/48). Scores 8 and 9 are not attested.

According to these results, IP boundaries occur at the right edge of dislocated object NP in more than 60% of cases, while ip boundaries occur in only 37.5 %. The AP boundary is observed only in 2.1% of the cases.

4.2 Dislocated constituents with two words in assertive sentences

As for the condition with two lexical words (Decl. 2w), all scores between 3 and 9 were attested. Score 7 was by far the most prominent one with a percentage of 46.7 (28/60), followed by score 4 with 30% (18/60). All other scores fell below 10%. In condition 2w, the percentages of scores ranging from 4 to 7 add up to 85% (51/60). In contrast to condition 1w, condition 2w had instances of scores 8 and 9, but no instance of score inferior to 3. Together with the very high percentage for score 7, the data show that the prosodic boundaries in condition 2w are generally slightly stronger than in 1w.

Despite this picture, IP and ip boundaries occur similarly frequently in both conditions. Note that some speakers even realize a stronger boundary in 1w condition than in 2w condition.

4.3 Questions

Table 1 shows that score 3 was most prominent with 25% (9/36); followed by score 4 with 19.4% (7/36), and score between 2 and 3 with 16.7% (6 cases out of36). These three scores add up to a total of 61.1% (22/36). This shows that the general degree of CLLD boundary strength is much lower in questions than in assertions - even though there were six instances of score 7 (16.7%) and 5 instances of score 5 (13.9%).

The interpretation of the results in terms of boundary level shows that an AP boundary occurs in 16.7 % of the cases, while an ip boundary occurs in 44.4 % and an IP in 38.9% of the cases. In comparison to assertive sentences, non-IP boundaries (i.e. AP or ip boundaries) are found to be more frequent in questions.

5. Discussion

Up to now, the results were mostly presented by relying on the ANALOR scores of prominence. In this section, we would like to highlight the interpretation of these scores in terms of prosodic constituents and prosodic structure: accentual phrase (AP), intermediate phrase (ip), and Intonational Phrase (IP).

Before running the experiment, the authors agreed to assign all scores between 2 and 3 to the AP, the scores between 3 and 4 to the ip level, and the scores equal and higher than 5 to the IP level. [22] and [23] made slightly different choices in assigning scores between 2 to 3 (2<score<3) to the AP level, score 3 to the ip level, and everything equal and greater than 4 to the IP level. Despite the difference in the affiliation of score 4, this score always signals that a clitic-left dislocation in French is not obligatorily followed by an IP boundary, but rather can be realized on either the ip or the IP level. Similarly, the prominence of the ultimate syllable in the dislocated constituent is subject to strong variation.

According to the two classifications, our results show that it is only in questions that the CLLD constituent needs not to be prosodically separated from the following sentence by an ip or an IP boundary. In assertions, the CLLD constituent is almost always separated from the sentence by an ip or an IP boundary (just one case of an AP boundary was observed in all the data). The impossibility of having an AP is even clearer with branching CLLD constituents, as these never group with the subsequent material in our data. Interestingly, the boundary at the right edge of the dislocation is the strongest boundary one can find sentence internally: (i) it is the strongest one within the dislocated XP (in case the XP consists of at least two distinct prosodic units, and (ii) it is also the strongest one with respect to the following sentence internal boundaries. This means, independently of the CLLD boundary being an ip or an IP boundary, it is always the strongest sentence-internal boundary. Only the sentence-final boundary might be stronger. The first point is not really surprising and confirmed what was said by [10], [15] and [16] among others. As for the second point, it shows that no restructuring with subsequent elements is possible in French, even when the boundary strength is not very important. To sum up, it is possible to say that a phrase boundary is always realized at the right edge of the CLLD in assertive sentences. The results show however that the mapping between syntax/information structure and prosody is not invariant with respect to the boundary strength, since it is either an ip or an IP boundary. In addition, the great variation in prominence and boundary strength in conditions 1w and 2w clearly indicates a lack of correlation between metrical complexity and phonetic realization of the boundary.

As for the behavior of CLLD in questions, we think that the sentence type may play a role: all questions were declarative questions, which end with a rising tonal contour. So, in order to distinguish this contour from the rising contour occurring after the CLLD, the speaker may realize a weaker sentence-internal rising contour, or even a falling one. These realizations confirm the proposal made in [16], [28] and [29]. They are also in accordance with more recent studies ([37], [38]) which show that global intonational patterns may be more important than the exact strength of a boundary as the "impact of prosodic boundaries depends on the other prosodic choices a speaker has made" ([38], p. 244). In the cases where a non-rising contour has been observed at the right edge of the dislocated XP, the assigned degree of prominence may result from the form of the contour. The calculation, as currently made has a tendency to underweight falling contours. Since such cases are not very frequent, they cannot invalidate the conclusions made here.

6. Conclusion and Perspectives

The analysis of our data shows that the choice between the different boundary levels cannot be systematically explained on the basis of the three criteria usually claimed to account for prosodic phrasing, i.e. syntactic and information structure, and metrical complexity. There is great variation in the realization of boundary strength. To our mind, this could be attributed to a sort of neutralization: in some contexts, the distinction between ip and IP boundaries is not really relevant, since the sentence as a whole does not request to distinguish three levels of prosodic structuring. In order to be the strongest sentence-internal break, the prosodic break occurring at the end of the dislocated XP could be an ip or and IP boundary.

Further research is necessary to understand the exact motivation behind this neutralization. One could for instance take into account CLLDs in embedded clauses in order to address some of the issues. In addition, a more comprehensive study of the phrasing in questions has to be achieved in order to see what motivates the occurrence of AP boundary in this context: is it due to tonal and realizational constraints, or to the expression of some difference in speaker 's attitude.

References

[1] Ladd, R. D., “Intonational Phonology”, 2nd edn, CUP, 2008.
.
[2] Pierrehumbert, J. and Beckman, M., “Japanese Tone Structure”, MIT Press, 1988.
[3] Nespor, M. and Vogel, I., “Prosodic Phonology”, Mouton de Gruyter, (1986: Foris), 1986/2007.
[4] Selkirk, E., “On derived domains in sentence phonology”, Phonology 3: 371-405, 1986.
[5] Selkirk, E., “The syntax-phonology interface”, in J. Goldsmith, J. Riggle and A. Yu [Eds], The Handbook of Phonological Theory, 435-484, 2nd edition, Blackwell, 2011.
[6] Truckenbrodt, H., “On the relation between syntactic phrases and phonological phrases”, Linguistic Inquiry, 30:219-255, 1999.
[7] Delais-Roussarie, E., “ Phonological Phrasing and Accentuation in French”, in M. Nespor and N. Smith [Eds], Dam Phonology : HIL Phonology Paper II, 1-38, Holland Academic Graphics, 1996.
[8] Feldhausen, I., “Sentential Form and Prosodic Structure of Catalan”, John Benjamins, 2010.
[9] Mertens, P., “Accentuation, intonation et morphosyntaxe”, Travaux de Linguistique, 26:21-69, 1993.
[10] Mertens, P. “Syntaxe, Prosodie et structure informationnelle: une approche prédictive pour l’analyse de l’intonation dans le discours”, Travaux de Linguistique, 56:97-124, 2008.
[11] Truckenbrodt, H., “Phonological Phrases: their Relation to Syntax, Focus, and Prominence”, Doctoral dissertation, 1995.
[12] Selkirk, E., “The Interaction of Constraints on Prosodic Phrasing”, in M. Horne [Ed], Prosody: Theory and Experiment, 231-261, Kluwer Academic Press, 2000.
[13] Jun, S.-A. and Lee, H.-J., “Phonetic and Phonological markers of Contrastive Focus in Korean”, in Proceedings of the 5th ICSLP, 4:1295-1298, 1998.
[14] Féry, C., “Focus as prosodic alignment”, To appear in NLLT 31:4.
[15] Martin, P., “Structure prosodique et structure rythmique pour la synthèse”, Actes des 15èmes Journées d’etudes sur la parole, 89-92, 1986.
[16] Martin, P., “Prosodic and rhythmic structure in French”, Linguistics 5(5):925-949, 1987.
[17] Elordieta, G., Frota, S., Prieto, P. and Vigario, M., “Effects of constituent weight and syntactic branching on intonational phrasing in Ibero-Romance”, in M.-J. Sole, D. Recasens and J. Romero [Eds], Proceedings of the 15th International Congress of Phonetic Sciences, 487–490, 2003.
[18] D'Imperio, M. and Michelas, A., “Embedded register levels and prosodic phrasing in French”, Speech Prosody, 2010.
[19] Michelas, A., “Caractérisation phonétique et phonologique du syntagme intermédiaire en français: de la production à la perception”, Thèse de doctorat d'Aix-Marseille Université, 2011.
[20] Delais-Roussarie, E. and Post, B., “Unités prosodiques et grammaire de l’intonation:vers une nouvelle approche”, Actes des Journées d’étude sur la Parole JEP-TALN, 2008.
[21] Astésano, C., Espesser, R. and Rossi-Gensane, N., “Quelques cas particuliers de détachement à gauche –ou la prosodie à l’aide de la syntaxe”, Actes des Journées d’étude sur la Parole JEP-TALN, 2008.
[22] Avanzi. M., “La dislocation à gauche en français parlé. Etude instrumentale”, Le français moderne, 2011(2), 2012.
[23] Avanzi, M., “L'interface prosodie/syntaxe en français”, Peter Lang, 2012.
[24] Vallduví, E., “The Informational Component”, Garland, 1992.
[25] Post, B., “Tonal and phrasal structures in French intonation”, Thesus, 2000.
[26] Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R. and Yoo, H.Y., “Developing a ToBI system for French“, in S. Frota and P. Prieto [Eds], Intonational Variation in Romance, chapter 3, Oxford University Press, to appear.
[27] Di Cristo, A., “Intonation in French”, in D.J. Hirst and A. Di Cristo [Eds], Intonation Systems: A Survey of twenty languages, 195-218, Cambridge University Press, 1998.
[28] Martin, P., “Pour une théorie de l’intonation”, in M. Rossi et al. [Eds], L’Intonation de l’acoustique à la sémantique, 234-271, Klincksieck, 1981.
[29] Martin, P., “Intonation du français”, Armand Colin, 2009.
[30] Delais-Roussarie, E., Doetjes, J. and Sleeman, P., “Dislocations in French”, in C. Francis and H. de Swart [Eds], Handbook of French semantics, 501-528, CSLI, 2004.
[31] Blasco-Dulbecco, M., “Les dislocations en français contemporain. Etude syntaxique”, Champion, 1999.
[32] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer [Computer program]”, http://www.praat.org/, 2013.
[33] Goldman, J-Ph., “EasyAlign: an automatic phonetic alignment tool under Praat”, InterSpeech, 2011.
[34] Avanzi, M., Lacheret-Dujour, A. and Victorri, B., “A corpus-based learning method for prominence detection in spontaneous speech”, Speech Prosody, 2010.
[35] Avanzi, M., Lacheret-Dujour, A., Obin, N. and Victorri, B., “Toward a Continuous Modeling of French Prosodic Structure: Using Acoustic Features to Predict Prominence Location and Prominence Degree”, Interspeech, 2033-2036, 2011.
[36] Post, B., “The multi-faceted relation between phrasing and intonation in French”, in Lleo, C. and C. Gabriel [Eds], Hamburger Studies in Multilingualism 10: Intonational Phrasing at the Interfaces: Cross-Linguistic and Bilingual Studies in Romance and Germanic, pp. 44-74. Amsterdam: John Benjamins, 2011.
[37] Frazier, L., Clifton, C. Jr. and Carlson, K., “Don’t break or do: Prosodic boundary preferences”, Lingua, 114:3–27, 2004.
[38] Frazier, L., Carlson, K. and Clifton, C. Jr., “Prosodic phrasing is central to language comprehension”, Trends in Cognitive Science, 10(6):244-249, 2006.

Acknowledgements

This work was funded by the French Investissements d’Avenir - Labex EFL program (ANR-10-LABX-0083).
We would like to thank Mathieu Avanzi for helping us with Analor, and Fabían Santiago Vargas for help during the recording session.

Tone Modeling Using Stress Information for HMM-Based Thai Speech Synthesis

Decha Moungsri[1], *Tomoki Koriyama*[1], *Takashi Nose*[2], *Takao Kobayashi*[1]

[1]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
[2]Graduate School of Engineering, Tohoku University, Japan

moungsri.d.aa@m.titech.ac.jp, koriyama@ip.titech.ac.jp,
tnose@m.tohoku.ac.jp, takao.kobayashi@ip.titech.ac.jp

Abstract

This paper describes a modeling technique of Thai tones for HMM-based speech synthesis. Tones are important prosodic features for tonal languages including Thai because the phonetically same words but with different tones give different meanings. Although there have been several approaches to improving tone correctness of synthetic speech by considering tone types, another significant factor, stress, was not used explicitly for prosody modeling. We incorporate stress/unstress information into the framework of the HMM-based speech synthesis. Objective and subjective evaluation results show that the use of stress information improves the performance in Thai tone modeling.

Index Terms: HMM-based speech synthesis, tone correctness, stress, context clustering

1. Introduction

In the speech synthesis of tonal languages including Chinese, Vietnamese, and Thai, tone correctness of the synthetic speech is a crucial point, because different tones give different meanings even if the phonetic information of words is the same. In this context, various techniques have been examined to improve tone correctness for Thai speech synthesis. A tone-separated tree structure was introduced for HMM-based Thai speech synthesis to reduce the tone-dependent effects in the context clustering process [1]. Moreover, to capture a variety of shapes of fundamental frequency (F0) contours effectively, phrase-intonation and tone-geometrical features derived from Fujisaki-model were used as the contexts in the HMM-based speech synthesis [2]. Another approach to modeling tone in Thai is the use of Tilt model [3]. The Tilt model was extended for modeling F0 contours in tonal languages and the modified Tilt model for Thai is called T-Tilt model [4]. To enhance T-Tilt model, an optimized T-Tilt model by expanding the Tilt curve over the whole syllable was also proposed [5]. If tone information is determined from transcriptions of speech data in the modeling, the quality of synthetic tones much depends on the accuracy of tone labeling. For this problem, a tone modeling technique using a quantized F0 context was proposed to reduce the tone distortion caused by inconsistent tonal labeling, in which quantized F0 symbols were utilized as the context for constructing the decision trees [6].

Although, these techniques improve the naturalness and tone intelligibility of synthetic speech, tone correctness is not perfect and the tones of some syllables in continuous synthetic speech are perceived to be obviously incorrect and unnatural. To alleviate this problem, we focus on one of other factors, which affects prosodic features and has not been considered explicitly as a context in the conventional HMM-based Thai speech synthesis. In [7], it is reported that stressed syllables have largely different characteristics from unstressed ones. Stressed syllables have typical F0 contours and long durations. In contrast, unstressed syllables are diverse. For example, the effect of speaking rate on the F0 contours of unstressed syllables is more extensive, both in terms of height and slope, than that of stressed syllables [8]. As described in [9], coarticulation and intonation affect tonal assimilation and declination. The coarticulatory effect will cause the change in F0 contour shape of a syllable to be assimilated with neighboring syllables.

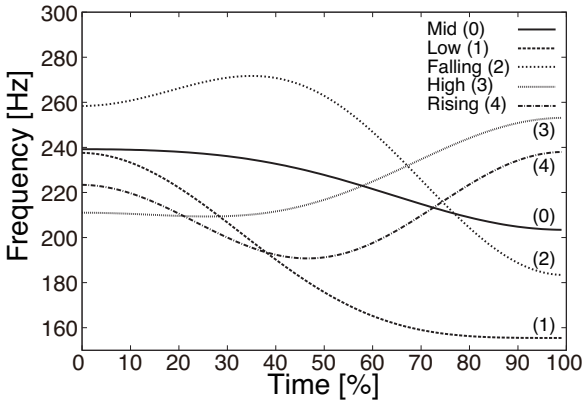


Figure 1: Typical F0 contours of Thai tones. Each contour represents the average of extracted F0 contours from the database TSynC-1.

In this paper, we examine stressed/unstressed effects on Thai tones, and propose a modeling technique using stress information to improve Thai tone correctness of HMM-based synthetic speech. We classify stressed and unstressed syllables manually by using the stress properties described in [7, 10], and incorporate the stress information into the context labeling. We compare the performance of the proposed and conventional methods through objective and subjective evaluations and show the effectiveness of the proposed method.

2. Tone and stress in Thai

Tone represents a change in the pitch of a syllable during its pronunciation. In Thai, every syllable is pronounced in one of five tones: mid (0), low (1), falling (2), high (3), or rising (4).

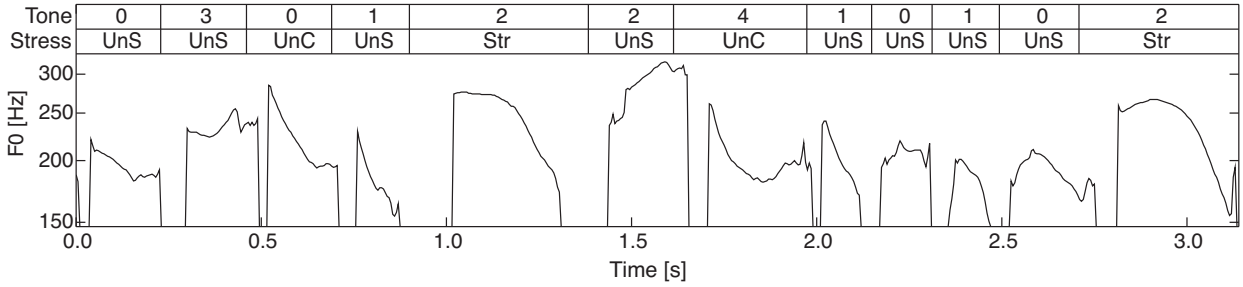
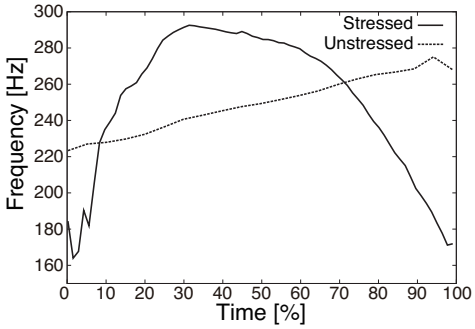
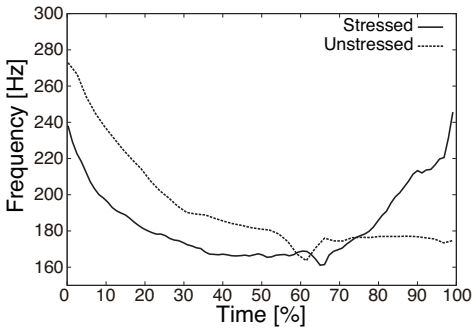


Figure 3: Stressed and unstressed syllables in natural speech (Str : Stressed syllable, UnS : Unstressed syllable, and UnC : Unclear syllable).



(a) Falling tone (2)

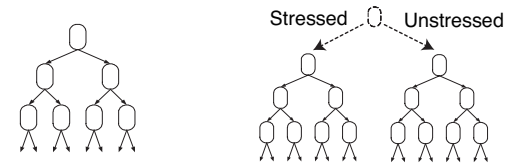


(b) Rising tone (4)

Figure 2: Example of F0 contours in (a) falling tone (2) and (b) rising tone (4) syllables.

The tone must be spoken correctly for the intended meaning of a word to be understood. The identification of a Thai tone relies on the shape of the F0 contour. Figure 1 shows the typical F0 contours of five different tones of isolated syllables. However every F0 contour shape does not always look like the typical one. The shapes depend on stress information of syllables [7]. Figure 2 shows an example of F0 contour shapes of falling tone (2) and rising tone (4) in stressed and unstressed syllables which were extracted from speech samples included in Thai speech database TSynC-1 [11].

The stressed F0 contour shapes are similar to the typical ones, whereas the shapes of unstressed syllables tend to be flat and have less movement of contour, especially in falling tone (2) and rising tone (4). Furthermore, the actual unstressed F0 contours are diverse. Several interacting factors affect F0 realization of tones, e.g., syllable structure, declination, tonal assimilation, stress, and speaking rate [7–10].



(a) Single tree (b) Stressed/Unstressed separated tree

Figure 4: Decision trees for context clustering: (a) single tree structure and (b) stressed/unstressed separated tree structure.

3. Tone modeling with stressed/unstressed context

3.1. Annotation of stress information

As described in [10], stressed syllables have following characteristics:

- Long duation
- F0 contour similar to the prototypes
- High energy

Duration is the predominant feature in the distinction between stressed and unstressed syllables in Thai. The secondary feature is the range of F0 movement [7]. Generally, stressed syllables appear in the end of utterances, isolated phrases, and emphasized words. In other words, the characteristics of unstressed syllables are the opposite ones of stressed syllables.

Figure 3 shows an example of the F0 contour that includes stressed, unstressed, and unclear syllables. The utterance contains 12 syllables, of which the fifth and the twelfth ones are stressed, the third and the seventh ones have unclear stress information, and the others are unstressed.

In this study, we classified stressed and unstressed syllables manually. First, we listened to each syllable individually. If it clearly has the properties of stressed syllables, and it can be classified into only one tone, we annotate a *stressed* label on it. Otherwise, *unstressed* is annotated. However, there are some syllables that are not easy to distinguish, because they are indistinctly uttered by the characteristics between the stressed and unstressed syllables. Currently, we annotate *unstressed* labels for such indistinctly uttered syllables.

Table 1: Average F0 distortion between original and generated speech samples.

Method	RMS error [cent]	# of leaf nodes
Conventional	139.1	2532
Single tree	132.3	2470
Separated tree	132.6	2685

3.2. Context clustering using stress information

We examine two tone modeling methods using stress information. Specifically, we incorporate the stress information into the context clustering that is an essential process in the HMM-based speech synthesis [12]. The first method is adding the stress information to the context set. By incorporating stress information to the context set, the different characteristics are automatically separated during the decision tree clustering. We refer to this method as *single tree*. The second method is to use two trees separated by stressed and unstressed syllables using a manner similar to the tone-separated tree structure proposed in [1]. It is based on the facts that the characteristics of stressed and unstressed syllables are largely different and the frequencies of the stressed and unstressed syllables are imbalanced. In this method, the stress contexts of preceding and succeeding syllables are taken into account based on the results of [1]. We refer to this method as *stressed/unstressed separated tree*. Figure 4 illustrates decision tree structures of the proposed two methods.

4. Experiment

4.1. Training condition

A set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC [11] was used for training and evaluation. The sentences in the database were uttered by a professional female speaker with clear articulation and standard Thai accent with reading style. A speaker-dependent model was trained using 340 utterances from the database. We used 29 utterances for evaluation, which were not included in the training set.

Speech signals were sampled at a rate of 16kHz. F0 and spectral features were extracted by STRAIGHT [13] with 5-ms frame shift. The feature vector of phone-unit HMM consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. We used hidden semi-Markov model (HSMM) which has explicit duration distributions. The model topology was 5-state left-to-right context-dependent HSMM. The conventional method uses context clustering as described in [14]. Context clustering of the proposed method was extended from the conventional one by including stress information in the context set.

4.2. Objective evaluation result

The proposed and conventional methods were evaluated objectively. The measurements for evaluation were average F0 distortions that were calculated by RMS error between generated and original log F0s.

The results are shown in Table 1. The number of leaf nodes of F0 decision trees for context clustering is also shown in the table. The F0 distortions of the proposed methods were lower than the conventional method. However, there was only slight difference between the single tree and stressed/unstressed separated tree. In addition, the numbers of leaf nodes were not much different for all methods. Figure 7 shows an example of the F0 contours generated by all methods compared to the original speech. The sixth and eleventh syllables are the stressed syllable in the falling tone. The F0 contours of these syllables in Fig. 7 (a), which were generated by the conventional method, are not like the falling tone because they do not fall at the end of the syllable, but they are similar to those of the mid tone. In Figs. 7 (b) and (c), the F0 contours were generated by the proposed methods and they are similar to the original ones. The ends of the F0 contours are falling and were perceived as the falling tone. There is a slight difference in F0 contours between the single tree and stressed/unstressed separated tree structure methods.

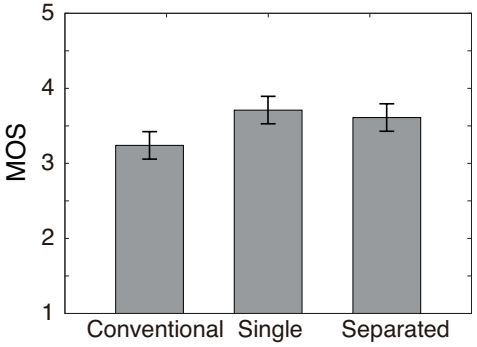


Figure 5: Mean opinion score of naturalness of synthetic tone comparison.

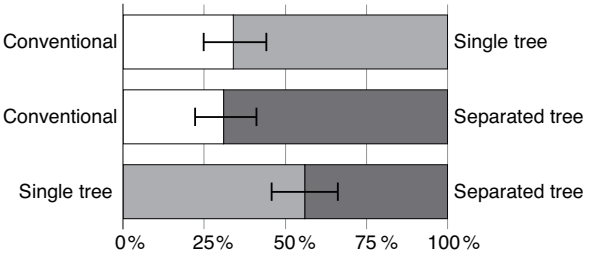
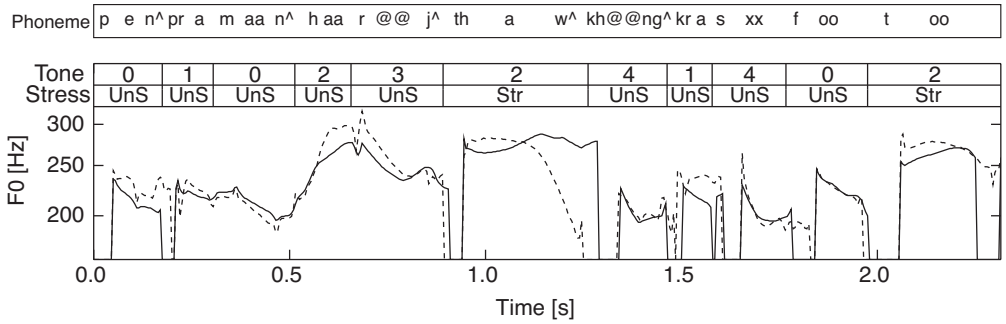


Figure 6: The results of forced choice preference test in tone intelligibility.

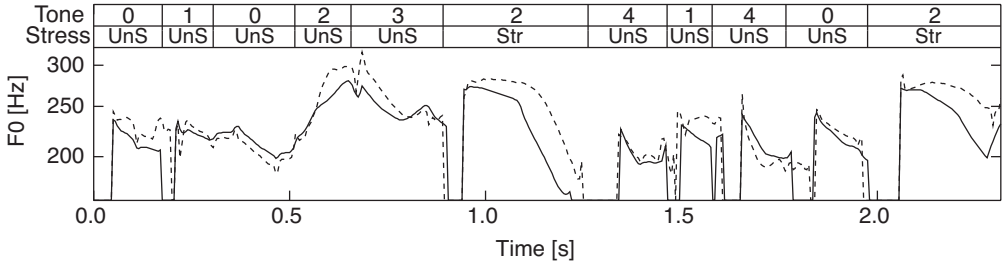
4.3. Subjective evaluation result

To ensure the effectiveness of the proposed methods, we evaluated the perceptual quality in terms of naturalness and tone intelligibility. Specifically, we employed mean opinion score (MOS) and forced choice preference tests.

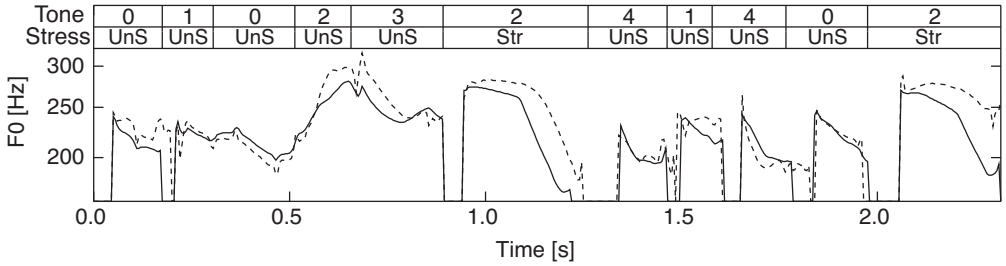
Ten utterances were randomly chosen from the synthetic speech samples used in the objective evaluation test. We assessed the synthetic speech from the proposed two methods and the conventional method. As a result, we compared three types of synthetic speech in the evaluation. Ten Thai native speakers listened to and evaluated the samples. In MOS test, the listeners evaluated each utterance on a five-point scale from 1 to 5 according to their satisfaction with the perceptual naturalness of tones. The definition of the rating was: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. Listeners could repeat sentences to evaluate as many times as they required for ensuring that they were accurately evaluating. Figure 5 shows the result with



(a) F0 contours comparison between the original one in dashed line and the conventional one in soild line.



(b) F0 contours comparison between the original one in dashed line and the single tree one in soild line.



(c) F0 contours comparison between the original one in dashed line and the separated tree one in soild line.

Figure 7: Examples of F0 contours generated by (a) the conventional method, (b) the single tree method, and (c) the separated tree method. The sentence means that "it is about 500 times of the photoelectric." The tone numbers represent mid tone (0), low tone (1), falling tone (2), high tone (3), and rising tone (4).

95% confidence intervals. It can be observed that the proposed methods outperformed the conventional method. The scores of both proposed methods are not significantly different and it is consistent with the result of objective evaluation.

In the forced choice preference test, the listeners were asked to choose more natural-sounding tone from each pair of synthetic speech. The listener could repeat sentences as many times as they required in the same way as the MOS test. The results of the forced choice preference test are shown in Fig. 6. It is seen again that the proposed methods outperformed the conventional method. When comparing between the single tree structure method and the stressed/unstressed separated tree structure method, the listeners preferred the single tree method, but the difference is statistically not significant.

5. Conclusion

This paper has described a modeling technique of Thai speech synthesis using the stress information of syllables. Although stress is an important factor for tone perception in Thai, stress information has not been included in the context clustering in conventional techniques. This causes generation of incorrect tones. To overcome the problem, we added stress information to context clustering. The objective evaluation showed that the proposed method can reduce the F0 distortion significantly. The subjective tests also yielded results that corresponded to those from the objective tests. Although we have confirmed that the stress information could improve the tone correctness, there still exist unnatural tones in synthetic speech. For improvement of the tone naturalness, syllable-level unit might not be appropriate. Thus, in future work, we will investigate a tone modeling technique based on longer unit such as word or phrase.

6. Acknowledgements

We would like to thank Dr. Vataya Chunwijitra of NECTEC, Thailand, for providing us with helpful discussion and the TSynC-1 speech corpora. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071.

7. References

[1] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis," *Speech Communication*, vol. 51, no. 4, pp. 330–343, 2009.

[2] S. Chomphan and T. Kobayashi, "Incorporation of phrase intonation to context clustering for average voice models in HMM-based Thai speech synthesis," in *Proc. ICASSP*, 2008, pp. 4637–4640.

[3] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *J. Acoust. Soc. Am.*, vol. 107, p. 1697, 2000.

[4] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, A. Rugchatjaroen, and S. Saychum, "T-Tilt: a modified tilt model for F0 analysis and synthesis in tonal languages," in *Proc. INTERSPEECH*, 2008, pp. 2270–2273.

[5] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol, and C. Wutiwiwatchai, "Optimization of T-Tilt F0 modeling," in *Proc. INTERSPEECH*, 2009, pp. 508–511.

[6] V. Chunwijitra, T. Nose, and T. Kobayashi, "A tone-modeling technique using a quantized F0 context to improve tone correctness in average-voice-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 245–255, 2012.

[7] S. Potisuk, J. Gandour, and M. Harper, "Acoustic correlates of stress in Thai," *Phonetica*, vol. 53, no. 4, pp. 200–220, 1996.

[8] J. Gandour, A. Tumtavitikul, and N. Satthamnuwong, "Effects of speaking rate on Thai tones," *Phonetica*, vol. 56, no. 3-4, pp. 123–134, 1999.

[9] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Tone recognition in Thai continuous speech based on coarticulaion, intonation and stress effects," in *Proc. INTERSPEECH*, 2002.

[10] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Stress and tone recognition of polysyllabic words in Thai speech," in *Proc. Int. Conf. Intelligent Technologies*, 2001, pp. 356–364.

[11] C. Hansakunbuntheung, A. Rugchatjaroen, and C. Wutiwiwatchai, "Space reduction of speech corpus based on quality perception for unit selection speech synthesis," *Proc. SNLP*, pp. 127–132, 2005.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[14] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," in *Proc. INTERSPEECH*, 2007, pp. 2849–2852.

Understanding the significance of different components of mimicry speech

D. Gomathi[1]*, P. Gangamohan*[2] *and B. Yegnanarayana*[3]

Speech and Vision Laboratory, International Institute of Information Technology,
Gachibowli, Hyderabad, India 500032

gomathi@research.iiit.ac.in[1], gangamohan.p@students.iiit.ac.in[2], yegna@iiit.ac.in[3]

Abstract

Voice conversion systems aim at finding a transformation function using statistical models. Mimicry (voice imitation) is a natural voice transformation technique which sounds convincing to the listeners. It thus seems advisable to study the transformation used by human beings who perform mimicry. The objective of this study is to examine the various components of speech that are modified during voice imitation. To transform a given speech utterance to sound like that of a target utterance, the process needs to be understood at both production and perception level. In this paper, the importance of source and system parameters and also the significance of different components of speech that contribute to the perception of imitation are studied. A flexible analysis-synthesis tool is used to modify the features of natural utterance and convert it to imitated utterance. Perceptual studies are carried out to understand if the modified features contribute to imitation. The results show that a combination of features is varied by the imitator to achieve imitation and they vary depending on the target speaker.

Index Terms: Mimicry, voice imitation, speech synthesis, speech prosody.

1. Introduction

Speech is a natural medium of communication among human beings. Speech signal carries information about the intended message, the identity of the speaker and the background. There are a few features in speech signal that enable us to differentiate between speakers. There are applications in gaming industry which require voices of celebrities for gaming avatars. Voice Conversion (VC) is a technique to transform an utterance of a source speaker so that it is perceived as if spoken by a specific target speaker. A variety of techniques have been proposed for the conversion function [1], such as artificial neural networks, dynamic frequency warping or Gaussian mixture model.

Human beings imitate voice for language acquisition, entertainment or for voice disguise [2]. Voice imitation is primarily used for entertainment, where the mimicry artist trains his voice to imitate the voice of a target speaker. An imitator cannot imitate all the features of the target speaker but tries to imitate features that are perceptually significant.

In the literature, analysis of voice imitation has been carried out and the closeness of features like fundamental frequency, duration, discrete Fourier transform spectra (DFT) and formant frequencies have been studied [3]. Different studies of professional imitators and their imitations suggest the possibility of getting close to target voice. Both perceptual and acoustic analysis confirm the flexibility of human voice [2]. It is also to be noted that an imitator might exaggerate a few important features while he may ignore a few less important features [4].

Mimicry is a natural voice transformation technique which takes into account the ignored aspects of speech synthesis research. Studying the way a professional imitator imitates will help in building better voice conversion systems. For voice transformation/conversion, the cues underlying a particular voice quality need to be identified and represented. It is not sufficient to just represent them; but they need to be modified in such a way that modified speech signal sounds natural. It is more likely that more than one feature is modified during imitation and these features vary depending on the target speaker. It is also possible that two different imitators may choose different features for the same target speaker. In this paper, the components of speech that are modified by an imitator are addressed by signal processing techniques. In this study, the speech data from a professional imitator who performed mimicry of various celebrities has been used [5]. The first study addresses the importance of source and system parameters in voice imitation. The second study examines the features modified by an imitator, few experiments are carried out to synthetically transform the natural utterance spoken by an imitator to the corresponding imitated utterance by varying different features of speech. This is carried out using a flexible analysis-synthesis tool (FAST) [6]. This tool was used to transform a neutral utterance to an emotional utterance and vice versa. In FAST, two utterances spoken by the same speaker are matched using dynamic time warping (DTW) algorithm to get two warping paths. These warping paths are used for understanding the way different features of speech are modified. The features correspond to both source and system characteristics of speech production mechanism.

The terminology used in the paper is similar to the one used in [7]. The utterance spoken by the Indian celebrity (actor) will be referred to as target (T). The utterance spoken by the imitator, when he imitates the target (celebrity), will be referred to as imitation (I). The utterance spoken by the imitator in his original voice will be referred to as natural (N). The terms mimicry and imitation are used interchangeably in this paper.

The paper is organized as follows: Section 2 discusses the data and the features of speech that will be used for modification of natural utterance. In Section 3, the following experiments are conducted: (i) to understand the significance of source and system parameters in imitation and (ii) to convert a natural utterance to an imitated utterance. The results of subjective studies on the experiments are also discussed. Section 4 gives the summary and conclusion.

2. Data and Feature Extraction

Database for the current study consists of recordings by a professional mimicry artist in Telugu language [5]. Voices of five popular Indian celebrities (MB, NG, PO, PR and SP) were chosen as target. These voices were collected from interviews and

movies. For each target voice, ten utterances of short duration were chosen. Utterances of short duration do not contain many prominent prosodic features, and the imitator has to be very good to imitate such utterances. All the target utterances were imitated by the professional imitator five times. Recording of the utterances was done in his natural voice as well. There are three parallel utterances corresponding to target (T), imitation (I) and natural utterance (N) in the database.

The source filter theory of speech states that speech can be described as the output of sound source being modulated by a dynamically varying filter. The speech signal carries information about the dynamic vocal tract system and the excitation source. The vocal tract system and excitation source parameters represent inherent characteristics of the speech signal, while duration and intonation are examples of acquired characteristics over a period of time. The features that are extracted for this study are vocal tract features, excitation source features and prosodic features.

The linear prediction coefficients (LPCs) are used to represent the vocal tract information. They are obtained using LP analysis method. The basic idea is that a speech sample at time instant n, can be approximated as a linear combination of the past p, speech samples.

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \tag{1}$$

where s[n] are speech samples, $\{a_k\}$ are the predictor coefficients and $\tilde{s}[n]$ are the predicted samples.

An all-pole filter H(z) is used to represent the vocal tract parameters of the speech signal in the frequency domain.

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2}$$

The error in prediction is given by

$$e[n] = s[n] - \tilde{s}[n] \tag{3}$$

The representation in frequency domain is given as

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

As A(z) is the reciprocal of H(z), LP residual is obtained by the inverse filtering of speech.

The LP residual obtained from LP analysis is used for the representation of excitation source information. The successive samples in the LP residual are less correlated compared to the samples in the speech signal.

Prosodic features are represented by the pitch contour and duration parameter. The pitch contour which is modified by the imitator is derived using zero frequency filtering (ZFF) method [8]. The method involves passing the differenced speech signal twice through a digital resonator having poles at zero frequency. The trend in the output is removed by local mean subtraction using a window length in the range of one to two pitch periods. The negative to positive zero-crossings in the zero frequency filtered output give the glottal closure instants or epochs. The reciprocal of the interval between two successive epochs gives the instantaneous fundamental frequency.

Duration parameter of natural and imitated utterances are mapped frame-wise using dynamic time warping (DTW) algorithm. DTW is an algorithm used for measuring optimal match between two utterances which may vary in time or speed. The utterances are represented by a sequence of vectors which correspond to the vocal tract features. The DTW algorithm is constrained as the labelling of data is done phoneme-wise.

3. Understanding the relative importance of components of speech in imitation

3.1. Significance of source and system parameters

During imitation, the imitator tries to position his articulators in some specific way in order to imitate a few target speakers. Though there are physiological constraints on the vocal tract, there is some flexibility in positioning tongue and some articulators. This brings the changes in his system (vocal tract) parameters. The imitator has to modify the way he excites his vocal folds to produce some of the voice characteristics. This brings the changes in the excitation source characteristics.

A study was performed to understand the importance of source and system parameters in performing voice imitation. A 10^{th} order short-term (20 ms frame size and 10 ms frame shift) LP analysis is performed to compute the residual signal and LP coefficients. The LP coefficients are converted to 20 dimensional linear prediction cepstral coefficients (LPCCs). The LPCCs and residual are extracted for 'T','I' and 'N'. The speech signals are time aligned using dynamic time warping (DTW) with LPCCs as feature vectors. For synthesis, the residual of the imitated utterance is passed through LP filter corresponding to the system parameters of the natural utterance. All combinations of residual and LP coefficients of 'T', 'I' and 'N' of all celebrities (MB, NG, PO, PR, SP) were used for synthesis to know the importance of source and system parameters.

Table 1: *Subjective evaluation results for all combinations of source and system parameters of 'T', 'I' and 'N'.*

Experiment	Source	System	MB	NG	PO	PR	SP
E1	I	T	1	1	1	1	1
E2	T	I	1	1	1	1	1
E3	N	T	1	0	0	0	1
E4	T	N	0	1	1	1	0
E5	N	I	1	0	0	0	0
E6	I	N	0	1	1	0	0

The synthesized files obtained after interchanging the corresponding source and system features for all cases mentioned above are assessed by subjective evaluation. The evaluation is carried out by twenty listeners in the age group of 21-30. Each subject was given six synthesized files and asked to give a score of '1' if it is target (T)/ imitation (I), '0' if it is natural (N). The results of the evaluation are presented in Table 1. The scores in the table are arrived by majority voting. All the synthesized speech files were presented in random order, and were not grouped in any particular order.

The rows E1 and E2 show that when source parameters belong to 'I' and the system parameters belong to 'T' or vice versa, the synthesized speech sounds similar to target for all celebrities.

For celebrity MB, when system parameters of 'T' or 'I' are used, the synthesized file sounds closer to 'T' as seen from rows E3 and E5. When system parameters of 'N' are used, the synthesized file sounds like an unknown speaker. The listeners reported that the characteristic pause of 'T' was missing hence the synthesized speech sounds like unknown speaker. So the system parameters seem to play a bigger role in this case.

In the case of celebrity 'NG', the voice quality is breathy. So whenever source parameters of 'T' or 'I' are used, even if the system parameters belong to 'N', there is breathiness in the synthesized speech which gives an impression that we are listening to 'T' or his imitation. This can be observed from rows E4 and E6.

The source parameters play an important role in the case of celebrity 'PO'. This is because there is an increase in loudness when the source features of 'T' are used. The listeners could make out the difference clearly between the experiments where source parameters of 'T' or 'I' were used. The effect of source features of 'I' is similar to 'T' in terms of intonation but the level of loudness is low. The use of source parameters of 'N' makes the synthesized file sound like 'N' or unknown speaker.

The results of celebrity 'PR' is similar to that of celebrity 'PO', except for row E6. This may be because the source parameters in imitation 'I' are not well imitated in case of celebrity 'PR'.

The imitations of celebrity 'SP' is similar to target for rows E1, E2 and E3. The synthesized files for rows E4, E5 and E6 sound like 'N' or unknown speaker. The expectation is when source or system parameters of 'T' or 'I' are used, the synthesized file should be similar to 'T' or 'I', but the files sounds like an unknown speaker. This may be because the imitations of celebrity 'SP' were not well imitated.

3.2. Perceptual significance of features

The aim of this study is to understand the features that are modified during imitation. The imitator's natural voice and imitation are compared and the differences in features are studied. The features from the imitated voice are incorporated into the natural utterance of the imitator so that imitated voice can be synthesized from natural voice. To modify the natural utterance in order to make it sounds similar to imitated utterance, a flexible analysis-synthesis tool (FAST) has been used. The main feature of FAST is that it can be used to match two utterances of same lexical content spoken by same speaker to determine the warping path (WP). After time alignment, modification of features is carried out as per the warping path. The modified features are then used to synthesize the imitated utterance. The synthesis is carried out using prosody modification program [9].

A 10^{th} order LP analysis is performed on a speech segment of 20 ms for every 10 ms. The 11 LP coefficients are converted to 20 dimensional LPCCs. Time alignment of the natural utterance with the imitated utterance is carried out using the DTW algorithm. The optimal warping path obtained by DTW represents the best mapping between the natural and imitated feature vectors. Two warping paths are obtained for each pair (natural and imitated) of utterances. Warping path 1 (WP1) corresponds to the one in which all frames of the imitated utterance are used. Usage of WP1 will automatically modify the duration. Warping path 2 (WP2) corresponds to the usage of all frames of natural utterance. Figures 1 and 2 show the warping paths WP1 and WP2. Pitch contour of the natural utterance is mapped with that of imitated utterance using the warping path 2 (WP2), as shown in Figure 3. Pitch and duration are modified using the prosody modification program [9]. In the case of LPC modification, the LPCs of each frame of natural utterance are replaced by the LPCs of imitated utterance. For LP residual modification, the residual between two epochs is replaced by an Liljencrants-Fant (LF) model estimate [10]. The set of experiments described in [11] were performed. The features modified and the warping paths used are mentioned in Table 2, '0' in the feature column

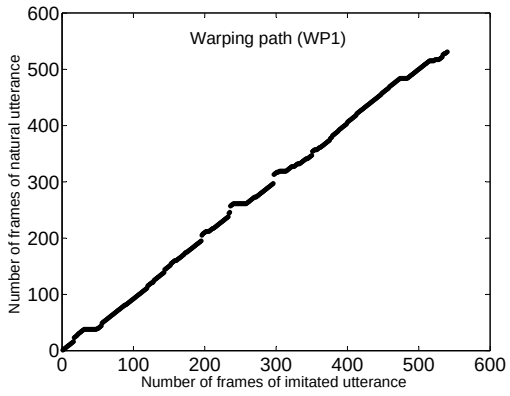


Figure 1: *Illustration of warping path (WP1) when imitated utterance is reference vector and natural utterance is test vector.*

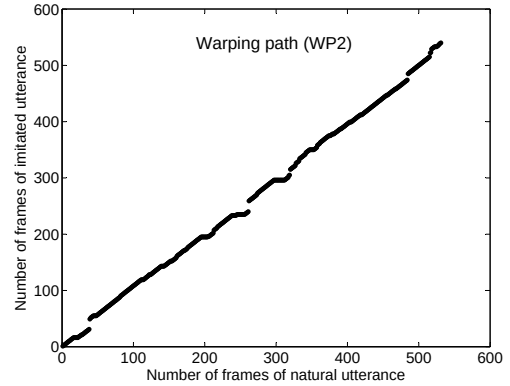


Figure 2: *Illustration of warping path (WP2) when natural utterance is reference vector and imitated utterance is test vector.*

indicates that the feature is not modified and '1' indicates that the corresponding feature is modified in this synthesis experiment.

Table 2: *Experiments and corresponding warping paths for modification of features of natural utterance.*

Experiment	Feature				Warping Path
	Residual	LPC	Duration	Pitch	
E1	0	0	0	1	WP2
E2	0	0	1	0	WP1
E3	0	1	0	0	WP2
E4	0	0	1	1	WP1
E5	0	1	1	0	WP1
E6	0	1	0	1	WP2
E7	0	1	1	1	WP1
E8	1	1	1	1	WP1

There are 10 imitated and 10 natural utterances for each of five celebrities in the database. For each utterance all the eight experiments listed in Table 2 are conducted. Ten listeners participated in the listening test to evaluate the synthesized speech

Table 3: Subjective evaluation results of synthesized imitated utterance.

Experiment	No. of features modified	Feature modified	MB-I	MB-T	NG-I	NG-T	PO-I	PO-T	PR-I	PR-T	SP-I	SP-T
E1	1	Pitch	2.89	1.71	3	2.71	2.85	1.71	3	2.28	2.67	2.28
E2	1	Duration	1.77	1.42	1.85	1.71	1.85	1.57	1.42	1.31	2.67	2.28
E3	1	LPC	1.97	1.32	1.85	1.14	1.85	1.28	1.71	1.71	1.83	1.28
E4	2	Pitch, Duration	2.78	2.31	2.85	2.42	2.71	2.28	3.42	2.14	2.5	2.5
E5	2	Duration,LPC	1.75	1.3	2.14	1.42	1.85	1.85	2.14	1.85	2.33	2.42
E6	2	Pitch, LPC	2.78	2.45	3.42	3.28	3.57	2.85	3.14	2.71	2.5	2.42
E7	3	Pitch, Duration, LPC	2.67	2.71	3	3	3.42	2.85	3.85	2.85	3	2.14
E8	4	Pitch, Duration, LPC, Residual	3.02	2.67	3.28	2.85	3.57	3.37	3.57	3.14	3.5	2.85

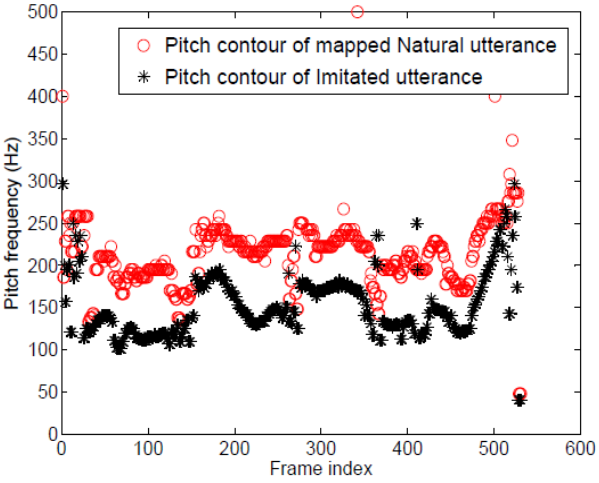


Figure 3: *Mapping of instantaneous pitch contour of natural utterance to that of imitated utterance.*

obtained after modification. The best imitated utterance of each celebrity is considered for subjective evaluation. Each subject is given a natural utterance, an imitated utterance, a target utterance and eight modified utterances which are synthesized by experiments 1 to 8 (presented in random order). The listener has to compare each synthesized utterance to imitated utterance and target utterance and give a score on a scale of 1-5 (1: highly dissimilar, 2: dissimilar, 3: somewhat similar and somewhat dissimilar, 4: similar, 5: highly similar). For example, if the synthesized utterance is compared to imitated utterance, a score of 5 indicates that synthesized file is very similar to imitated, while a score of 1 indicates synthesized file is very different from imitated. The results presented in Table 3 are the mean scores of all 10 listeners. The comparison of synthesized utterance to 'I' is performed and scores are presented in columns MB-I, NG-I, PO-I, PR-I, SP-I. Similarly the comparison between synthesized utterance to 'T' is performed and scores are presented in columns MB-T, NG-T, PO-T, PR-T, SP-T. It is expected that the scores for similarity of synthesized file to imitated utterance will be higher than the scores for similarity of synthesized file to target utterance.

The following observations are made from Table 3. The rows E1, E2 and E3 correspond to modification of one feature at a time namely pitch, duration and LPCs. The high scores in E1 indicates that pitch is a major suprasegmental feature that an imitator can modify easily and contributes more to the perception of imitation. The rows E2 and E3 show us that modification of duration and LPCs alone do not contribute as significantly as pitch modification. The rows E4, E5 and E6 correspond to modification of two features at a time. The rows E4 and E6 where pitch is modified along with duration and LPCs has better scores than E5 in which pitch is not modified. The row E4 gives lower scores for all targets except PR. This might be because duration modification is not aiding the feature pitch in perceiving imitation. The row E6 where pitch and LPCs are modified, shows significant high scores for 'NG','PO' and 'PR'. The combination of pitch, duration and LPCs seem to give significant improvement in the synthesized imitated utterance as can be seen from E7 especially for voices of 'PR' and 'SP'. E8 corresponding to the case where LP Residual is replaced by an LF model has higher scores. Though the residual is absent in this case, the perceptual scores are still high for celebrity 'MB' and 'SP'. In section 3.1, it was shown that system parameters play a big role in the imitation of celebrity 'MB', hence the absence of residual does not seem to affect the perceptual scores. The combination of pitch and LPC features give high scores for celebrities 'NG' and 'PO' but the combination of pitch, duration and LPC gives high score for celebrity 'PR'. The above results show us that the combination of features vary as per the target speaker.

4. Summary and Conclusion

In this paper, the various features of speech that contribute to the perception of imitation have been studied. The first study was to identify the contribution of source and system parameters. The subjective evaluation by listeners confirmed that source parameters were important for target speakers like 'NG' and 'PO' while system parameters were important for target 'MB'. The second study was modification of excitation source, vocal tract and prosodic features. The modification was performed using flexible analysis-synthesis tool. The synthesized files were evaluated for their closeness to imitation and target. The prosodic feature pitch contour seems to play a major role in contributing to the perception of imitation. Though duration and linear prediction coefficients individually do not contribute much to imitation but their combination along with pitch contour gives a good amount of similarity to imitation. The above observations are general ones but the combination of features also vary with the target speaker. The same combination of features need not give high perceptual scores for all target speakers. Further studies can be carried out by collecting mimicry speech from many professional artists to examine whether same features are modified by all of them for a given target speaker.

5. References

[1] Y. Stylianou, "Voice transformation: A survey", *Proc. Int. conference on acoustics, speech, and signal processing (ICASSP)*, Taipei, Taiwan, pp. 3585-3588, April 2009.

[2] E. Zetterholm, " *Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success*", Doctoral dissertation, Travaux de linstitut de linguistique de Lund 44, Lund University,2003.

[3] Tatsuya Kitamura, "Acoustic Analysis of Imitated Voice Produced by a Professional Impersonator", *Interspeech*, pp. 813 – 816, September 2008.

[4] E. Zetterholm, "Detection of speaker characteristics using voice imitation", *In C. Mller and S.Schtz (eds.) Speaker Classification, Springer LNCS/LNAI series*, 2006.

[5] D. Gomathi, Sathya Adithya Thati, Karthik Venkat Sridaran and B. Yegnanarayana, "Analysis of mimicry speech", *Interspeech*, Portland, USA, September 2012.

[6] P. Gangamohan, V.K. Mittal, and B. Yegnanarayana, "A Flexible Analysis Synthesis Tool (FAST) for studying the characteristic features of emotion in speech", *Proc. 9th Annual IEEE Consumer Communications and Networking Conference - Special Session Affective Computing for Future Consumer Electronics*, Las Vegas, USA, pp. 266-270, 2012.

[7] Gal Ashour and Isak Gath, "Characterization of Speech during Imitation", *Eurospeech99*, Budapest, Hungary, September 1999.

[8] K.S.R. Murthy and B.Yegnanarayana, "Epoch Extraction from speech signals", *IEEE Trans. Audio,Speech, Lang Process.*, vol.16, no. 8, pp. 1602 – 1613, November 2008.

[9] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, pp. 972–980, May 2006.

[10] G. Fant, J. Liljencrants and Q. Lin,"A four-parameter model of glottal flow", *Quarterly Progress Status Report*, Speech Trans. Lab., KTH-Sweden, vol. 26, no. 4, pp. 001-013, 1985.

[11] P. Gangamohan, V.K. Mittal and B. Yegnanarayana,"Relative Importance of Different Components of Speech Contributing to Perception of Emotion", *Proc. Int. Conf. Speech Prosody*, pp. 657–660, May 2012.

The Cartoon Task – Exploring Auditory-Visual Prosody in Dialogs

Hansjörg Mixdorff [1], *Angelika Hönemann*[1], *Grégory Zelic*[2], *Jeesun Kim*[2], *Chris Davis*[2]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] MARCS Institute, University of Western Sydney, Australia

[mixdorff:ahoenemann]@beuth-hochschule.de, [G.Zelic J.Kim:chris.davis]@uws.edu.au

Abstract

This paper introduces and analyses a collaborative task for eliciting auditory-visual dialogs based on the viewing of two versions of the same cartoon film. The original film was edited and cut in such a way that the story must be reconstructed by joining information from two incomplete versions which however share between them all the scenes in a consecutive fashion. Our intention is to elicit a relatively balanced dialog between the two participants throughout the conversation as they are piecing together the story from the beginning to the end. The current paper describes the production of the auditory-visual corpus using audio, video and motion capturing of 22 pairs of Australian English speaking participants, and presents first results regarding turn-distribution and raw prosodic features. Our analysis shows that the task is indeed relatively balanced between talkers though this does not apply equally to all pairs. Analysis of raw prosodic features does not suggest convergence throughout the conversation, but replicates, for instance earlier findings of similarity between partners as compared to others.

Index Terms: auditory-visual prosody, dialog, turn-taking, F0, intensity, entrainment

1. Introduction

It is well established that seeing a talker (visual speech) influences auditory speech processing. Typically, research has focused on the perception of segmental information and has demonstrated that visual speech facilitates speech perception [1]. Indeed, the McGurk effect shows that information processing from the two senses is strongly connected and conflicting cues are resolved to form the most likely percept [1]. It has also been shown that the provision of visual speech can improve the perception of lexical tone in noise [3]. Moreover, recent research we have conducted suggests that visual speech influences the perception of speech prosody in interesting but possibly complex ways [4]. This work was based upon a corpus of spontaneous Auditory-Visual A/V monologs that was collected and annotated in terms of both acoustic as well as the visual properties. In addition, motion capture data was recorded and evaluated for non-verbal gestures.

In the analysis of this corpus, which involved the alignment of acoustic landmarks such as accents and boundaries with visible non-speech movements, a question arose as to which way the anchoring of movements should be achieved. In an initial approach only movements that occurred during accented syllables or syllables preceding a boundary were taken into account. However, this left a number of movements unanchored, where, for instance, these were located in syllables neighboring accented syllables. In order to determine how the alignment of acoustic and visual cues reinforce the perceived prominence of the same underlying syllable(s), and when separate events of prominence are perceived, a perceptual rating experiment was designed in which the distance between auditory and visual cues for prominence was systematically varied [5]. The results of this work were in good agreement with a separate production study that examined the timing of head and eyebrow movement with respect to the expression of corrective focus [6].

At this stage, however, it is unclear how the results of the above controlled experimental studies applies to spoken dialogus, since a limitation of corpus collected in [4] was that it only consisted of monologs that had been delivered to a (mute) listener. Plausibly, non-verbal gestures may play an important role in structuring dialogues, so we decided to collect a corpus of spontaneous dialogs in order to examine more closely how non-verbal gestures facilitate discourse and interact with prosodic cues (e.g., in negotiating turn exchanges).

In the current study we examine this corpus with respect to the balancedness of speaker contributions. As a first application of the data we explore the effect of entrainment, the phenomenon that talkers engaged in a dialog adjust their speech to one another, e.g., such as synchronizing (turn-by-turn coordination between interlocutors), or where speech properties become more alike, that is, the talkers attain convergence [7].

The remainder of this paper is structured as follows: In Section 2 we introduce the cartoon task and the collected corpus. Section 3 presents statistical results based on the structures of the resulting dialogs. Section 4 discusses analyses and preliminary results regarding the prosodic entrainment between the participants in the dialog, as reflected by their *F0* and intensity contours, as well as their voice quality. Section 5 offers discussion and conclusions.

2. Experiment Design and Corpus

A large number of different paradigms exist for eliciting spontaneous dialog data. These paradigms range from completely unrestricted designs, in which at most only a general topic is given, to guided exchanges based on structured task solving. Some of the authors of this paper have applied the well-known Map Task [8] in their prosodic studies [9]. Although this task has been thoroughly studied and documented, its nature produces relatively unbalanced dialogs, as the Giver usually supplies most of the information for guiding the Follower to the desired location and the Follower's reactions often consist of one-word acknowledgments such as "yeah", "alright". In contrast, the Video Task developed by Benno Peters [10] involves the interlocutors in a discussion about specially edited diverging versions of an episode of a soap opera. The resulting dialogs are relatively natural and balanced regarding the contributions of the two talkers.

This task, however, requires that interlocutors are familiar with the particular series and also know each other well. The idea of discussing conflicting video presentations is appealing; however we wanted the task to be more focused and generalizable, i.e., not requiring any previous knowledge of

the material or familiarity with the topic. Furthermore, since we ultimately plan to apply the same paradigm in different language and cultural environments, we selected an animated cartoon film of approximately eight minutes that had no dialog.

2.1 *Participants*. Twenty-two pairs of participants (five of them male, 14 female and three mixed) were tested. Participants were recruited from the University of Western Sydney, aged between 17 and 53 and native speakers of Australian English. Participants were either students or university graduates and knew each other previously. Most of the students participated for course credit, the remainder were paid.

2.2 *Materials*. Two (approximately) five minute versions of the film were created in which the first and last scenes were common, but subsequent shots were present only in one or the other. In this way, the complete story was only recoverable when information from both versions was combined.

2.3 *Procedure*. We informed participants that the experiment was about maintaining concentration and collaborating on a cognitive task. Participants were tested in pairs and were told that each person would view a different version of a short silent movie and that the versions were cut in such a way that they were going to see some scenes that their partner would not and vice versa. The cuts in the movie were made so that when a scene was missing the picture would cross-fade into the next scene and the missing scenes also recognizable by interruptions to the background music. We asked participants to memorize the sequence of events and the details of the scenes; they were told that subsequently they would be requested to interact with their partner in reconstructing the story. Specifically, participants were instructed that the story should be recovered cooperatively in chronological order and that they should avoid disclosing all the information they possessed at once, but rather piece together the sequence of scenes as the story develops.

For each participant of a dialog pair, 23 infra-red faces markers were applied in a standard configuration and three markers affixed to a head-worn rig (to track rigid head motion). Participants sat in a sound-treated room facing each other at a distance of about 1.5 m. Each was equipped with a DPA 4066-B head-worn microphone. In order for the facial markers not to be obscured the participants were asked not to raise their hands to their faces if possible.

After calibrating and adjusting the Vicon motion capture system (Lake Forest, CA) which consisted of 8 cameras (4 MX40; 4 MXF40), participants were provided with laptops and head phones for viewing the videos. After participants had finished viewing the video, we started two Sony HDR-PJ200E HD video cameras manually (MPEG4-AVC/H.264 - 1920 x 1080/50i) to have a visual record of each participant (see Figure 1). Following this, the motion capture system was started, capturing audio at 45kHz/16bit and marker motion at a frame rate of 100Hz and the participants were given a signal to begin. During the dialog no instruction were given to the participants. The recording was halted when the participants had decided that they had recovered the story as well as possible.

3. Analysis of Temporal Characteristics

The resulting two videos of each conversation were synchronized with the high quality audio from the motion capture system and joined in a single video that displayed both talkers along-side each other (see Figure 1). Then we performed text level transcription of inter-pausal units on the audio and also annotated non-verbal gestures such as audible breathing, smacks and laughter using the *Praat TextGrid* editor [11]. Based on these transcriptions we performed an analysis of talkers' contributions to the dialog in order to investigate whether the task was balanced.

Figure 1: *Combined videos of talkers A and B of Pair01.*

Table 1 provides information on the resulting 22 dialogs, including the total durations, each participant's percentage of contributions, as well as the percentage of overlaps and silent pauses. Figure 2 displays sample graphic representations of turns along the time axis for a duration of four minutes. In each panel the black areas indicate activity of talker A and the grey areas indicate activity of talker B. As can be seen, Pairs 11 and 17 are balanced with regard to overall contributions by talkers A and B. However, the pairs differ greatly with respect to the distribution of turns. In Pair 11, both talkers produce longer stretches of speech and have fewer turn exchanges, in contrast to the talkers in Pair 17. This indicates that talkers apply different strategies for reconstructing the film. In Pair 11 talker A begins the dialog and talks about several scenes of his version, and only after that talker B presents his observations. The entire dialog continues in this way. Talkers in Pair 17 reconstruct the film more collaboratively by providing shorter pieces of information consecutively and ask each other for missing facts. They interrupt one another more frequently in order to take turns. This is the reason why Pair 11 exhibits only 7% of overlaps, but Pair 17 15%, as can be seen in Table 1. These two examples are representative for most others of the 22 dialogues. Both strategies to reconstruct the film appear to be successful technically, but Pair 17 obviously shared a more vivid exchange and followed our instructions better than Pair 11, hence providing more instances of turn exchanges that we wish to study. We checked whether the version of the video influenced the percentage of talkers' contributions. On average talker A speaks for 48%, and talker B for 41% of the total dialog time. Paired-samples T-test suggests that this tendency is small, but significant ($T=2.239$, $df=21$, $p < 0.036$).

4. Acoustic Analysis

For the subsequent analysis of prosodic features we selected ten pairs (nos. 1, 2, 3, 7, 11, 12, 17, 18, 20, and 22) where the contribution of the two talkers was relatively balanced and where the minimum discourse duration was at least four minutes.

Due to the close proximity between the two talkers during the recording there was audible cross-talk in each of the audio channels, the channel separation being approximately 15-20 dB. By applying audio source separation [12] we yielded a gain of 6-8dB without audible deterioration of the speech signal.

Table 1: *Overview of the 22 dialogs with total durations, percentage talking time of talkers A and B, percentage common pauses and overlap between A and B.*

#	total dur. [s]	% A	% B	% common pause	% overlap	#	total dur. [s]	% A	% B	% common pause	% overlap
01	524	43	37	30	10	12	349	52	40	19	11
02	256	39	37	33	9	13	273	68	36	9	13
03	349	44	40	27	11	14	312	62	33	21	16
04	293	48	28	32	9	15	296	52	38	22	12
05	274	28	64	17	9	16	109	37	49	17	3
06	113	34	41	37	12	17	403	54	50	11	15
07	288	51	45	19	14	18	428	56	41	11	8
08	101	45	24	37	6	19	212	48	50	18	17
09	265	54	25	26	6	20	264	44	52	17	13
10	290	43	38	31	11	21	181	63	35	17	15
11	275	42	48	17	7	22	287	57	43	16	15

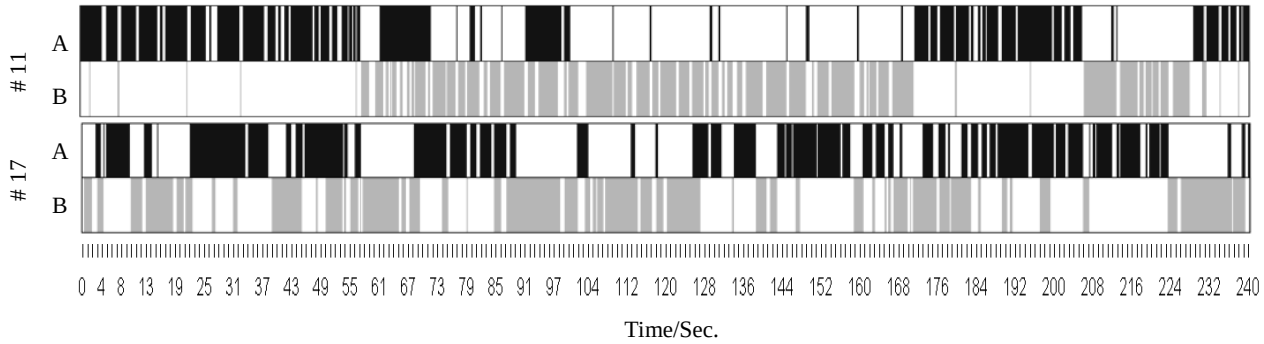


Figure 2: *Graphic representations of turns for two selected conversations displayed for chunks of four minutes. In each panel the black areas indicate activity of talker A and the grey areas indicate activity of talker B.*

We then extracted *F0* contours at a step of 10ms employing the *Praat* standard algorithm [11] with different *F0* floors and ceilings for male (50-300Hz) and female participants (130-400Hz). Along with the *F0* values the *Praat PitchObject* contains information on frame intensity as well as periodicity, a measure comparable to the harmonics-to-noise ratio. For each of the features *F0*, intensity and periodicity we calculated z-scores with respect to the male and female mean and standard deviations, respectively.

In principle our analysis follows the approach presented in [7]: in order to examine the prosodic entrainment between the two talkers in each conversation we calculated means, standard deviation as well as minimum and maximum values of the resulting feature z-scores for chunks of constant length in the conversations (since we do not yet dispose of a detailed transcription of inter-pausal units as well as annotation of turn exchanges).

We then performed two types of analysis: (1) Correlation analysis between the sequences of chunk-wise features for the entire conversation; (2) Statistical analysis of absolute differences between chunk-wise features depending on the talker, the pair, the distance between chunks, as well as the start time of the chunk with respect to the conversation.

After experimenting with several chunk sizes we employed durations of 20s for our subsequent tests. In order to ensure that chunk parameter sets contained averaged values from a sufficient number of speech frames, we required a chunk to contain at least 6s of speech by the talker examined, a speech frame being defined by the intensity reaching a fixed threshold.

On the conversation level, as a test for proximity, we calculated the correlations between sequences of chunk parameters by partners as well as non-partners. Since conversations varied in length, the number N of chunks employed for each analysis varied as well. Results are displayed in Table 2.

Table 2: *Conversation-wise inter-partner correlations (Pearson's r) of mean intensity and mean F0 for selected pairs.*

pair	N	r(mean int.)	p	r(mean F0)	P
01	24	.70	.001	.37	.072
02	10	.72	.020	-.12	n.s.
03	14	.91	.001	.53	.053
07	8	.11	n.s.	.86	.007
11	4	-.97	.035	.93	.067
12	10	-.10	n.s.	.36	n.s.
17	18	.31	n.s.	.39	n.s.
18	15	.39	n.s.	-.02	n.s.
20	8	.34	n.s.	.47	n.s.
22	12	.18	n.s.	.79	.003

Of all features only mean intensity and mean *F0* yielded inter-partner correlations that were significant or approached significance for some of the pairs. Results for pair 11 may be unreliable due to the small number of chunks in which both partners have a sufficiently high number of speech frames.

In our analysis of chunk-wise parameter differences we first calculated means and standard deviations of feature differences between chunks of the same talker (*self*) as compared to those by others (*other*, see Table 3). As expected, talkers were much more similar to themselves than to others.

We then performed intra-talker correlation analysis of chunk-wise feature differences as a function of the distance between the chunks compared. Only mean intensity (Pearson's r =.11, p <.001), intensity s.d. (r = 0.10, p < 0.005) and mean periodicity (r = 0.09, p < 0.02) indicated a weak tendency of the talker to be more dissimilar to him/herself between chunks in discourses that were spaced further apart.

Table 3: *Feature difference means and standard deviations self vs. other.*

	F0 mean	F0 sd	F0 max	int. mean	int. sd	int. max	per. mean	per. sd	per. max
self mean	.24	.22	1.67	.24	.22	1.04	.17	.09	.01
self s.d.	.24	.18	1.25	.22	.21	.92	.14	.08	.08
other mean	.55	.28	2.08	.35	.31	1.50	.29	.16	.18
other s.d.	.42	.22	1.46	.29	.27	1.07	.23	.14	.17

Turning to the relationship between talkers who were engaged in the same conversation (*partner*) as opposed to those in others (*other*), we conducted T-Tests on chunk differences. The results shown in Table 4 indicate that for most features the differences between talkers in the same conversation (*partner*) were smaller compared with talkers from a different conversation. For mean *F0* the difference between *partner* and *other* was significant as well, though the feature differences proper were larger between partners of the same pair.

An intra-pair correlation analysis of chunk-wise inter-talker differences was performed to see whether chunks spaced further apart were more dissimilar. However, we only found a rather weak dependency of mean *F0* on the distance between chunks (Pearson's r = 0.13, p < 0.002).

If we compare intra-pair chunk-wise parameter differences as a function of the onset times of the chunks by only including pairs of chunks occurring at the same time or neighboring one another, mean *F0* (r = 0.28, p < 0.001), intensity max (r = -0.20, p < 0.004) and periodicity s.d. (r = -0.22, p < 0.002) were weakly correlated with the onset times of the chunks in the conversation.

Table 4: *T-tests partner vs. other differences*

Feature	t	df	p-value	Sig.
intensity max	-2.8	7273	<0.001	*
intensity mean	-3.1	7273	0.002	
intensity s.d.	-3.2	7273	0.001	*
F0 max	-5.3	7273	<0.001	*
F0 mean	10.5	7273	<0.001	*
F0 s.d.	-2.7	7273	<0.001	*
periodicity max	-20.4	7273	0.005	
periodicity mean	-0.2	7273	N.S.	
periodicity s.d.	-2.5	7273	0.012	*

As a test of whether or not talkers in the same pair converged during the conversation we examined chunk differences calculated for chunks located in minutes 1 and 2 of the discourse with those in minutes 3 and 4, as only a few of the conversations were considerably longer than four minutes. Mann-Whitney independent sample U-Test suggests differences for intensity mean (p < 0.039) and intensity max (p < 0.017), however, the tendency was for talkers to become more dissimilar with respect to these features later in the discourse.

5. Discussion and Conclusions

This paper presented the first results from an auditory-visual corpus of spontaneous dialogs based on a collaborative task centered on the reconstruction of a cartoon film. Based on transcriptions of inter-pausal units and the inspection of graphical representations of discourse structures we found that conversations overall are relatively balanced between talkers, although pairs differed with respect to the total duration of the discourse as well as turn durations and the amount of overlap.

For a subgroup of relatively well-balanced pairs we examined the prosodic features *F0*, intensity and periodicity with respect to entrainment, the adaptation that can occur between talkers engaged in a conversation. We calculated chunk-wise means, standard deviations as well as min and max values of feature z-scores and examined the relationships between these features for chunks of 20s length. With respect to the whole discourse intensity means exhibited the highest correlations between partners, followed by *F0* max, however this was the case only in some of the pairs. This might reflect individual differences in discourse strategy between pairs. For example, in some dialogues, one partner took the lead and presented most of the information s/he had before granting a turn exchange. In these cases adjustment by the partner may be more difficult than in pairs where the information was delivered in balanced turns.

We investigated chunk-wise feature differences between talkers and themselves, their partners and talkers with whom they had not conversed. With respect to a number of features, especially intensity and *F0*, talkers were more similar to their partners than to other talkers. The similarity seemed to decrease with the distance between chunks in time, though the dependency was relatively weak. We did not find evidence of talkers converging during a conversation though this might simply be due to the short durations of most dialogs. It rather seemed that talkers diverged with respect to intensity, for instance. We believe that it will be necessary to perform a detailed annotation of turns and turn exchanges to better pin-point possible places of stronger coordination. We also require word, syllable and phone-based segmentations in order to test for the entrainment of duration information. In addition, future work will involve annotations of non-verbal facial or head movements followed by the analysis and modeling of the motion capture data.

6. Acknowledgements

We would like to thank Simone Simonetti at MARCS Institute for organizing participants and checking transcriptions and Yuanfu Liao of NTUT Taiwan for applying channel separation to the audio recordings. This work was supported by Deutsche Forschungsgemeinschaft international research grant Mi625/24 funding a stay by Mixdorff and Hönemann at MARCS Institute.

7. References

[1] Sumby, W. H., & Pollack, I.,"Visual contribution to speech intelligibility in noise. JASA, 26, 212-215, 1954.

[2] McGurk, H., & MacDonald, J., "Hearing Lips and seeing voices", in: Nature, Volume 264, pp. 746-748, 1976.

[3] Mixdorff, H., Charnvivit, P. and Burnham, D., "Auditory-Visual Perception of Syllabic Tones in Thai," in Proceedings of AVSP 2005, pp. 3 - 8, Parksville, Canada, 2005.

[4] Hönemann, A. & Mixdorff, H. and Fagel, S., "A preliminary analysis of prosodic features for a predictive model of facial movements in speech visualization", Proceedings of Nordic Prosody 2012, Tartu, Estonia, 2012.

[5] Mixdorff, H., Hönemann, A. and Fagel, S., "Integration of Acoustic and Visual Cues in Prominence Perception", Proceedings of AVSP 2013. Annecy, France, 2013.

[6] Kim, J., Cvejic, E. and Davis, C.,"Tracking eyebrows and head gestures associated with spoken prosody. Speech Communication, 57, 317–330, 2014.

[7] Levitan, R. and Hirschberg, J., "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions", Proceedings of Interspeech 2011. Florence, Italy, 2011.

[8] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G.M., Garrod, S., Isard, S. D., Kowtko, J. C., McAllister, J., Miller, J., Sotillo, C. F., Thompson, H. S. & Weinert, R., "The HCRC Map Task Corpus. In: Language and Speech 34, pp. 351-366, 1991.

[9] Mixdorff, H., Pech, U., Davis, C. and Kim, J., "Map Task Dialogs in Noise - a Paradigm for Examining Lombard speech". Proceedings of ICPHS07, Saarbrücken, Germany, 2007.

[10] Kohler, K. J., B. Peters, and M. Scheffers (Eds.), "The Kiel Corpus of Spontaneous Speech IV, German: Video Task Scenario (Kiel-DVD1)", Kiel: IPDS, Christian-Albrechts-University, 2006.

[11] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International 5, 341-345, 2001.

[12] Ozerov, O. and Févotte, C. "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 3, MARCH 2010.

A preliminary study on the prosody of broadcast news in Hong Kong Cantonese

Peggy Pik Ki Mok, Holly Sze Ho Fung, Jingwen Li

Department of Linguistics and Modern Languages, The Chinese University of Hong Kong

peggymok@cuhk.edu.hk, hollyfung@cuhk.edu.hk, joanneljw@gmail.com

Abstract

Broadcast news is a distinctive register. Previous studies only provided some general descriptions of the prosodic features in broadcast news but with few concrete data. Most of them were also on English news. This study investigated the prosodic features of Cantonese TV broadcast news using acoustic data. Speech using the same materials from two groups was compared: eight Hong Kong professional TV news anchors, and a control group consisting of eight university students. The results show clear differences between the two groups in terms of speech rate, pitch range and variability of syllable duration (speech rhythm). It was found that the news anchors spoke significantly faster than the control group, also with an enlarged pitch range. They also produced more variability in syllable duration. There is clearly more prosodic variation in the news register than ordinary speech. Finally, we provide some possible reasons for these features, as well as directions for future studies.

Index Terms: news, broadcast, prosody, Cantonese

1. Introduction

This study investigates the prosody of Cantonese news broadcast in television read by professional news announcers and compares it to that read by non-professional speakers. Broadcast news has a distinctive style/register in different languages, but not many studies have examined its specific prosodic features. Most previous studies were on English news, while so far no study has examined the prosody of Cantonese broadcast news. This preliminary aims to fill this research gap.

The broadcast news register is easily recognizable even by lay people. Before the 1970s, news readers spoke in a very formal and solemn way, delivering news with much authority. Women were excluded from news broadcast due to their voices being too high-pitched or lack in authority [1]. After the 1970s, with technological advancements in voice capturing techniques and various social changes (e.g., the introduction of television in common households), the style of broadcast news has changed dramatically. Summarizing some previous descriptions on vocal characteristics of newsreaders, [2] suggested that an effective newsreading voice is characterized by a rich, warm, and resonant tone. News readers need to deliver the news clearly and succinctly, and with enthusiasm to capture the audience's attention, while at the same time sounding professionally distant and knowledgeable.

Cotter [3] proposed three factors which affect news prosody: text structure (condensed text written to convey newsworthiness and capture attention); an unseen audience and medium constraints (using voice only to signal various boundaries and emphasis). These factors contribute to the unique prosody of broadcast news register, which, according to [4], is between reading and spontaneous speech, having all the features of reading and some features of spontaneous speech.

Despite the easy identification of the news register, and the literature on general perception and opinions of good practice in newsreading (see summary in [2, 5]), surprisingly few studies have examined the prosodic features of broadcast news in detail. Bolinger [6] noticed that some American radio newscasters intentionally distorted expected sentence stress and put accentual emphasis on words whether semantically justified or not in order to attract attention. Cotter [3] compared news read by radio announcers and that by volunteers, and found that news prosody makes use of speech rate, pauses, pitch movement and other paralinguistic features in an identifiable way, incorporating features of spontaneous and public discourse. [2] compared professional newsreaders with student newsreaders and controls, and found that female newsreaders had a higher pitch than both students and controls; professionals also had greater pitch variability, spoke faster and made fewer errors than both students and controls. The professional newsreaders were also rated significantly higher by the judges on phrasing and overall performance. Moreover, more students reported consciously altering their voices and using more effort in newsreading than the professionals did.

The above studies are all based on broadcast news in English. Zou [7, 8] examined the prosody of broadcast news in Mandarin Chinese. He found that the mean syllable duration is longer in news announcements than in conversation spoken by the same presenters, i.e., the presenters spoke more slowly and clearly in news announcements, compared with their own conversation. Moreover, there is a wider pitch range variation in news announcements than in conversation. His data show that a distinct news register is also present in Chinese. However, it is unclear if these prosodic features in Mandarin news are also applicable to broadcast news in Cantonese, as broadcasters in China are specially trained for the profession with very strict standards on their speech characteristics, while it is not the case in Hong Kong. Also, news prosody can vary between languages [1, 9] and also between different ages of the same language [9, 10]. Therefore, more studies on the prosody of broadcast news in different languages are needed to thoroughly investigate the various prosodic features of this distinct speech style.

2. Method

The present study compares the news announced by professional journalists (anchors) in television news broadcast with those read by non-professional speakers (controls). The controls read exactly the same materials as the anchors for direct comparison.

2.1. Speakers and materials

Eight news anchors (four male, four female) from a mainstream local television broadcasting company in Hong Kong (TVB) were chosen. It is unknown if the company has any special voice/speech training for their news anchors, but the company is known for its insistence on a high standard of pronunciation, e.g., their anchors preserve some contrasts that

are already lost or used interchangeably by most Hong Kong people. Approximately one-minute of news announced by each anchor based on four different stories each was chosen from the 'News at Six-Thirty' programme aired between 23 March and 23 July 2013, videos of which could be found in YouTube. Criteria for choosing the anchors include their popularity and the availability of videos in YouTube. Table 1 summarizes the information about the speech samples of the eight anchors.

Table 1. *Summary of news materials by the anchors.*

Female				Male			
Anchors	# σ	# IP	# S	Anchors	# σ	# IP	# S
Cheng	246	20	5	Fong	278	16	5
Chow	246	19	5	Lau	262	22	6
Lau	266	15	6	Ng	279	19	5
Law	277	22	5	Pun	303	24	5

σ = syllables; IP = intonational phrases; S = sentences

The control group consists of eight local university students, also four male and four female, all native speakers of Hong Kong Cantonese. They were aged between 18 to 27, and none of them reported any speech or hearing impairment. Two speakers received course credits and another speaker was paid for their participation. The other five speakers participated on a voluntary basis.

2.2. Procedure

A pilot trial with one speaker (not in the control group) indicated that it would be very difficult and very slow for the control speakers to read the news materials fluently on the spot. In order to solve this problem, the control speakers were given the news materials a couple of days in advance and were asked to practise them well before the actual recording. Therefore, the control speakers knew the purpose of the experiment. This is unavoidable given the logistic constraints. This also means that the control speakers were performing at their best as 'mock anchors', based on their own perception of the news register.

The recording took place in a quiet room at the Chinese University of Hong Kong. A solid state recorder was placed approximately 20 cm away from the speakers. The speech materials were recorded with a sampling rate of 44100 Hz and 16 bits. The control speakers were asked to read aloud the stimulus sentences displayed on a computer screen one by one in a random order. If they made a mistake, they were prompted to re-read the sentence again immediately, so the recorded materials contain no speech errors or disfluencies. Two repetitions of the materials were collected.

In order to control for the gender difference, the control speakers only read the news materials produced by the anchors of the same sex.

2.3. Data analysis

As the sentences written for the news scripts were very condensed and very long, the sentences were divided into intonational phrases for analysis (see Table 1). Three types of data were examined to compare the prosodic features between anchors and controls: speech rate, pitch range and speech rhythm. Speech rate was calculated based on the average number of syllables per second. Since there is much individual variation between speakers' habitual pitch levels, pitch range is a better indicator of any stylistic pitch variation in news prosody than mean pitch is. We used two methods to estimate the pitch range. As Cantonese is a tone language, a good estimation of the pitch range can be obtained by comparing the F0 values at the midpoint of all Tone 1 syllables (a high level tone [55]) with those of Tone 4 (a low falling tone [21]), which respectively represent the highest and lowest pitch levels in a speaker's voice. The pitch level at the end of Tone 4 falls so low that it often results in creakiness. Measuring F0 values at the midpoint can ensure more valid data. In addition to the pitch differences between T1 and T4, we also calculated the pitch ratios between T1 and T4 as a normalized measure for all speakers.

We used the acoustic metrics on syllable duration, VarcoS and nPVIS, to compare speech rhythm. VarcoS shows the standard deviation of syllable duration in an utterance normalized for speech rate [11], while nPVIS shows the normalized pairwise variability of syllable duration [12]. VarcoS measures durational variability globally, while nPVIS measures variability locally between each pair of unit. Mok [13, 14] have shown that these metrics on syllable duration can be more robust than metrics on consonantal and vocalic durations in reflecting rhythmic differences. Moreover, the syllable structure of Cantonese is very simple. Segmenting the speech streams into syllables is quite easy and straight forward. It is noted that the rhythmic metrics have been under much criticism in recent years [15], but it is justified in our case because we compared exactly the same speech materials between anchors and controls. Cantonese is a typical syllable-timed language [13, 14, 16]. It will be interesting to see if more durational variability would be employed by the news anchors to enhance liveliness and to capture audience's attention.

3. Results

3.1. Speech rate

Table 2 shows the average speech rate of the eight anchors and the control speakers. It is obvious that the anchors spoke faster than the controls [$t(14) = 7.484$, $p < 0.0001$]. It is worth noting that the controls were already performing at their best as 'mock anchors', and no disfluency was included in their recordings. Nevertheless, they still spoke significantly slower than the anchors.

Table 2. *Average speech rate (# syllable / second).*

Source	Anchors	Controls*
Cheng	6.10	4.66
Chow	5.75	4.76
Lau	5.68	4.53
Law	5.19	4.53
Fong	5.51	4.97
Lau	5.62	4.84
Ng	6.19	4.90
Pun	5.75	4.93

* The control data are averaged across the reading of the same anchor's materials by four control speakers of the same sex.

3.2. Pitch range

Figure 1 shows the pitch differences in Hz of the anchors and those averaged across the four control speakers reading the materials by the same anchor. Each bar represents one

anchor's materials. Female and male data are shown separately. It can be seen that female anchors have a significantly larger pitch difference than female controls do [$t(6) = 7.147$, $p < 0.001$]. The pitch range is also generally larger for male anchors than male controls, although the difference is not significant [$t(6) = 1.669$, $p = 0.146$]. The larger pitch range indicates that the (female) professional anchors used more pitch variation in their news reading. It is interesting to note that there is also more individual variation in pitch range among the anchors than among the controls.

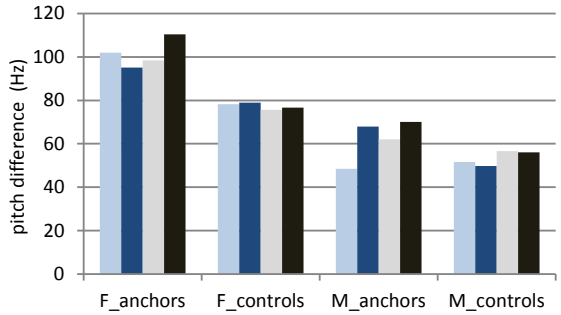


Figure 1: *Pitch difference of anchors and controls.*

In addition to the differences in Hz, we have also calculated the ratios of T1/T4 as a normalized measure of pitch range for both anchors and controls to verify the patterns observed above. Figure 2 shows the pitch ratios of both groups of speakers. Again, anchors have a larger pitch range than controls do, and the difference is significant for the female speakers [$t(6) = 5.344$, $p=0.002$], and is tending towards significance for male speakers [$t(6) = 2.075$, $p = 0.083$].

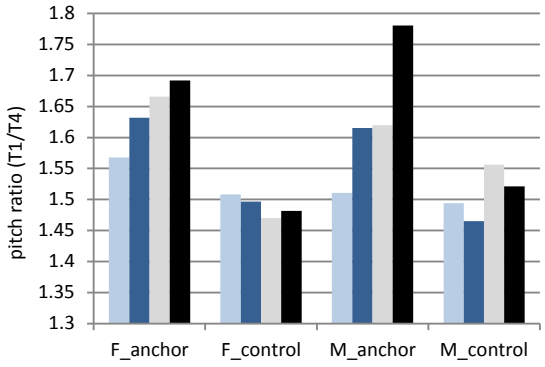


Figure 2: *Pitch ratio of anchors and controls.*

3.3. Rhythmic metrics

Table 3 shows the values of VarcoS and nPVIS for the anchors and those averaged across the four control speakers reading the materials by the same anchors. Anchors have significantly higher variability of syllable duration than controls do for both rhythmic metrics: VarcoS [$t(14) = 5.90$, $p < 0.0001$] and nPVIS [$t(14) = 4.503$, $p = 0.001$]. Moreover, if we examine the values more closely, those produced by the anchors are always higher than those averaged among the controls, even though they read exactly the same news materials. It is also true if we compare the anchor's data with each individual control speaker.

Table 3. *Rhythmic metrics on syllable duration.*

Source	Anchors		Controls	
	VarcoS	nPVIS	VarcoS	nPVIS
Cheng	28.59	30.17	23.31	25.05
Chow	25.20	28.44	22.89	25.77
Lau	28.11	27.95	24.22	25.84
Law	29.24	32.05	24.97	27.90
Fong	30.62	33.61	23.73	26.31
Lau	27.62	31.42	25.72	30.02
Ng	29.68	30.83	25.39	28.10
Pun	28.42	31.01	25.18	28.06

4. Discussion

The study has found clear differences in the prosodic features between news read by professional anchors and by control speakers. Some differences can be expected, e.g., speech rate, while others are rather interesting, e.g., speech rhythm.

The best versions produced by the control speakers are still significantly slower than those produced by the anchors. In order to speak clearly as 'mock anchors', the controls produced the materials carefully and thus had slowed down the speech rate, as what most people would do. However, the anchors could speak clearly and quickly, as they need to deliver as much information as possible in a short period of time. The demands on time necessitate a faster speech rate. So in addition to the three factors (text structure, unseen audience and medium constraints) proposed by Cotter [3] which influences news prosody, time pressure is another important factor contributing to the unique prosody of broadcast news.

In addition to saying things more slowly, listening to the recordings tells us that another reason why the controls were slower is that they made more pauses and/or paused longer than the anchors.

Furthermore, given that news anchors can speak clearly and quickly, it will be interesting to examine their speech articulatorily to see if there are differences between their articulation in news reading and that of control, and also between the news speech and the 'normal' speech by the same anchors. Do anchors simply move their articulators faster or do they have different articulatory strategies in news reading? For example, would there be more gestural overlaps in their news reading? A faster speech rate usually results in more gestural overlaps [17], but the demand for clarity would result in less overlaps [18]. Would the anchors increase the magnitude of their gestures in order to accommodate these contradicting demands? Interesting patterns are likely to be found.

Professional news readers need to capture and sustain the attention of the audience, and to highlight different important information in the scripts that are already very condensed. A useful strategy is to increase variability in speech. Our data show that the anchors did so in both pitch and duration.

The pitch ranges produced by the female anchors were significantly larger than those by the female controls. Thus, the 'tone space' is larger for the anchors, and their lexical tones would be more distinct than those by the controls. In addition to having more distinct tones, more fluctuating intonation patterns would also result in a larger pitch range. It is quite likely that both aspects have jointly contributed to their larger pitch range. Further studies should devise ways to evaluate the contribution of lexical tone and intonation

separately.

It is interesting to notice that although the male anchors had a similar pattern of generally having a larger pitch range than the male controls did, the difference is not significant. The pitch range for male speakers is naturally smaller than that of female speakers, so the increase in pitch variability would probably be proportionally smaller as well, although Figure 2 shows one male anchor having a higher ratio than the female anchors. Probably, the more subtle increase is harder to show up statistically.

A further reason for the insignificant result is that one male control speaker had a pitch range larger than other male speakers, including three male anchors. The authors were familiar with this control speaker. When listening to his recordings, it is obvious that he had prepared the materials well and spoke with more pitch variation in the news reading than in his ordinary speech. He performed well as 'mock anchors'. This shows us that even lay speakers have a hunch of the pitch patterns and variation in the news register, which are confirmed statistically by our female data.

Last, but certainly not the least, our results indicate that the anchors had also increased the variability of syllable duration in their news reading. This finding is particularly interesting given that Cantonese is a very typical syllable-timed language, even more so than French and Italian [19]. One may expect that variability of syllable duration would be the least likely aspect to be manipulated. Our significant findings confirm that even such unlikely aspect is used in news prosody. All the anchors consistently had higher variability than the controls as a group had, and also consistently higher variability than each individual control speaker.

When we listened to the recordings, it is obvious that the anchors were skillful in compressing some common or familiar phrases, while speaking more clearly and slowly for other important information. For example, the title and the name of the President of China, Xi Jinping (國家主席習近平) produced by one female anchor had the following characteristics: she had parsed the whole phrase into two parts (the title with four syllables and the name with three syllables). The duration for the first part with four syllables is 531ms, while it is 552ms for the second part with only three syllables. In addition to such strategic compression and extension, since the speech rate is so fast, the anchors may be actively using syllable duration for phrasing and creating boundaries. Also, they may produce strong focus when important terms come up, resulting in longer syllable durations in certain phrases. All this explains very well why the rhythmic metrics for syllable duration are significantly higher for all anchors than for the controls. Much skill and experience in information selection is needed in order to manipulate this prosodic feature well. Further studies on Cantonese news prosody should especially examine this interesting feature.

Our study is a preliminary study on the news prosody in Hong Kong Cantonese. It provides concrete acoustic data as evidence, supplementing the many general descriptions of features of news prosody in the literature. Nevertheless, one shortcoming is that the data, although well controlled for comparison, is quite limited. Further studies using a larger data set are needed to corroborate the findings in this study.

5. Conclusions

A distinct register of broadcast news is found in Cantonese. Professional news readers speak significantly faster than control speakers do. They also have more stylistic variations in both pitch and syllable duration. It will be interesting to compare the news register in different languages in the future to confirm which prosodic features are common for news reading in general, and which features are language-specific devices to cater for this special speech style.

6. References

[1] Price, J. (2008). New News Old News: A Sociophonetic Study of Spoken Australian English in News Broadcast Speech. *Arbeiten aus Anglistik und Amerikanistik,* 33(2), 285-309.

[2] Neil, E., Worral, L., Day, A., & Hickson, L. (2003). Voice and speech characteristics and vocal hygience in novice and professional broadcast journalists. *Advances in Speech-Language Pathology,* 5(1), 1-14.

[3] Cotter, C. (1993). Prosodic Aspects of Broadcast News Register. In *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Semantic Typology and Semantic Universals* (pp. 90-100). Berkeley.

[4] Levin, H., Schaffer, C., & Snow, C. (1982). The prosodic and paralinguistic features of reading and telling stories. *Language and Speech,* 25(1), 43-54.

[5] Warhurst, S., McCabe, P., & Madill, C. (2013). What Makes a Good Voice for Radio: Perceptions of Radio Employers and Educators. *Journal of Voice,* 27(2), 217-224.

[6] Bolinger, D. (1982). The network tone of voice. *Journal of Broadcasting,* 26, 726-728.

[7] Zuo, Y., Li, X., & Hou, M. (2006). Comparison of News Announcing and Talking Styles in Broadcast Speech, *International Symposium on Chinese Spoken Language Processing (ISCSLP 2006)*. Singapore.

[8] Zou, Y. (2007). A Formal Study on Prosody of Presenter's Spoken Language Based on Broadcast Speech Corpus (in Chinese). PhD thesis, Communication University of China.

[9] Boula de Mareüil, P., Rilliard, A., & Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. *Language and Speech,* 55(2), 263-293.

[10] Zou, Y., Wang, Y., & He, W. (2012). Diachronic contrastive analysis on read speech in broadcast news: evidence from pitch and duration, *The International Symposium on Chines Language Processing (ISCSLP 2012)* (pp. 291-295). Hong Kong.

[11] Dellwo, V. (2006). Rhythm and speech rate: a variation coefficient for ΔC. In P. Karnowski & I. Szigeti (Eds.), *Language and Language-Processing* (pp. 231-241). Frankfurt am Main: Peter Lang.

[12] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology VII* (pp. 515-546). Berlin: Mouton de Gruyter.

[13] Mok, P. (2011) The acquisition of speech rhythm by three-year-old bilingual and monolingual children: Cantonese and English. *Bilingualism: Language and Cognition*, 14: 458-472.

[14] Mok, P. (2013) Speech rhythm of monolingual and bilingual children at 2;06: Cantonese and English. *Bilingualism: Language and Cognition,* 16: 693-703.

[15] Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica,* 66(1-2), 46-63.

[16] Mok, P. (2009). On the syllable-timing of Cantonese and Beijing Mandarin. *Chinese Journal of Phonetics,* 2, 148-154.

[17] Byrd, D., & Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics,* 24(2), 263-282.

[18] Tasko, S. M., & Greilick, K. (2010). Acoustic and Articulatory Features of Diphthong Production: A Speech Clarity Study. *Journal of Speech, Langauge and Hearing Research,* 53, 84-99.

[19] Mok, P. & Dellwo, V. (2008) Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In *Proceedings of the 4th Speech Prosody (2008)*. 423-426. Campinas, Brazil.

Prosodic chunking algorithm for dictation with the use of speech synthesis

Sebastien Le Maguer[1], *Elisabeth Delais-Roussarie*[2], *Nelly Barbot*[1], *Mathieu Avanzi*[2], *Olivier Rosec*[3], *Damien Lolive*[1]

[1]IRISA, Université de Rennes 1, Lannion, France
[2] UMR 7110-Laboratoire de Linguistique Formelle, Université Paris-Diderot, France
[3] VOXYGEN, Lannion, France

Sebastien.Le_maguer@irisa.fr, Elisabeth.roussarie@wanadoo.fr, Nelly.Barbot@irisa.fr, mathieu.avanzi@gmail.com, olivier.rosec@voxygen.fr, Damien.Lolive@irisa.fr

Abstract

The aim of this paper is to present an algorithm that automatically segment a text in prosodic chunks for a dictation by conforming to the rules and procedures used in real settings to dictate a text to primary school children. A better understanding and modeling of these rules and procedures is crucial to develop robust automatic tools that could be used in autonomy by children to improve their spelling skills through dictation with the use of speech synthesis. The different steps used to derive the prosodic chunks from a given text will be explained through concrete examples. The proposal made here relies on the analysis of a corpus of 10 dictations given to children in French and French Canadian elementary schools, and more precisely during their first three years in elementary school (i.e. cycle 2 in the French school system). The phrasing observed in the data is described. It is thus simplified in order to develop an algorithm that automatically generates prosodic chunks from texts.
Index Terms: speech synthesis, prosodic phrasing, automatic parsing.

1. Introduction

The use of software and automatic tools in language teaching offers several advantages among which we may mention the ability to adapt to the learner needs and to provide an environement for him to work in autonomy. Despite these advantages, there are nowadays in France few softwares which are currently used in primary schools to teach reading and writing skills with the use of speech synthesis (e.g. Lectramini[1] and PLATON[2]).

One of the goal of the collaborative ANR research project Phorevox (`http://www.phorevox.fr/`) is to develop this kind of tools, with special emphasis given to the acquisition of writing skills. The software will automatically propose some practice exercises which allows children improving their written skills. Thus, dictations may be given to the children to work on specific grammatical or orthographical aspects.

In order to use automatic procedures and speech synthesis systems for dictation, it is necessary to (i) provide a very intelligible synthesized speech to allow children to hear all the words and sounds to be written, (ii) divide the texts to dictate into chunks that allow accessing all relevant grammatical information, while being of a reasonable size, and (iii) provide a user-friendly typing environment that takes into account the typing speed of the different children. Among these issues, we will focus in this paper on the ability to automatically divide a text to dictate into chunks. To develop an automatic chunking procedure for dictation, we have first analyzed a set of dictations made in primary school. The results of the data observation were used to select and formalize the rules used by the algorithm which generates the dictation.

The paper is organized as follows. In section 2, the rules that are used to derive the prosodic chunks in standard French are briefly presented, and the main characteristics of prosodic phrasing in French are described. The segmentation procedure used by our speech synthesis system are then explained. In section 3, the data and methodology chosen to study the phrasing patterns of dictations in French are presented. The phrasing rules extracted from the observation of the data are listed in section 4. Section 5 explains which pieces of information are taken into account to develop the algorithm that automatically provides a chunking to any text.

2. Background

Studies on French prosody have traditionally pointed out that accentuation, phrasing and intonation are closely intertwined (see, among others, [3], [4] and [5]). In French, the lack of lexical stress causes a syncretism between intonation and accentuation.

In the studies on prosodic phrasing, two or three distinct levels of phrasing above the word are argued for. The lower level, i.e the minor phrase (MiP) – which is also called accentual phrase (see [6] and [7], among others), phonological phrase (see [4] among others) or rhythmic group (see [8] among others) – plays a crucial role. This unit is characterized by the realization of a phrasal stress on its last metrical syllable, which indicates its right edge. In the literature, there is a broad consensus about the definition of this unit: it corresponds minimally to a lexical word and to all the function words that this word governs (see, among others, [4], [6], [7] and [8]). The sentences in (1) are segmented in Minor Phrases as shown in (2).

(1) a. *Les enfants sont venus dans l'après-midi.*
The children came in the afternoon.

b. *Bernard est rentré de son voyage en Asie.*
Bernard came back from his travel to Asia.

(2) a. (Les enfants)$_{MiP}$ (sont venus)$_{MiP}$ (dans l'après-midi)$_{MiP}$.

b. (Bernard)$_{MiP}$ (est rentré)$_{MiP}$ (de son voyage)$_{MiP}$ (en Asie)$_{MiP}$.

In addition to the level of the MaP, two additional levels of phrasing are often referred to: the intermediate phrase or ip (see,

among others, [7] and [9]), which is also called major phrase (see, among others, [10]) or restructured phonological phrase (as in [4]); and the intonational phrase or IP (see, among others, [3], [4], [6] and [7]), which is also called the Breath group. Even if there is no broad consensus on the existence of the ip, this level of phrasing is often requested when the morphosyntactic structure is relatively complex. As to the Intonational phrase or IP, it is the largest prosodic unit in the prosodic hierarchy. It is characterized by a presence of an intonational contour at its right edge, the strongest degree of phrase-final lengthening, and also often followed by a pause. In sequences of clauses, each clause is normally phrased as an independent IP.

In section 4, the prosodic phrasing observed in the data, and the information requested to generate these phrases will be explained in details. It will allow evaluating what differentiate phrasing in standard French from phrasing in dictations.

3. Corpus and methodology

3.1. Corpus

To study prosodic chunking and intonation in dictation, a set of dictations has been gathered. This set consists of dictations that have been given to children enrolled in first to third year elementary school in France and Quebec (Canada). The data come from three different sources:

- Four short dictations come from the website of the Canadian association "Fondation Paul Gerin-Lajoie"[11], in particular, the dictations for first year level (CP in France);
- Four dictations come from the French website La-dictée.fr[12], which offers a wide range of dictations and grammatical exercises to school children;
- Two dictations have been recorded in class situations by some researchers belonging to the Phorevox project.

The choice of the dictations has been done in order to cover the various levels we are interested in for the software development. Moreover, as they come from different sources, some variation may occur in the way to dictate a text, in particular for the repetition of certain sequences. Some teachers repeat each sequence a few times (two or even more), while some others don't. The software will allow the user to configure such repetitions for any given dictation.

3.2. Methodology

The dictations have been annotated by two of the authors. The annotation indicates for each text two distinct types of information: the phrasing obtained (i.e. the way the texts were segmented into chunks during the dictation); and the form of the tonal contours that occur at the end of the various chunks.

The annotation has been achieved by means of a perceptual and instrumental data analysis. The perceptual analysis was done by a careful listening of the data, and allowed determining the chunking and the form of the pitch contours at the end of the various chunks. The acoustic analysis, achieved with the Praat software [13], confirmed what was perceived. Special attention was given to the occurrence of pauses to determine the segmentation in chunks, and to the form of the tonal contours occurring at the end of the prosodic chunks.

4. Data analysis: defining chunking rules

Before describing in details the rules used to derive the prosodic chunks, and the procedures at stake to repeat and introduce new chunks, we want to mention three major features that have been observed in all the analyzed dictations. First, the title and the whole text are said once at a relatively slow rate at the beginning, and then the dictation proper begins. Second, during the dictation proper, the whole sentence is usually repeated once after all its parts have been dictated in separate chunks. Third, punctuation marks are pronounced at the position they occur in the written text for the children to encode them as shown in (3) and (4), where the orthographic transcription of what has been pronounced is given.

(3) *Avec Papa. Point. Je marche dans la nature avec Papa. Point.*
With Daddy. Full stop. I am walking in the country with Daddy. Full stop.

(4) *A l'école, virgule, je travaille toujours avec lui. Point.*
At school, comma, I am working with him. Full stop.

Nevertheless, variation occurs in the pronunciation of the punctuation marks: the pronunciation of commas may be omitted when the sentence as a whole is repeated for the last time, whereas the pronunciation of stops may be done only when the sentence as a whole is produced. Since the latter realizations are not systematic, they will not be taken into account in the elaboration of the dictation algorithm presented in section 5: punctuation marks will always be pronounced in the position where they occur in the written text.

4.1. Phrasing and prosodic structure observed

In the observed data, the chunks used during the dictation proper correspond, in more than 95% of the cases, to minor phrases (see section 2), that is to lexical words preceded by the function words they syntactically govern. In a prepositional phrase, for instance, the noun is always phrased with the preposition and the determinant as in (5a), and, in the same vein, an auxiliary is phrased with the verb as in (5b).

(5) a. *Je marche dans la nature avec Papa* $\rightarrow$ (Je marche)$_{MiP}$ (dans la nature)$_{MiP}$ (avec papa)$_{MiP}$

b. *Il se demande si sa maman a trouvé les bons médicaments* $\rightarrow$ (Il se demande)$_{MiP}$ (si sa maman)$_{MiP}$ (a trouvé)$_{MiP}$ (les bons médicaments)$_{MiP}$

Note however that two Minor Phrases, which are derived from morphosyntactic information, may be restructured in a single one when the size of the phrase is inferior to two syllables. In many cases, for instance the copula *être* is restructured with what follows as shown in (6). The restructuring fails sometimes to apply, in particular when the resulting phrase would be relatively long as in (7).

(6) *Le lac est bleu* $\rightarrow$ [Le lac] [est bleu]

(7) *C'est mon meilleur ami* $\rightarrow$ [C'est] [mon meilleur ami]

As for the prosodic realization, each prosodic chunk is treated as an IP, be it in isolation as in (8) or integrated in a sentence as in (9), when the whole sentence is produced at the beginning or at the end.

(8) *Avec mon ami.* $\rightarrow$ [avec mon ami]$_{IP}$ [point]$_{IP}$

(9) *Je joue dans l'eau avec mon ami* → [Je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

This means that in dictation as a speaking style any prosodic group that would be a minor phrase in a normal reading style is realized with the major prosodic features of an IP such as an important final lengthening and a presence of a pause. Such a phrasing has been described as completely appropriate in French by [8]. It results from what [8] called the *élasticité prosodique* (i.e. prosodic elasticity), and account for the fact that any MiP could be realized as an IP, without any further restructuring.

The segmentation procedure used to dictate the text could thus be based on a parsing which introduces a major break after the words categorized as nouns, verbs, adjectives or adverbs if they are not modifying the following word. This latter principle should allow phrasing together in the same IP prenominal adjective and noun as in "le petit garcon", modifier adverb and adjective as in "très ennuyeux", or verbal auxiliary and past participle as in "est arrivé".

4.2. Procedures of repetition

Apart from the segmentation and the pronunciation in chunks, new chunks and sentences are introduced by special procedures that will be described in the following subsections. From the observation of the data, it was possible to infer three distinct procedures

4.2.1. Procedure IP by IP

This procedure consists in pronouncing the whole sentence where each MiP is realized as an IP first, and then to produce each IP in isolation (be they repeated or not). When all IPs have been uttered, the whole sentence is pronounced once again. For the sentence in (10), this procedure will lead to the chunking and the realization in (11). Each line break indicates that the chunk is uttered isolated from the preceding ones.

(10) Je joue dans l'eau avec mon ami.

(11) [je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$
[dans l'eau]$_{IP}$
[avec mon ami]$_{IP}$ [point]$_{IP}$
[je joue]$_{IP}$ [dans l'eau]$_{IP}$ [avec mon ami]$_{IP}$ [point]$_{IP}$

This procedure has mostly been used for the dictations given to the younger children (first year in elementary school).

4.2.2. Procedure by IP chaining

The procedure by IP chaining relies on a dictation of the sentence IP by IP, each IP consisting of what would be a MiP in colloquial speech. When the first IP has been realized once, it is repeated followed by the next IP, then the added IP is also produced in isolation once. For the sentence (12), this procedure will lead to the chunking and pronunciation in (13).

(12) Le chien s'étire sur le tapis.

(13) [Le chien]$_{IP}$
[Le chien s'étire]$_{IP}$
[S'étire]$_{IP}$
[S'étire sur le tapis]$_{IP}$ [Point]$_{IP}$
[Sur le tapis]$_{IP}$ [Point.]$_{IP}$
[Le chien]$_{IP}$ [s'étire]$_{IP}$ [sur le tapis]$_{IP}$ [point]$_{IP}$

Even if this procedure has been used in approximately 25% of the case in our data, it has some serious drawbacks, in particular in case of more complex sentences. In a sentence with a branching NP subject the verb is not uttered with the subject as shown in (14). Such a way of dictating is really error prone, as the subject-verb agreement cannot be interpreted in a straightforward manner.

(14) La porte de la chambre s'ouvre.
[La porte]$_{IP}$
[La porte]$_{IP}$ [de la chambre]$_{IP}$
[De la chambre]$_{IP}$
[De la chambre]$_{IP}$ [s'ouvre]$_{IP}$ [point]$_{IP}$
[S'ouvre]$_{IP}$ [point]$_{IP}$
[La porte]$_{IP}$ [de la chambre]$_{IP}$ [s'ouvre]$_{IP}$ [point]$_{IP}$

4.2.3. Procedure sentence by sentence

The last procedure consists in dictating the text sentence by sentence. Each sentence is pronounced once or twice, depending on its size, at a relatively slow rate. The segmentation in IP should be clearly realized as in (15), where a succession of two sentences is given.

(15) Le lac est bleu. J'aime le lac.
[Le lac]$_{IP}$ [est bleu]$_{IP}$ [point]$_{IP}$
[J'aime]$_{IP}$ [le lac]$_{IP}$ [point]$_{IP}$

In cases of complex or long sentences, the segmentation proceeds clause by clause. Clause refers here to different types of elements:

- Comma clause such as peripheral adjunct followed by a clause as the underlined sequence in (16).
- Subordinated or coordinated clauses as in (17)

(16) A l'école, je travaille toujours avec lui.

(17) Mon père rentre très tard à la maison parce qu'il est musicien.

When such a sub-segmentation is used, the sentence as a whole is repeated once when all parts have been dictated.

5. Adaptation of the rules and procedures for speech synthesis

The achieved segmentation and observed procedures have been used to automatically dictate any text with a speech synthesizer. The implementation of the various elements just described was achieved by generating the chunks, and producing the text while respecting the various features mentioned at the beginning of section 4.

5.1. Chunk generation

To explain how we generate the chunk list from an input text, we are going to take the following French sentence as an example:

(18) *En réalité, c'est un hélicoptère. Avec une cheminée qui crache de la fumée, comme une locomotive à vapeur.*
Actually, this is an helicopter.With a smokestack to expel smoke, like a steam locomotive.

5.1.1. *Main algorithm*

To generate a word chunk list, we first need to determine the syntax tree associated to the dictation text. This is given by the Synapse pos tagger[14]. We suppose that the nodes are corresponding to a syntactic phrase and each leaf is associated to a word.

For a given node N_0, we identify the children of N_0 by $(N_1 \dots N_n)$. Each node N_j represent a chunk of words. So N_j is defined by $N_j = (s_j, ej)$ where s_j is the first word's index and e_j the index of the chunk's last word. The syntax tree has the following properties which implies an order between children: $s_0 = s_1$; $e_0 = e_n$ and $s_j = e_{j-1} + 1$

The goal of the main algorithm is to find the "syntactic group" sequence which could be used as baseline chunks. To achieve this goal, we suppose a user defined parameter w representing the ideal number of words contained in a chunk. Based on this parameter, we define a cost function $C(N_j)$ associated to a node:

$$C(N_j) = |(e_j - s_j + 1) - w|$$

By using this cost function, the idea is to locally determine if, by splitting the current group N_0 into smaller syntactic groups $(N_1 \dots N_n)$, we approach the ideal chunk size or not.

This recursive algorithm distinguishes three cases:

1. if N_0 is a leaf, the recursion is stopped and the chunk is defined by the word contained in N_0,
2. if $(C(N_0) < \sum_{j=1}^{n} C(N_j))$ and N_0 is not a leaf, we consider two cases:
   - if $(N_1 \dots N_n)$ contains only leaves then the chunk is defined by N_0,
   - else we try to recompose the node sequence $(N_1 \dots N_n)$ by eliminating leaves.

     To do this, we try to merge each leaf to the preceding group in an incremental way. We identify by $(N'_1 \dots N'_{n'})$ the obtained node sequence. If $(C(N_0) >= \sum_{j=1}^{n} C(N'_j))$ then we apply the current algorithm by considering $N_0 = N'_j$. In the other case the chunk is defined as N_0 without considering the merging step.
3. in other cases, we apply the current algorithm by considering $N_0 = N_j$ for each N_j in $(N_1, \dots, N_n)$;

By applying this algorithm on the previous example, we achieved the segmentation presented in figure 1(b).

5.1.2. *Post-processing*

The previous stage results in a chunk sequence which is not yet optimal. If we consider the example, we can see that some chunks are composed by only one word ("comme"), other chunks contains punctuation in the middle which is not good in a dictation context ("en réalité, c'est").

In order to improve the consistency of the chunks, we have defined a post-processing stage, using a rule-based approach. Three steps are achieved in this stage: a splitting step, a merging step and finally an annotation step.

The splitting step goal is to isolate punctuations and split large chunk according to part-of-speech information. Punctuations have to be isolated because of their special treatment in dictations. Furthermore, for the moment, we split a chunk into two parts only before a coordinating conjunction. By applying this step, we achieved the segmentation presented in figure 1(c).

The merging step goal is to deal with two constraints. The first one aims to avoid any isolated word or any chunk starting with a non-alphabetical character. The objective of the second rule is to assess a minimum number of syllables in each chunk. As we don't have access to a segmentation in syllables (since we deal with a written text), we made the assumption that number of non-consecutive vowels in the text gives an approximation of the number of syllables. The minimum number of syllables in a chunk is then defined in a parametric way. Consequently, we merge a chunk to the previous one if it contains only one word, if it starts with a non-alphabetical character or if the number of consecutive vowels included in both chunks are inferior to the number selected as parameter. The result of this step is presented in figure 1(d).

5.2. Entire text dictation procedure

Once the text has been segmented into chunks, it is possible to automatically dictate it. To do so, the entire text is first copied while using the chunking automatically generated. In a second stage, each chunk is annotated in such a way as to be pronounced in isolation (followed by pause), the punctuation marks being made explicit. This leads for (18) to the final segmentation and the pronunciation given in Fig. 1(d).The text is then given to the TTS system that can treat it, while using a synthesized voice that has been specifically designed for dictation.

6. Conclusion and perspectives

The paper presents a chunking algorithm that allows segmenting any text into chunks that are comparable to the ones we observed in dictation data. Further research is currently achieved in order to (i) decide which procedure is more appropriate depending on the level of the pupils; an (ii) provide a closer analysis of the intonation contours used at the end of the different non-final IPs.

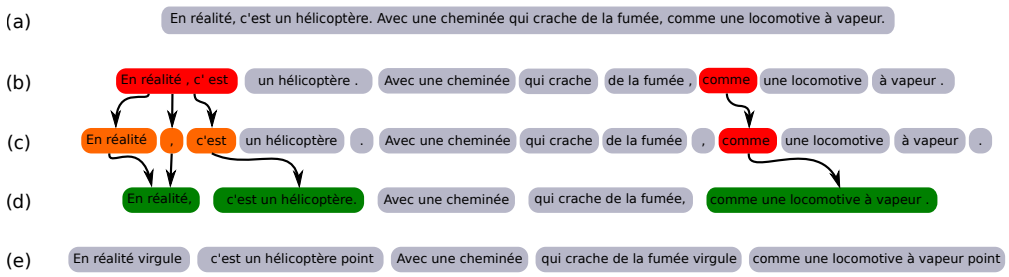


Figure 1: Chunking procedure for (18)

7. Acknowledgement

The work presented here is related to the research project ANR-CONTINT 2011 *PHOREVOX* funded by ANR/CGI.

8. References

[1] "Lectramini," http://www.lectramini.com/.

[2] R. Beaufort and S. Roekhaut, "Automation of dictation exercises. a working combination of call and nlp," *Computational Linguistics in the Netherlands Journal*, vol. 1, pp. 1–20, 2011.

[3] A. Di Cristo, "Intonation in french," *Intonation systems: A survey of twenty languages*, pp. 195–218, 1998.

[4] B. Post, *Tonal and phrasal structures in French intonation*. Thesus, 2000, vol. 34.

[5] P. Martin, *Intonation du français*. A. Colin, 2009.

[6] S.-A. Jun and C. Fougeron, "A phonological model of french intonation," in *Intonation*. Springer, 2000, pp. 209–242.

[7] *Developing a ToBI system for French*. Oxford University Press, Accepted, ch. 3.

[8] S. P. M. Verluyten, *Investigations on French prosodics and metrics*. University Microfilms, 1982.

[9] A. Michelas, "Caractérisation phonétique et phonologique du syntagme intermédiaire en français: de la production à la perception." Ph.D. dissertation, Université de Provence-Aix-Marseille I, 2011.

[10] E. Selkirk, "On derived domains in sentence phonology," *Phonology yearbook*, vol. 3, no. 1986, pp. 371–405, 1986.

[11] F. P. Gerin-Lajoie, "La dictée p.g.l." http://fondationpgl.ca/audio/.

[12] Ladictée, http://www.ladictee.fr/.

[13] P. Boersma and D. Weenink, "Praat (version 5.5)," *Amsterdam: Institute of Phonetic Sciences*, 2012.

[14] Synapse, "Documentation technique: Composant d'étiquetage et lemmatisation," 2011. [Online]. Available: http://www.synapse-fr.com/API/API_Etiquetage_lemmatisation.htm

Within- and Between-Speaker Variability of Parameters Expressing Short-Term Voice Quality

Jitka Vaňková, Radek Skarnitzl

Institute of Phonetics, Faculty of Arts, Charles University in Prague, Czech Republic
jitka.vanka@gmail.com, radek.skarnitzl@ff.cuni.cz

Abstract

This study focuses on short-term acoustic correlates of voice quality. It assesses the within-speaker stability (across different speaking styles) and between-speaker variability of measurements which compare the amplitudes of various spectral events – H1*-H2*, H2*-H4*, H1*-A1*, H1*-A2* and H1*-A3*. Although speakers do differ with regard to the compactness of the parameters in read and spontaneous speaking styles, the parameters H1*-H2*, H1*-A1* and H1*-A2* appear both considerably stable for one speaker in different speaking styles and efficient in between-speaker comparisons. Though not directly applicable in forensic settings, these glottal parameters outperformed vowel formants in classification using LDA.

Index Terms: voice quality, spectrum, speaking styles, Czech

1. Introduction

Voice quality has long been recognized as an independent and full-fledged prosodic dimension [1]. It is a multidimensional phenomenon, which has made it difficult to describe in other than negative terms (i.e., what it is not) [2]. Since Laver's seminal work [3], voice quality has been defined in two ways: narrowly, referring only to the vibration of the vocal folds and its perceptual impact; and broadly, referring also to the perceptual impact of the movements and longer-term settings of supraglottal organs. In this study, we are interested in the narrower sense of the term voice quality, in phonatory modifications.

Since the perceptual evaluation of voice quality is far from straightforward (see e.g. [4], [5]), various acoustic correlates of different aspects of voice quality have been proposed. In this regard, we may talk about both long- and short-term acoustic manifestations of voice quality. The long-term average spectrum (LTAS) shows the frequency distribution of the speech signal over a longer (typically at least 30 seconds) stretch of speech [6], [7], [8]. By averaging over a long portion of speech, spectral differences due to individual segments are evened out, and the method thus yields information pertaining to general voice quality. LTAS has been successfully applied in various phonetic and speech pathological tasks (see e.g. [8] for a summary), with various parameterizations of the LTAS being proposed. Most of these reflect spectral slope, or spectral tilt, in other words the energy decrease with increasing frequency (e.g., the alpha value [9], the Hammarberg index [10], or more recent attempts [11]). In addition, the prominence of a specific peak in the LTAS – called the singer's or speaker's formant – has been correlated with qualities like resonance or sonority of the voice [12], [13].

Short-term manifestations of voice quality may be extracted from individual speech sounds, typically vowels. Jitter and shimmer quantify the degree of fluctuations of the voice source, in the frequency and amplitude domain, respectively; however, it has been suggested that these do not constitute useful correlates of voice quality [14]. Another group of parameters concerns the degree of additive noise in voice: harmonicity, or harmonics-to-noise ratio (HNR) whose measurement has been proposed in various domains [15], [16], [17], the glottal-to-noise excitation ratio (GNE) [18], and others. The last group of parameters to be mentioned here reflects, similarly to the LTAS, spectral slope. In this case, however, it is the spectral slope of a single vowel that is quantified [19], by means of comparing the amplitudes of various events in the acoustic spectrum (see [20] or [21] for an extensive review). The parameter H1-H2 (the amplitude of the first harmonic relative to that of the second) has been correlated with the open quotient. H1-A1 (the amplitude of the first harmonic relative to that of the first formant) is regarded as an indication of F1 bandwidth (B1), which is in turn an indication of the degree to which the glottis fails to close completely during the closing phase. H1-A3 (the amplitude of H1 relative to that of F3) is a reflection of spectral slope. These measures on the speech spectrum are most frequent, although others have also been proposed, such as H2-H4 or H1-A2 (e.g. [22], [23], [24]).

In the current study, we are interested in the stability of voice quality parameters across different speaking styles within one speaker and, at the same time, in inter-speaker differences. We will focus on those parameters which rely on the comparison of the amplitudes of various harmonics (or spectral peaks) – i.e. H1-H2, H1-A1 etc. – one of the reasons for this choice being the fact that Hanson [20] does mention a considerable degree of speaker specificity of these parameters but does not explore this question further.

Theoretically, a low degree of within-speaker variability and a high degree of inter-speaker variability may be useful in phonetic speaker recognition [25], [26], [27]. However, it appears that direct applicability of even such a positive finding in most forensic phonetic casework is problematic, as illustrated by Nolan [28]. The greatest drawback consists in the fact that most recordings of unknown speakers are telephone speech in which (at least) the first two harmonics of male voices are lost. In addition, the widespread use of mobile phones leads to various kinds and levels of background noise. While it is also true that voice quality may differ significantly with speaking style and a mismatch in speaking style thus may lead to false eliminations [28], we believe that the within-speaker variability of parameters like H1-H2 or H1-A3 – which would belong to the long-term segmental strand in Nolan's model [29] – does merit further investigation.

In her study, Hanson [20] suggests that, in order to enable comparison of these measures across different speakers (and vowels), the amplitudes of the first and second harmonic, H1 and H2, need to be corrected for the boosting effects of the first formant (frequency and bandwidth), while F3 amplitude needs to be corrected for the boosting effects of the lower formants. The corrected values are then denoted with an asterisk, thus for instance H1*-H2*, H1*-A1, or H1*-A3*. It will be these corrected measures that will be applied in this study. Specifically, we are interested in the stability of these measures within speakers across speaking styles, as well as in differences across speakers. The performance of the target measures will be compared with that of mean formant values, which will serve as a sort of benchmark here.

2. Method

2.1. Material & subjects

The material for this study was taken from the VASST corpus, which focuses on the variability of speaking styles and which has been collected in various regions of the Czech Republic. Recordings were obtained in quiet rooms in people's homes via a professional portable recorder Edirol HR-09, with a 48-kHz sampling frequency (later down-sampled to 32 kHz). In the present study, we analyzed recordings of spontaneous and read speech produced by six adult male native speakers of Czech aged 28–65 (mean age = 40).

The spontaneous speech sample involved a semi-structured interview in which the speaker was encouraged to speak freely about selected topics. As for the read speech sample, the speakers were asked to read a coherent text in a natural way after sufficient preparation.

We analyzed the Czech open central monophthongs /a/ and /aː/ in various consonantal contexts – only vowels in the context of /ɦ/ were excluded, as the glottal fricative may introduce additional breathiness into the spectrum of the vowel. The boundaries of the target segments were manually adjusted following the suggestions of [30] in Praat [31]. Each token was marked for syllable status with respect to word stress and for its position in utterance (final or non-final). For each speaker and style, we analyzed 50 vowel items, yielding the total of 600 tokens (6 speakers × 2 styles × 50 items). 10 of those had to be removed from analyses since they were not assigned glottal parameter values (see below).

2.2. Parameter extraction and analyses

All parameter values (spectral magnitudes of H1, H2, H4, A1, A2 and A3, as well as the formant frequencies of F1–F4) were automatically extracted by VoiceSauce (VS) [32], [33], a free stand-alone software, using the labelled Praat TextGrids.

To locate and measure the harmonics, VS relies on the extraction of F0. The default algorithm for F0 extraction in VS is STRAIGHT [34], which was also used in our study. In traditional FFT analysis, changing the analysis window can change the features of the extracted spectrum. Here, amplitudes of the harmonics are computed pitch-synchronously (over a 3-cycle window), which eliminates much of the variability in spectra computed over a fixed time window. The method is equivalent to using a very long FFT window but enables considerably more accurate measurements without relying on large FFT calculations [32], [33]. As only male voices were examined, the settings were slightly adjusted: maximum F0 was lowered to 400 Hz and minimum F0 raised to 60 Hz. All other default settings have been preserved.

As for the formant frequencies (F1–F4), they were likewise automatically extracted by VS, using the default algorithm for formant detection, the Snack Sound Toolkit [35]. Snack is an algorithm based on LPC, which uses as defaults the covariance method, pre-emphasis of 0.96, window length of 25 ms, and frame shift of 1 ms, so as to match the F0 estimation by the STRAIGHT algorithm [36].

From the values extracted at 1-ms intervals, the mean value was computed from the middle third (33–67%) of each vowel. Subsequently, 10% of the lowest and 10% of the highest values of the four formants and F0 were manually checked for extraction errors and, if necessary, corrected by direct estimation from the spectrogram (in the case of formants) and the waveform (for F0). Extraction mistakes were not numerous, apart from one speaker whose vowels occasionally manifested diplophonia [37]. The corrected F0 values were then reloaded into VS and the glottal parameters of the corresponding items computed again.

The computation of glottal parameters in VS differs slightly from that mentioned in [20]. VS uses for the corrected measures an algorithm developed by Iseli et al. [38] where H1*-H2* is corrected for the boosting effect of not only F1 but also F2, and F1 through F3 are used for the computation of H1*-A3*. In addition, VS computes H1*-A1* (*cf.* H1*-A1 in [20]).

To investigate the within- and between-speaker variability of these parameters, we performed several analyses. First, the stability of the glottal parameters within a speaker and across the two speaking styles was examined by means of the Kolmogorov-Smirnov (K-S) test, which compares two distributions of values. Second, the K-S test was also applied, in pairwise comparisons, to examine between-speaker variability. Finally, the effectiveness of the glottal parameters to discriminate between speakers was compared to that of formant frequencies by means of Linear Discriminant Analysis (LDA).

3. Results

The main aim of this study was to assess the stability of short-term voice quality parameters across different speaking styles (read and spontaneous) within one speaker, as well as their between-speaker variability. For these purposes, the Kolmogorov-Smirnov (K-S) test was used as it is also sensitive to differences in the general shapes of the distributions (such as differences in dispersion and skewness) in the compared samples.

3.1. Within-speaker stability

Let us first have a look at how stable the parameters are within one speaker. Figure 1 displays for each parameter which of the speakers (labelled S1–S6) did not yield any significant differences across the two styles ($p > 0.05$; above the line,

denoted with a +) and which of the speakers did yield significant differences ($p < 0.05$; below the line, denoted with a –). As we can see, speakers differ with respect to parameter stability: while the parameter differences in the two speaking styles are always insignificant for S1 – i.e., the values are stable in the two styles – S4, on the other hand, yields significant differences in 4 out of the 5 parameters. The figure also suggests that the most stable parameter is H1*-A2* followed by H1*-H2* and H1*-A1*, while H2*-H4* and H1*-A3* appear the least successful in expressing within-speaker stability of our sample.

	H1*-H2*	**H2*-H4***	**H1*-A1***	**H1*-A2***	**H1*-A3***	
				S1		
	S1		S1	S2		
	S2	S1	S3	S3	S1	
	S3	S3	S5	S5	S2	
+	S6	S4	S6	S6	S5	+
–	S4	S2	S2	S4	S3	–
	S5	S5	S4		S4	
		S6			S6	

Figure 1. *Within-speaker stability of voice quality parameters in the two analyzed speaking styles: insignificant differences between the styles appear above the line (also denoted with a +), significant ones below the line (with a –).*

Example distributions are given in Figures 2 and 3; Figure 2 shows the values of parameter H1*-A1* (in dB) for speaker S5 whose distribution did not differ across the two speaking styles ($p > 0.05$), while Figure 3 shows the values of H1*-H2* for speaker S4 which differed significantly ($p < 0.05$).

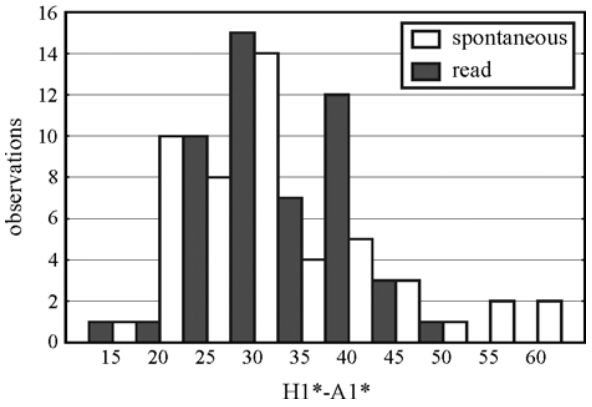


Figure 2. *The distribution of speaker S5's H1*-A1* (in dB) in read and spontaneous speech.*

3.2. Between-speaker variability

We were further interested to what extent these parameters can capture differences between speakers. To illustrate this, we present the results of pairwise comparisons between speakers for the most successful parameter in expressing between-speaker variability, H1*-H2*, in Table 1. The table shows that all 15 possible comparisons – having 6 speakers allows 15 pairwise comparisons – are statistically significant (denoted by an *), most of them highly significant (denoted by **). Moreover, speakers S1 and S5 show statistically highly significant differences from all other speakers, thus being clearly discriminated by their distribution of H1*-H2* values from the others.

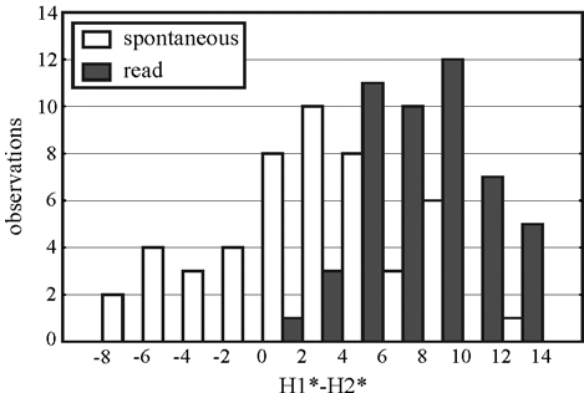


Figure 3. *The distribution of speaker S4's H1*-H2* (in dB) in read and spontaneous speech.*

	S1	S2	S3	S4	S5	S6
S1	X	**	**	**	**	**
S2		X	*	**	**	**
S3			X	**	**	**
S4				X	**	*
S5					X	**
S6						X

Table 1. *Between-speaker variability of H1*-H2* in pairwise comparisons (** $p < 0.001$; * $p < 0.05$).*

Not only H1*-H2* but also the other parameters appear to be efficient in expressing between-speaker differences. The overview of between-speaker pairwise comparisons for all 5 parameters is presented in Table 2. As already stated above, H1*-H2* is the most successful parameter in this respect, though it can be seen that also H1*-A1* yields statistically significant differences for all possible comparisons. H1*-A2* and H1*-A3* perform only slightly worse (one statistically insignificant comparison), while H2*-H4* turns out to reflect between-speaker variability the least, with 4 of the 15 comparisons being statistically insignificant.

	$p < 0.001$	$p < 0.05$	$p > 0.05$
H1*-H2*	13	2	0
H2*-H4*	8	3	4
H1*-A1*	12	3	0
H1*-A2*	14	0	1
H1*-A3*	14	0	1

Table 2. *Significance levels of between-speaker pairwise comparisons for all analyzed parameters.*

Figures 4 and 5 again provide example distributions. Figure 4 shows a similar distribution of H2*-H4* of speakers S1 and S3, while Figure 5 shows distinct distributions of H1*-H2* of speakers S1 and S6 (note that the depicted parameters did not differ in these two speakers across the two speaking styles; see Figure 1). Speaker S6's voice thus appears to be breathier, as suggested by the positive values of H1*-H2* [20].

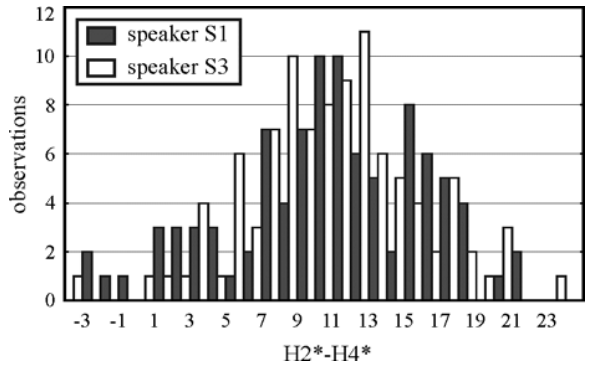


Figure 4. *The distribution of H2*-H4* (in dB) of speakers S1 and S3.*

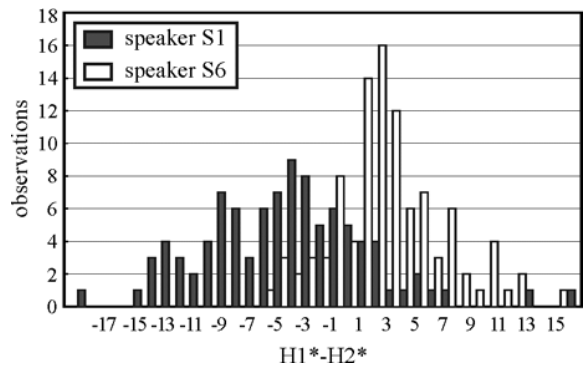


Figure 5. *The distribution of H1*-H2* (in dB) of speakers S1 and S6.*

3.3. Comparison with formant frequencies

Our final objective was to compare the effectiveness of the glottal parameters in discriminating between speakers with that of formants (F1–F4), which are often employed for these purposes and were thus used as a benchmark. This comparison was based on Linear Discriminant Analysis (LDA).

The classification rates for the two sets of parameters are presented in Table 3. The glottal parameters perform slightly better (52.6%) than formants (48.3%), both being well above the chance classification rate of 16.7%. Also, both models are statistically highly significant: F (25,2148) = 36.7; $p < 0.001$ for the glottal parameters, and F (20,1927) = 32.1; $p < 0.001$ for the formants. Table 3 also reveals considerable differences between the discriminability of individual speakers, as well as differences in the performance of the two models for the six speakers, which is most marked for speaker S1.

The values of Wilks' λ complement these tendencies: overall λ equals 0.267 for the glottal parameters and 0.385 for formants, which indicates that the variability in our data is better accounted for by the former.

Speaker	Glottal parameters	Formants
S1	62.5	12.5
S2	39.2	56.1
S3	24.2	39.4
S4	51.5	74.5
S5	84.0	72.0
S6	53.5	34.3
Total	**52.6**	**48.3**

Table 3. *Classification rates (in %) for the glottal parameters and formant values.*

The values of Wilks' λ for the individual parameters also support our findings that the most useful parameter for differentiating between speakers in our model is H1*-H2* as its removal would impair its efficiency the most, while H2*-H4* contributes to its efficiency the least. As for formants, F3 appears to be most and F1 least useful.

4. General discussion and Conclusions

This study analyzed short-term voice quality parameters from the viewpoint of their within- and between-speaker variability, as well as of their potential to discriminate between speakers.

As for within-speaker stability, our results suggested considerable differences between speakers with regard to their compactness in read and spontaneous speaking styles (*cf.* speakers S1 and S4 in Fig. 1). The results also indicate that H1*-A2* is the most stable parameter, followed by H1*-H2* and H1*-A1*. The same parameters also manifested high between-speaker variability (Table 2). This finding points to the importance of the relative amplitude of H1 for distinguishing voice quality (*cf.* [20], [23], [24]).

Between-speaker comparisons also revealed that some speakers are clearly differentiated from all others, specifically speaker S1 and S5 (see Table 1 and also Table 3). Interestingly, S1 was our youngest subject (28 years old), while S5 was our oldest one (65 years old). Our results are thus in accordance with previous studies [38] which showed an age dependency of H1*-H2* and H1*-A3*.

The comparison of the speaker-discriminating potential of the glottal parameters and formant values suggested that the glottal parameters slightly outperform formants overall, though individual differences may be observed. By way of conclusion, let us repeat, however, that the findings cannot be directly applicable in forensic casework due to the band-limited telephone signal, and that our main point of interest was the stability of the voice parameters across the two speaking styles; in this respect, we believe, our study indicated their usefulness for future research.

5. Acknowledgements

This research was supported by the Czech Science Foundation (GACR 406/12/0298) and the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation.

6. References

[1] Campbell, N. and Mokhtari, P., "Voice quality: the 4th prosodic dimension", Proc 15th ICPhS, Barcelona, 2417-2420, 2003.

[2] Kreiman, J. and Sidtis, D., "Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception", Oxford: Wiley-Blackwell, 2011.

[3] Laver, J., "The Phonetic Description of Voice Quality", Cambridge: Cambridge University Press, 1980.

[4] Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A. and Berke, G. S., "Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research", J Speech Hearing Res, 36: 21-40, 1993.

[5] Bele, I.V., "Reliability in perceptual analysis of voice quality", J Voice, 19: 555-573, 2005.

[6] Löfqvist, A., "The long-time-average spectrum as a tool in voice research", J Phon, 14: 471-475, 1986.

[7] Master, S., De Biase, N., Pedrosa, V. and Chiari, B. M., "The long-term average spectrum in research and in the clinical practice of speech therapists", Pró-Fono Revista de Atualização Científica, 18: 111-120, 2006.

[8] Leino, T., "Long-term average spectrum in screening of voice quality in speech: Untrained male university students", J Voice, 23: 671-676, 2009.

[9] Frøkjær-Jensen, B. and Prytz, S., "Registration of voice quality", Brüel Kjær Technological Review, 3: 3-17, 1976.

[10] Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J. and Wedin, L., "Perceptual and acoustic correlates of abnormal voice qualities", Acta Otolaryngologica, 90: 441-451, 1980.

[11] Volín, J., Weingartová, L. and Skarnitzl, R., "Spectral characteristics of schwa in Czech accented English", Research in Language, 11: 31-39, 2013.

[12] Bele, I. V., "The speaker's formant", J Voice, 20: 555-578, 2006.

[13] Leino, T., Laukkanen, A.-M. and Radolf, V., "Formation of the actor's/speaker's formant: A study applying spectrum analysis and computer modeling", J Voice, 25: 150–158, 2011.

[14] Kreiman, J. and Gerratt, B. R., "Jitter, shimmer, and noise in pathological voice quality perception", Proc VOQUAL'03, Geneva, 57-61, 2003.

[15] Yumoto, E., Gould, W. J. and Baer, T., "Harmonics-to-noise ratio as an index of the degree of hoarseness", J Acoust Soc Am, 71: 1544-1550, 1982.

[16] de Krom, G., "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals", J Speech Hearing Res, 36: 254-266, 1993.

[17] Yu, A.-T. and Wang, H.-C., "New speech harmonic structure measure and its applications to speech processing", J Acoust Soc Am, 120: 2938-2949, 2006.

[18] Michaelis, D., Gramss, T. and Strube, H. W., "Glottal-to-noise excitation ratio – a new measure for describing pathological voices", Acta Acustica, 83: 700-706, 1997.

[19] Weingartová, L. and Volín, J., "Spectral measurements of vowels for speaker identification in Czech", Studie z aplikované lingvistiky, 1/2013: 21-36, 2013.

[20] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates", J Acoust Soc Am, 101: 466-481, 1997.

[21] Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y. and Slifka, J., "Towards models of phonation", J Phon, 29: 451-480, 2001.

[22] Kreiman, J., Gerratt, B. R. and Antoñanzas-Barroso, N., "Measures of the glottal source spectrum", J Speech Language Hearing Res, 50: 595-610, 2007.

[23] Keating, P. A. and Esposito, C., "Linguistic Voice Quality", UCLA Working Papers in Phonetics, 105: 85-91, 2007.

[24] Keating, P., Esposito, C., Garellek, M., Khan, S. ud D. and Kuang, J., "Phonation contrasts across languages", UCLA Working Papers in Phonetics, 108: 188-202, 2010.

[25] Nolan, F., "Speaker Recognition and Forensic Phonetics". In W. J. Hardcastle and J. Laver [Eds], Handbook of Phonetic Sciences. Oxford: Blackwell, 744-767, 1997.

[26] Hollien, H., "Forensic Voice Identification", San Diego: Academic Press, 2002.

[27] Jessen, M., "Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs", München: LINCOM, 2013.

[28] Nolan, F., "Forensic speaker identification and the phonetic description of voice quality". In W. Hardcastle and J. Beck [Eds], A Figure of Speech. Mahwah, New Jersey: Erlbaum, 385-411, 2005.

[29] Nolan, F., "The phonetic bases of speaker recognition", Cambridge: Cambridge University Press, 1983.

[30] Machač, P. and Skarnitzl, R., "Principles of Phonetic Segmentation", Praha: Epocha, 2009.

[31] Boersma, P. and Weenink, D., "Praat - Doing phonetics by computer" (Version 5.3.53.). Online: http://www.praat.org, accessed on 11 July, 2013.

[32] Shue, Y., Keating, P., Vicenik, C. and Yu, K., "VoiceSauce: A program for voice analysis", Proc 17th ICPhS, Hong Kong, 1846-1849, 2011.

[33] Shue, Y., "VoiceSauce: A program for voice analysis" (Version 1.14). Online: http://www.seas.ucla.edu/spapl/voicesauce/, last updated on May 30, 2013, accessed on 7 October, 2013.

[34] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A., "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction", Spe Com, 27: 187-207, 1999.

[35] Sjölander, K. (2004), Snack sound toolkit. KTH Stockholm, Sweden. Online: http://www.speech.kth.se/snack, accessed on 10 October, 2013.

[36] VoiceSauce Manual. Online: http://www.seas.ucla.edu/spapl/voicesauce/documentation/parameters.html#formants, accessed on 20 November, 2013.

[37] Cavalli, L. and Hirson, A., "Diplophonia Reappraised", J Voice, 13: 542-556, 1999

[38] Iseli, M., Shue, Y.-L. and Alwan, A., "Age, sex, and vowel dependencies of acoustic measures related to the voice source", J Acoust Soc Am, 121: 2283–2295, 2007.

Do Korean L2 learners have a "foreign accent" when they speak French? Production and perception experiments on rhythm and intonation

Bénédicte Grandon, Hiyon Yoo

Laboratoire de Linguistique Formelle, Labex – EFL, Université Paris Diderot
b_grandon@yahoo.fr, yoo@linguist.univ-paris-diderot.fr

Abstract

French and Korean are two languages with similar prosodic characteristics as far as rhythm and intonation are concerned. In this paper, we present the results of production and perception tests where we describe the prosodic characteristics of Korean L2 learners of French. Our aim is to analyze the impression of "foreign accent" for two prosodic components (intonation and rhythm) of speech produced by Korean L2 learners of French and the perception of this "accent" by native listeners of French (L1). We show that the productions of Korean learners and French native speakers present minor differences but that they do no translate into cues for determining clearly the presence of a "foreign accent".
Index Terms: L2 prosody, intonation, rhythm, production, perception, Korean, French

1. Introduction

Through the past decades, several second language (L2) acquisition models have shown that similarity is more problematic than difference for the L2 learner. Researchers, among others [1], in his Speech Learning Model but also [2] and [3] for intonation and [4] for rhythm, posit that L2 acquisition is more difficult for sounds and prosodic units where the contrast with L1 is poor. What happens then when L1 and L2 share most prosodic characteristics? In this paper, we investigate the impression of "foreign accent" (i.e. non-native production, see among others [5] and [6]) for two prosodic components (intonation and rhythm) in speech produced by Korean L2 learners of French and the perception of this "accent" by native listeners of French (L1).

French and Korean are both described as "syllable-timed" languages ([7] for French, [8] for Korean), with common prosodic features: (1) Primary stress, realized through syllabic lengthening, is located on the last syllable of the last lexical word of a phrase (among others [9] and ([7] for French, [8] for Korean), (2) non-stressed syllables have a constant duration ([7] for French, [10] for Korean), and (3) declarative utterances have a falling pitch contour beginning on the first accented syllable in French [7] and on the utterance's second syllable in Korean [10] and continuing through the end of the sentence. Furthermore, the intonation of modality is seen as the result of the realization of the fundamental frequency (F0) at the end of utterances both in French ([11], [9]) and Korean ([8]). Finally, the declination slope is steadily declining in both languages, no matter the length of the utterances ([12] for Korean and [13] for French).

Since French and Korean share these prosodic characteristics, we expect that (1) French native speakers and Korean learners of French (L2) produce overall similar intonational and durational patterns of vowels at the end of chunks and (2) that these similarities, along with minor differences in the production of Korean speakers cannot be perceived as the manifestations of a "foreign accent". In the next sections, we present the production and perception analyses we conducted to test these hypotheses.

2. Rhythm and intonation in L2 Korean productions of French utterances

In order to determine the prosodic proximity between French L1 and L2 productions, we built a comparative reading experiment with L2 learners of French and French native speakers. The aim of this experiment is to locate the prosodic differences concerning intonational patterns and rhythmic organization that can be seen in L2 productions and to measure variation from native L1 productions.

2.1. Corpus and experimental procedure

The corpus consisted in declarative utterances following the pattern "$NP_{subject}$-V-NP_{Object}", where each chunk has an identical number of syllables (varying from 1 to 10 syllables). Three examples (from the shorter to the longest utterance) are given below in French, with English translation:

(1) utterance 1x3 (9 syllables)
$(\text{Barbara})_{SUBJ}$ $(\text{a perdu})_{V}$ $(\text{son vélo})_{OBJ}$
(Barbara) $_{SUBJ}$ *(lost)* $_{V}$ *(her bike)* $_{OBJ}$
(2) utterance 5x3 (15 syllables)
$(\text{Les amis d'Alice})_{SUBJ}$ $(\text{ont voulu cueillir})_{V}$ $(\text{les pommes du jardin})_{OBJ}$
(Alice's friends) $_{SUJ}$ *(wanted to pick)* $_{V}$ *(the apples from the garden)* $_{OBJ}$
(3) utterance 10x3 (30 syllables)
$(\text{La voisine de ma cousine Annabelle})_{SUBJ}$ $(\text{a vraiment dû hésiter à porter})_{V}$ $(\text{cette affreuse paire de chaussures noires et vertes})_{OBJ}$
(The neighbour of my cousin Annabelle) $_{SUBJ}$ *(must really have hesitated in wearing)* $_{V}$ *(this ugly pair of black and green shoes)* $_{OBJ}$

Two female native speakers of Standard French and four native speakers of Standard Seoul Korean (three female and one male), with variable proficiency levels in French were asked to read the sentences that were presented to them. All speakers were students in their twenties living in Seoul at the time of the recordings. The corpus gathers a total of 20 utterances presented five times in a random order and mixed with distractors that were not taken into account (a total of 600 utterances were analyzed).

The recordings took place in a quiet room, using the Audacity software [14] (in mono, using a sampling frequency of 22050 Hz ad 32bits) on a laptop, with an external microphone. Annotations of the sentences and extraction of durations and F0 values were done first automatically with the Easyalign software [15] and Praat scripts [16], and then checked manually.

Three acoustic parameters were analyzed: rhythm, intonation patterns and declination. For better objectivity in the normalization of the data, we chose the vowel over the syllable as the unit of analysis; thus for rhythm, we used a ratio of the duration of each occurrence produced by the speaker divided by the mean duration of the corresponding vowel in all her/his productions, which allowed us to eliminate both "inter-speaker" (resulting from speakers' different speech rates) and

"intra-speaker" (resulting from different intrinsic vocalic duration) variations, as well as variation resulting from different syllabic structures. For intonation, F0 raw values (three measures per vowel) were converted into semi-tones relative to each speaker's mean F0[1].
The R software [17] was used to run the ANOVA tests.

2.2. Results and discussion[2]

2.2.1. Rhythm

For the final vowel of each chunk (subject, verb and object), we considered that a vowel is lengthened when its normalized duration is 1.2 or above (mean + 20%). We considered that choosing the mean value was not sufficient enough to determine a lengthening compared to the threshold of 1.2, above which lengthening can clearly be perceived
Figure 1 shows the variation of mean vocalic durations for the two groups of speakers (French L1 and Korean L2) for three sentences (3x3=9, 3x5=15 and 3x10=30 syllables).
In most cases, both French L1 speakers and Korean L2 learners lengthen the last vowel of the subject chunk. French speakers almost never lengthen the end of a verb, which shows that they tend to group the verb with its object and to place lengthening only at the end of the sentence. Korean speakers present more diverse results, with vocalic lengthening found in six cases out of ten (sentences with 3x4, 3x5, 3x6, 3x7, 3x8, 3x10 syllables), which might correspond to a more frequent segmentation of the utterance for learners than for native speakers. Vowels at the end of object chunks (which represents also the end of utterances) are systematically longer for French speakers while Korean learners do not produce this expected lengthening (vocalic lengthening of 1,2 can be seen only for sentences with 3x2, 3x4 and 3x10 syllables).
We ran ANOVA tests to compare the realizations of vowels at the end of chunks for the two groups of speakers. Thus, for sentence 3x3 syllables, it appears that the third and the ninth vowels produced by French speakers is longer than the adjacent vowels, even though the normalized values never cross the threshold of 1,2. Results of ANOVA tests show a significant difference for both groups, except for the utterance-final vowel ($F(1,56) = 5{,}737$ $p=.0199$).
The durations of 3x5 syllable utterances, are similar for both groups of speakers: both groups lengthen the 5th and the 15th vowels (for this last vowel, the lengthening is much more important for French speakers than for Korean learners), while the 10th vowel is lengthened only by Korean speakers. Notice that the ANOVA test is significant between the two groups only for the 10th syllable ($F(1,50) = 14{,}657$ $p=.0004$).
For the 3x10 syllable utterances, there is a lengthening of the 10th, 15th and 30th vowels for French speakers and of the 3rd, 10th, 16th, 20th and 30th vowels for Korean speakers. The differences between the two groups are significant for the 10th vowel ($F(1,51) = 5{,}666$ $p=.0211$) and for the 30th vowel ($F(1,51) = 9{,}16$ $p=.0039$) but not for the 20th vowel ($F(1,51) = 0{,}866$, $p=.3565$).

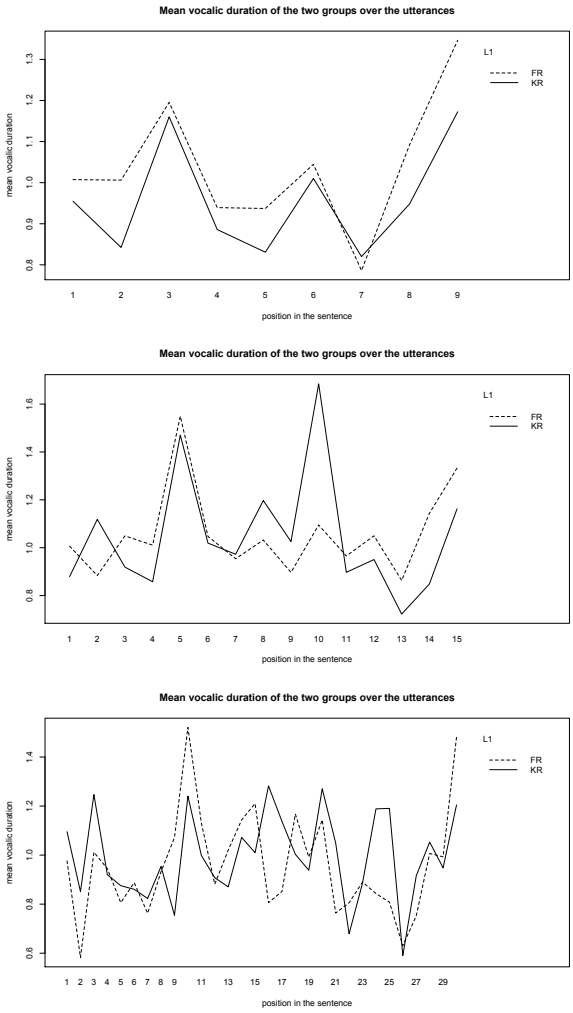


Figure 1: Mean vocalic duration of the two groups for 3x3, 3x5 and 3x10 syllable utterances.

Results show a main difference between the two groups for the final vowel of the utterance, which is lengthened systematically by French speakers while Korean speakers present less homogeneous results for vocalic durations at this position. The findings correspond to what is expected for French rhythm ([7], [9]) where lengthening of the last syllable of an accentual group is a main rhythmic characteristic in French. Lengthening at subject-final level seems to indicate the presence of an accentual group boundary on this position, while the absence of lengthening at verb-final level indicate a grouping of the verb and its object, putting the two chunks into a unique group with only a final lengthening.
Korean learners produce vocalic lengthening less systematically but more frequently. Rhythm is organized in terms of group weight: They have a tendency to re-segment the utterances in order to obtain a maximum of six/seven syllables per group ([9] among others posit that accentual groups in French contain seven syllables), and to lengthen the end of these new groups, no matter the syntactic description. Moreover, lengthening is less marked for Korean learners than for French speakers.

[1] We used the following formula taken in [9] for the computation of semi tones:
F0(ST) = 12*(Log(F0/speaker's_meanF0))/Log(2.00)
[2] ANOVA tests were conducted for every type of sentences, but because of limits of space, we show the results for three utterances, illustrating our purpose.

2.2.2. *Intonation*

For intonation, we compared the F0 values for the last vowel of each chunk (subject, verb and object). Results show that the two groups (French native speakers and Korean learners of French) produce very close patterns.

Thus, at the end of utterances (i.e. at the end of object phrases), both French and Korean speakers produce massively a HL pattern (utterances 3x2, 3x3, 3x6, 3x8, 3x9, 3x10), which can be followed by a small rising (utterances 3x2, 3x6, 3x8). Korean learners have more random productions with more final risings, and even a rising pattern for utterance 3x7. However, the ANOVA test reveals a non-significant difference for F0 realizations of the two groups of speakers on the last vowel of the utterance.

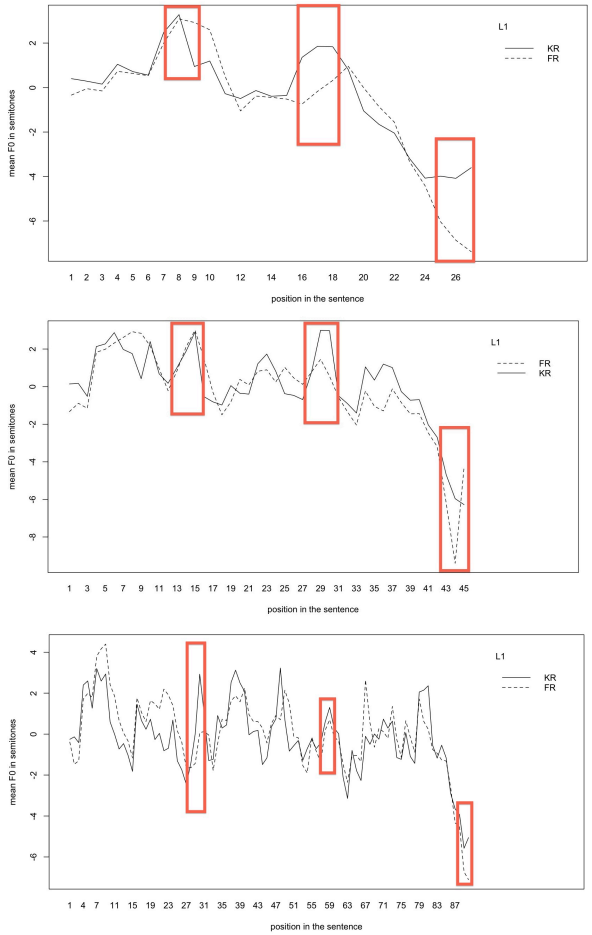


Figure 2: Mean F0 of the two groups for 3x3, 3x5 and 3x10 syllable utterances.

The F0 realization on the last vowel of the subject is also similar for both groups. However, the type of pattern can vary, with a fall-rise pattern for utterance 3x1, a rising pattern for utterances 3x5, 3x6 and 3x10, and a falling pattern for utterance 3x7. The realizations of the two groups of speakers differ for utterances 3x2, 3x3, 3x8 and 3x9 syllables

Results are less homogeneous at the end of verbs. Within the group of French speakers, the last vowel at this position is produced as a flat pattern in short sentences (3x1 and 3x2 syllables), a rising pattern in utterances 3x3, 3x7 et 3x9 syllables, a rise-fall pattern in utterances 3x5 and 3x10 and a falling pattern in utterances 3x4, 3x6 et 3x8. Korean learners realize also the same patterns but not for the same utterances: rising pattern is found in utterances 3x3, 3x5 and 3x7, a flat pattern in utterances 3x1 and 3x8, a rise-fall pattern in utterances 3x2, 3x9 and 3x10, a falling pattern in utterance 3x4, and a falling pattern followed by a rise in utterances 3x6.

Figure 2 illustrates the differences and similarities of the F0 measures for the two groups of speakers (French L1 and Korean L2 speakers) for three utterances (3x3=9, 3x5=15 and 3x10=30 syllables).

Thus, for utterances 3x3, even though the main shape of the F0 curve is similar for the two groups, the ANOVA test reveal significant results for: 1) the last vowel of the verb (at onset ($F(1,39) = 11,479$ $p=.0016$), at mid-point ($F(1,56) = 10,045$ $p=.0025$) and at end-point ($F(1,44) = 9,303$ $p=.0039$) of the vowel) and 2) at end-point of the vowel of the object ($F(1,33) = 5,883$ $p=.0209$).

For utterances 3x5, the ANOVA test shows that there is a significant difference for the verb (at mid-point ($F(1,51) = 14,255$ $p=.0004$) and at end-point of the final vowel of the verb ($F(1,40) = 13,974$ $p=.0006$)) and the object (at mid-point of the final vowel of the object $F(1,43) = 5,453$ $p=.0243$). The ANOVA test is not significant at subject–final level.

For utterances 3x10, the main F0 shapes are similar for both groups. However, the ANOVA test reveals a significant difference at the subject position (at mid-point ($F(1,52) = 11,827$ $p=.0012$) and at end-point ($F(1,35) = 36,713$ $p<.0001$) on the last vowel of the subject), mainly because of the more important rise produced by the Korean learners.

For intonation, it appears that overall F0 shape is similar for both Korean learners and French native speakers, but statistical analysis show significant differences in some positions of contour, especially at the end of the verb and at the end of utterances. The F0 contours produced by Korean learners are less systematic and present more variety than those produced by French speakers.

2.2.3. *F0 declination slope*

We used a regression-analysis in order to measure the overall declination slope of the F0 ([18] and [13]). The slopes of the regression line were calculated by taking into account the three F0 values (converted in semi-tones) on the last vowel of each group. Results given in Table 1 show that the two groups have the same attitude in regards to this prosodic characteristic, with an overall progressing F0 declination from the beginning to the end of utterances.

However, results show that the declination slope is slightly more important for French native speakers than Korean learners (the slopes of regression line are always negative, and have a higher absolute value for the French speakers than for the Korean speakers). This observation can also be seen in the study of [2] who noticed that declination slope is less marked in L2 speakers than L1 speakers. Finally, contrarily to what has been observed by [12] for Korean, there is a correlation between the length of the utterances and the declination slope: when the utterance is longer, the declination slope is less steep, but the highest F0 point at the beginning of the utterance and the lowest point and the end of the utterance remain the same, no matter the length of the utterance.

Since the production analyses revealed slight but statistically significant differences between French native speakers and Korean learners, the next step of our procedure was to test if these differences are perceived, and if they represent sufficient

cues to identify a foreign accent. We present the results of the perception experiment in the next section.

Number of syllables per chunk	Slopes of regression line for French speakers	Slopes of regression line for Korean learners
1	-1,201	-0,722
2	-0,393	-0,377
3	-0,228	-0,187
4	-0,116	-0,110
5	-0,112	-0,080
6	-0,980	-0,690
7	-0,940	-0,048
8	-0,055	-0,032
9	-0,054	-0,029
10	-0,036	-0,025

Table 1: Slopes of regression line for all utterances and for the two groups of speakers (French an Korean)

3. Perception

Since the production experiment showed only slight differences in production in French by French and Korean speakers both for rhythm (vocalic durations) and intonation (F0 contours and declination lines), we expect that these differences in the Korean speakers' production at a prosodic level are not sufficient enough to be perceived by naive French listeners as a « foreign accent » in comparison to the French speakers' production. In this section we present the results of a pilot experiment built to test the perception of this "accent".

3.1. Experimental procedure

The perception experiment is built in three different conditions. First, we neutralized segmental information: we synthesized the intonation and rhythm of two French and two Korean speakers who participated at the production experiment with the voice of a French native speaker, which allowed us keeping segmental information but changing the prosodic profile of the utterances (we thus obtained synthesized sentences with different prosodic profiles on a same French voice). Second, segmental and rhythmic information have been neutralized: thus, we coupled a long /a/ sound with the intonational tier of the four speakers used in the first condition while rhythm is neutralized for all syllables: in this test, we chose to keep only part of the initial stimuli (3x4, 3x6 and 3x8 syllables were not used) since the listening task of utterances with no segmental information is quite difficult to perform for naïve listeners. The third test is the control condition: stimuli remain unmodified productions of the same speakers, without any manipulation of segmental information, rhythm and intonation. The perception experiment is created and run on Praat [16]. Eight native speakers of French took part in the experiment which was conducted on a laptop and using headphones. The experiment is a forced choice task in which they are asked to judge if the stimulus is pronounced by a speaker with a foreign accent or by a French native speaker.

3.2. Results and discussion

When segmental information has been neutralized and only rhythm and intonation remain (test 1 of Tables 2 and 3), answers for "foreign accent" are only slightly above chance (62.5 to 67.2%) for shorter sentences (1 to 3 syllables per group). Results gradually improve for longer sentences (4 to 9 syllables per group) and reach 87.5% for "foreign accent".

When only the parameter of intonation is kept as the only cue, results for shorter sentences are low (42.2 to 54.7% for "foreign accent"), improve when sentences are longer (5 and 7 syllables per group) but remain around average for the longest tested sentences (9 syllables per group). The analyses of answers chosen per stimulus show that listeners identify the stimuli of native speakers of French correctly in only 42.7% of cases, which indicate that in this test, choices are made randomly.

In the control condition (Test 3 of Tables 2 and 3), results are as expected for all types of sentences (above 95.3%): segmental variation has not been analyzed in this study, yet it seems to be strong enough for the listeners to identify a foreign accent.

Number of syllables in the sentence	**Condition**		
	Test 1	Test 2	Test 3
3x1 = 3	62,5%	53,1%	100%
3x2 = 6	67,2%	42,2%	98,9%
3x3 = 9	62,5%	54,7%	95,3%
3x4 = 12	82,8%	-	98,4%
3x5 = 15	78,1%	67,2%	100%
3x6 = 18	73,4%	-	96,9%
3x7 = 21	79,7%	60,9%	96,9%
3x8 = 24	87,5%	-	100%
3x9 = 27	87,5%	50%	87,5%

Table 2: Percent of answers for "foreign accent" for the three conditions for each utterance (1/L2 rhythm and intonation, 2/Intonation only, 3/ no modification)

		Stimulus produced by	
		French speaker	Korean speaker
Answers for Test 1	French speaker's production	82,6%	31,3%
	Foreign Accent	17,4%	68,7%
Answers for Test 2	French speaker's production	42,7%	33,3%
	Foreign Accent	57,3%	66,7%
Answers for Test 3	French speaker's production	95,5%	1,7%
	Foreign Accent	4,5%	98,3%

Table 3: Results sorted by types of stimuli (produced by French or Korean speaker)

4. Conclusion

The production experiment showed that as far as rhythm is concerned, Korean learners of French tend to change boundaries and to lengthen the end of these groups, no matter the syntactic structure. Otherwise, strategies for intonation and declination slope are quite similar for both language groups.

This pilot perception experiment gives a first indication of the difficulty for native listeners to clearly perceive slight, randomly distributed differences in the production of L2 learners of French. It also reveals that the identification of "foreign accent" is easier when utterances are longer. This result is consistent with [3]'s results on German and US-English production and perception. It appears that the perception of a "foreign accent" is linked to prosodic proximity of the two tested languages. Ongoing work includes more listeners but also the dissociation between the two main parameters, rhythm and intonation. Moreover, a further production and perception study with the same material in French with speakers of a typologically different language could help understanding and confirming our findings.

5. References[i]

[1] Flege, J., 1995, “Second Language Speech Learning: Theory, Findings, and Problems”. In W. Strange (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research (pp. 233 – 272). Timonium, MD: York Press.

[2] Mennen, I., 2007. “Phonological and Phonetic Influences in Non-native Intonation”, In J. Trouvain and U. Gut (Eds.): Non-Native Prosody - Phonetic Description and Teaching Practice. Mouton De Gruyter, Berlin, pp 53-76

[3] Jilka, M., 2007, “Different Manifestations and Perceptions of Foreign Accent in Intonation”, in J. Trouvain and U. Gut (Eds.): Non-Native Prosody - Phonetic Description and Teaching Practice. Mouton De Gruyter, Berlin, pp. 77 – 96

[4] Barry, W.J., 2007. “Rhythm as an L2 Problem: How prosodic is it?” In J. Trouvain and U. Gut (Eds.): Non-Native Prosody - Phonetic Description and Teaching Practice. Mouton De Gruyter, Berlin, pp 97-120

[5] Piske T., MacKay, I.R.A, Flege, J.E, 2001, “Factors affecting degree of foreign accent in an L2: a review? Journal of Phonetics 29:191-215

[6] Vaissière, J. & Boula de Mareüil, 2004.“Identifying a language or an accent: from segment to prosody, in Proceedings of the Modelling for the Identification of Languages (MIDL) Workshop, Partis 1-6

[7] Di Cristo, A., 1999, “Intonation in French”. In Daniel Hirst and Albert Di Cristo (eds), Intonation Systems: A Survey of Twenty Languages. Cambridge: Cambridge University Press

[8] Jun, S.-A., 1993, “The Phonetics and Phonology of Korean Prosody”, PhD Dissertation, Ohio State University

[9] Martin, P., 2009, “Intonation du Français”, Paris, Armand Colin

[10] Lee, H.-Y., 1990, “The structure of Korean Prosody”, PhD Dissertation, University College of London

[11] Delattre, P., 1966, “Les dix Intonations de Base du Français”, French Review 40, p1-14

[12] Ko, D-H., 1988, Declarative Intonation in Korean: An Acoustical Study of F(o) Declination, Ph.D. Thesis, U. of Kansas

[13] Schmid, C. Gendrot, C., & M. Adda-Decker. 2012. "Une comparaison de la déclinaison de F0 entre le français et l'allemand journalistiques" Actes des 28èmes Journée d’Etude sur la Parole, Grenoble, France, 4-8 juin 2012, pp. 329–336.

[14] Audacity Version2.0, retrieved 01 November 2013 from http://audacity.sourceforge.net

[15] Goldman J.-Ph. 2011. « EasyAlign: an automatic phonetic alignment tool under Praat » in Proceedings of InterSpeech, September 2011, Firenze, Italy

[16] Boersma, P. & Weenink, D., 2013. “Praat: doing phonetics by computer” [Computer program]. Version 5.3.59, retrieved 20 November 2013 from http://www.praat.org/

[17] R Development Core Team (2012). R: A language and environment for statistical computing. R, Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[18] Lieberman P. et al., 1985. “Measures of the sentence intonation of read and spontaneous speech in American English" Journal of the Acoustical Society of America 77, 649-657.

[i] This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the “Investissements d’Avenir”-Labex EFL program (reference: ANR-10-LABX-0083)

Prosodic processing in the first year of life: an ERP study

Linda Garami[1], Anett Ragó[2,1], Ferenc Honbolygó[1,2], Valéria Csépe[1]

[1] Institute of Cognitive Neuroscience and Psychology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary
[2] Institute of Psychology, University of Eötvös Loránd, Budapest, Hungary
garami.linda@ttk.mta.hu, www.humlab.cogpsyphy.hu

Abstract

From early months of life prosody has a prominent contribution to segmentation: prosodic boundaries overlap with syntactic ones and facilitate the extraction of syntactic regularities both at word and at phrase level. Therefore, the long-term representation of rhythmic features of the native language, especially the stress templates derived from regularities are assumed to play a particular role in pre-lexical processing. We examined the nature of early stress representation in a language with a fixed stress pattern in an electrophysiological experiment (acoustic passive odd-ball paradigm, 10 month-olds: 28 infants; 6 month-olds: 21 infants, 400 items, deviant: p=20%) using bi-syllabic Hungarian pseudo-words to follow how prosodic features contribute to processing saliency and how word stress templates based on regularities may emerge. We used legally and illegally stressed stimulus both in standard and deviant positions in separate conditions.

In the legal standard condition two mismatch responses (MMRs) temporally synchronized to each syllable could be recorded. On the contrary, in the illegal standard condition no significant response was found. It seems that language environment influences the processing of speech prosody and the MMR correlates of word stress processing are related both to saliency and to stress templates emerging during the first year of life.

Index Terms: language development, prosody, mismatch response

1. Introduction

Infants acquire their native language quickly and accurately among suboptimal conditions, as the utterances of adults around them are usually imperfect, non-segmented and changing with personal characteristics. One of the first problems infants need to solve is how to extract meaningful units from the continuous speech stream, without having any lexical knowledge. Originally Gleitman and Wanner [1] then later Anne Christophe and her colleagues [2] proposed that since prosodic cues like lexical stress may reliably signal word boundaries they can act as bootstrapping mechanisms by helping to locate words in the speech stream. These presumed prosodic bootstrapping mechanisms help infants to generate and apply rules extracted from distributional patterns of spoken language [3, 4, 5]. Therefore prosody, among other cues, such as allophonic, phonotactic and statistical or distributional cues [6], is assumed to facilitate early language acquisition. The extraction of prosodic regularities is based on acquiring the rhythmic segmentation procedure that allows infants segmenting their first patterns [7, 8, 9] which soon issue in that infants show a strong trochaic bias in languages that apply this pattern (i.e. English, German or Hungarian) in contrast to infants whose native language applies an iambic one. This implies that infants are able to detect specific patterns of distinct acoustic features related to prosody, and that the exact nature of this mechanism is mostly dependent on the native language specificity. Our research focuses on the development of representation resulted from extracting language-specific patterns.

The early segmentation hypothesis relies on adults' segmenting ability. According to the electrophysiological data of Honbolygó and Csépe the perception of stress pattern is based on long-term, pre- lexical, language-specific representations of stress information among Hungarian adults that they called stress templates [10]. Development of such a template should occur already in early infancy, as infants are sensitive to specific salient acoustic features and are able to extract rules given in patterns quite early. However, neither the relationship of these acoustic cues nor the developmental timing of them is clear yet. In our study we were interested in when the representation of native language's specific prosodic pattern signaling word boundaries could be abstracted and how these patterns are utilized when listening to new utterances. Hungarian language has some special characteristics as it is a syllable-based language like French, although has a strong stress initial rule as German, while vowel duration is a segmental feature in contrast to both.

Studies using electrophysiological methods refined our understanding of prosodic information by examining the underlying mechanisms without relying on behavioural responses. This method enables us to filter out the behavioral readiness of infants, as it measures only the brain activity correlating with automatic linguistic process that is independent of willpower. Most of the electrophysiological experiments applying the method of passive oddball paradigm are looking for deflections in event-related brain potentials (ERP) related to stress processing. Among adults, the Mismatch Negativity (MMN) event-related brain potential component is elicited when an unexpected change occurs in the auditory environment. The expectations can be generated either via forming a short-term memory trace of the actual stimulus set (with a series of identical stimuli) or resulted from this trace affected by long-term memory representations [11]. The unexpected change of suprasegmental information (prosody) evokes mismatch responses, where latency, amplitude, and polarity, all depend on the familiarity of the stimulus [12]. A cross-linguistic study [13] underpinned that language environment promotes different discrimination abilities no later than at the age of 4 months as French and German infants showed different ERP waveforms to different stress patterns by four-months of age. This bias was congruent with the deviation of language related stress cues and regularities: while in German the dominant stress pattern is stress on the first syllable, in French stress is on the second one (among bi-syllabic words). In an oddball paradigm infants responded with a positive mismatch response (MMR) only to the illegally stressed pseudo-words of their own native

language. The authors used salient cues typical for German such as duration and intensity. However, syllabic stress is not determined by the same acoustic features in other languages, so stress is not uniform from an acoustic point of view. For example infants show high sensitivity to duration [14, 15] very early, while intensity or other features may show a different developmental trajectory. Dividing salient acoustic feature processing and prosodic template processing is a critical issue in all languages. The most important acoustic features defining stress pattern vary across languages, so discriminating phoneme duration or pitch or f0 alone does not allow us to shed light on detecting stress pattern deviations. This ability is derived as a rule extraction based on the early language experience. The ERP studies focusing on prosodic information varied mostly the duration of speech sounds; consonants [16] or vowels [13, 17, 18]. In Hungarian vowel duration is a segmental cue, where young learners have to extract a typical stress pattern while taking complex variations (acoustic perturbations in case of long vowels) into account. However, we took advantage of the expressed regularity characteristic for Hungarian with word-initial stress for all words (except for compound words), a pattern correlating with linguistic units, and therefore providing a strong unvarying environment for forming stress-related expectations [19, 20].

As lexical stress used in the Hungarian language (always word-initial) is associated perfectly with the words' beginning, this allows us to follow how these features contribute to processing saliency and how word stress templates based on regularities might emerge.

The objective of our ERP study was to reveal the ability 6 and 10 month-old infants to use acoustically rich stress information, including pitch and f0 changes, the two most important features characteristic for syllabic stress in Hungarian. We assumed that besides processing salient distinct features only, 6 month old infants have already extracted a specific rule regarding the stress pattern of their native language and violation is detected on a higher level as a long- term representation of stress.

2. Methods

We used the experimental paradigm of Honbolygó and Csépe [1] designed for testing the saliency versus template hypothesis. Legally and illegally stressed pseudo-words were presented in a passive oddball paradigm. We recorded ERPs in two conditions: the standard and deviant stimuli were simply reversed in order to see if only the absolute differences lead the discrimination or the relative pattern was also taken into account.

2.1. Participants

A total of 60 infants were recruited for the experiment, 48 were included in the statistical analyses. 12 infants were excluded due to extensive artifacts. The recordings were taken at the mean age of 196 days (*SD*=13) in the 6 month-olds' group and 316 days old (*SD*=13) in the 10 month-olds' group. Mean GA was 39.2 weeks while mean birth weight was 3346 g with 548 g standard deviation, with no significant difference between the two age groups. All infants were born to monolingual families and raised in a monolingual environment. None of them had known hearing problems, neurological impairments or any known developmental delay. Parents gave written consent for their child's participation after having detailed information. The experiment was approved by the Ethical Review Committee for Research in Psychology.

2.2. Stimuli

Two types of stimuli, stress variants of a Hungarian pseudo-word ('bebe') with duration of 539 ms, were used. The two stimulus types differed only in their stress pattern. In Hungarian stress is always on the first syllable [21], so for the legal version stress was on the first syllable ('BEbe'). The illegal version was created by reversing the order of the two syllables ('beBE') in order to minimize the difference between the two pseudo-words apart from stress pattern (using Praat [22]). The two syllables were spoken digitalized utterances that differed in three features: maximum intensity (2.42 dB), maximum f0 (15.77 Hz) and rise time (16 ms) respectively, and in contrast to former studies, not in duration. The source stimulus was uttered by a native female speaker in sentence context and edited for computing its illegal variation (for further details read Honbolygó and Csépe [10]).

2.3. Procedure

The procedure was the same as in the Honbolygó and Csépe [1] study. Stimuli were presented in a passive oddball paradigm in random order (deviant probability of 20%). In order to avoid rhythmic affects stimulus onset asynchrony (SOA) varied randomly between 730 and 830ms. We used two stimulus presentation conditions:

• *Legal standard* condition: the legal stimulus was the standard and the illegally stressed pseudo-word was the deviant stimulus.

• *Illegal standard* condition: the stimulus positions were changed, the illegal stimulus was presented as the standard and the legal one was the deviant stimulus.

The order of the two conditions was counterbalanced across subjects. Each condition contained 100 deviants in two blocks. The adult version of the experiment was shortened in order to adapt it to the infant participants [23]. Recording lasted approximately 12 minutes, in order to avoid fatigue among the infants. Stimuli were presented via loudspeaker (Soundkey MS-310, 70 *dB*) that was placed at the distance of 100 cm from the subjects. The experiment was performed using Presentation software (version 12.1, http://www.neurobs.com). The total experimental time was 1 hour including preparation and pauses. Infants were sitting on their parents' lap and were kept calm by presenting cartoons and puppets silently by an assistant.

2.4. Data collection

The EEG was recorded at 500 Hz sampling rate with Ag-AgCl electrodes using an appropriate sized electrode cap (BrainVision Recorder, BrainAmp amplifier, BrainProducts GmbH). Electrodes were attached to F3, Fz, F4, C3, C4, T3, T4, P3, Pz, P4, O1, O2, M1, M2, the reference electrode was Cz. Ground was placed between Fz and Fpz on the midline. The electrode locations corresponded to the international 10-20 system. Offline data analyses were done with BrainVision Analyzer software (BrainProducts GmbH). Recordings were re-referenced to the average reference of M1 and M2 and were band-pass filtered (0,5-20 Hz, 24 dB/oct). Raw EEG data were segmented into 800 ms epochs, including a 100 ms pre-stimulus baseline. Electrophysiological responses to deviant

stimuli and to standard appearing right before the deviant were taken into analysis. EEG responses exceeding ±150μV within a sliding window of 300 ms in any channel were rejected automatically. 12 infants' data were rejected because of contaminated recordings. Statistical analyses were carried out in 21 six-month-old and 27 ten-month-old infants for two time windows (300-400 ms, 450-550 ms) based on the grand averages. Epochs were averaged separately for each condition, electrode and participant for all deviants, and for the preceding standards.

3. Results

Difference waves computed by subtracting the averaged standard responses from the deviants are displayed on Figure 1.

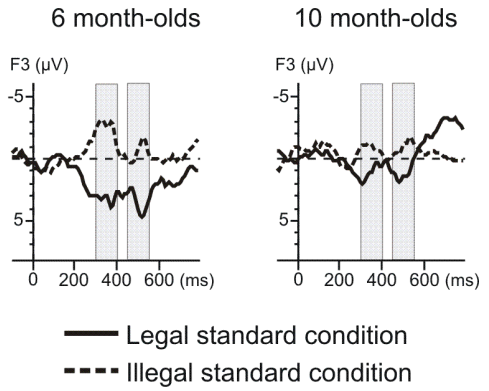


Figure 1: *Difference waves shown for the F3 electrode in two experimental conditions. Responses obtained in the legal standard condition are shown with thick lines, in the illegal standard condition with dotted lines. The time windows -– marked by grey bars show the latency ranges (300-400ms and 450-550ms) where the statistical analysis was run.*

For statistical analysis average areas were computed in the two time windows (300-400 ms; 450-500 ms – right after the syllables). We analyzed the main effect of the conditions performing a 3×2×2 mixed ANOVA with within factors Electrodes (F3, Fz, F4) and Stimulus (standard vs. deviant) and Age (6 MO vs. 10 MO) as a grouping variable in both legal standard and illegal standard conditions. Because of a possible violation of the sphericity assumption we used Greenhouse-Geisser (G-G) adjusted univariate tests where it was necessary [24]. Main effects are reported, effect size were calculated.

In the legal standard condition a Stimulus main effect was revealed confirming the presence of a positive mismatch response (MMR) in both time windows (300-400 ms: $F(1,46)=4.96$, $p<.05$, $r=.31$; 450-550 ms: $F(1,46)=4.76$, $p<.05$, $r=.31$). No Age effect ($p=.76$), or any other significant effect was found.

The statistical analysis did not reveal any significant effect in the illegal standard condition.

4. Discussion

Different ERPs were obtained in the different conditions (role as standard or deviant changed) with the two stimuli. While in the legal standard condition two MMRs were seen, though without any age difference, no significant difference was found in the illegal standard condition. If the responses were based on simple salient acoustic processing only, the automatic change detection reflected by the MMRs would rely on short-term (often called sensory) trace of the stimuli contributing to very similar responses in the two conditions, as the physical differences to detect were the same In contrast, repeating the legal stress pattern as standard enhanced the detection accuracy, so that the illegal pattern used as deviant could activate both a comparison with the long-term representation and with the actual trace of the salient acoustic feature of the stressed initial syllable Moreover, participants had difficulties in detecting stimuli when the illegal version served as standard and the legal one was contrasted with it. In the illegal standard condition we repeated a pattern that contrasted with the long-term representation even if only a weak template has emerged yet, so contrasting it with the legal form as deviant could rely neither on saliency nor on template. Therefore, we interpret our MMR data as electrophysiological evidence of a delicate developmental stage where saliency and emerging stress template representation show a particular dynamics in the first year of life.

Conclusion

Fitting in with the former electrophysiological results in case of other stress- and syllable-timed languages our results have strengthen the view that language specific environmental cues strongly influence the early sensitivity to suprasegmental patterns. Hungarian is a syllable-based language as French, with a word-initial stress pattern as German. Hungarian infants hear trochaic pattern more often, their result are in line with the German infants. Our results imply the existence of a possibly universal developmental pattern of rule abstraction. Relying on rhythmic segmentation procedure infants extract language-specific patterns from their environment, and start to generate expectations. The exact timing of the emergence of a long-term representation is still not clear, although the development of this processing seems to occur in a rather early stage of language acquisition. However, the crucial features of the assumed template as well as their interference with other acoustic attributes of spoken utterances are still unclear. There is a great difference of processing the cues depending on the native language characteristics. According to earlier results, stress based languages provide good conditions for early discrimination of stress, however there are controversial differences regarding syllable based languages [25]. Further investigational question is the nature of long-term stress representation in case of syllable-based languages in accordance with earlier results [26, 27].

5. Acknowledgements

We thank to our colleagues, Gabi Baliga, Kinga Kreif, Orsolya Kolozsvári, Ágoston Török and Andrea Kóbor for their valuable contribution as well to the participating families.

This study was supported by the Hungarian Research Fund project (OTKA-NK No. 101 087, PI: Valéria Csépe).

6. References

[1] Gleitman, L. and Wanner, E. (1982). "Language acquisition: The state of the state of the art". In E. Wanner and L. Gleitman [Ed], Language Acquisition: The state of the art, 3-48, Cambridge University Press, 1982.

[2] Christophe, A., Dupoux, E., Bertoncini, J. and Mehler, J., "Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition", J. Acoust. Soc. Am., 95(3):1570–1580, 1994.

[3] Gervain, J. and Mehler, J., "Speech perception and language acquisition in the first year of life", Annu. Rev. Psychol., 61:191–218, 2010.

[4] Yoshida, K. A., Iversen, J. R., Patel, A. D., Mazuka, R., Nito, H., Gervain, J. and Werker, J. F., "The development of perceptual grouping biases in infancy: a Japanese-English cross-linguistic study" Cognition, 115:356–361, 2010.

[5] Gervain, J., Nespor, M., Mazuka, R., Horie, R. and Mehler, J., „Bootstrapping word order in prelexical infants: a Japanese-Italian cross- linguistic study", Cogn. Psychol., 57:56-74, 2008.

[6] Christiansen, M. H., Allen, J. and Seidenberg, M. S., „Learning to segment speech using multiple cues: A connectionist model" Lang. Cognitive Proc., 13:221–268, 1998.

[7] Nazzi, T., Bertoncini, J. and Mehler, J., "Language discrimination by newborns: Toward an understanding of the role of rhythm", J. Exp.l Psychol: Human Perception and Performance", 24(3): 756–766, 1998.

[8] Nazzi, T., Kemler Nelson, D. G., Jusczyk, P. W. and Jusczyk, A. M., "Six month olds' detection of clauses embedded in continuous speech: effects of prosodic well-formedness", Infancy 1: 123–147., 2000.

[9] Nazzi, T., Iakimova, G, Fredonie, S. and Alcantara, C., "Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences", J. Mem. Lang, 54(3): 283–299., 2006.

[10] Honbolygó, F. and Csépe, V., "Saliency or template? ERP evidence for long-term representation of word stress", Int. J. Psychophysiol., 87:165-172, 2013.

[11] Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csépe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, H., Iivonen, A. and Näätänen, R., "Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations", Cogn. Brain Res., 73:357-69, 1999.

[12] Winkler, I., Denham, S. L. and Nelken, I., „Modeling the auditory scene: predictive regularity representations and perceptual objects", Trends Cogn. Sci., 1312:532-540, 2009.

[13] Friederici, A. F., Friedrich, M. and Christophe, A., "Brain responses in 4-month old infants are already language specific", Curr. Biol., 17:1208–1211, 2007.

[14] Kushnerenko, E., Ceponiene, R., Fellman, V., Huotilainen, M. and Winkler, I., „Event-related potential correlates of sound duration: similar pattern from birth to adulthood", NeuroReport, 12(17):3777-3781, 2001.

[15] Trainor, L. J. and Adams, B., "Infants' and Adults' use of duration and intensity cues in the segmentation of tone patterns", Percept Psychophys., 62: 333–340., 2000.

[16] Leppänen, P. H. T., Richardson, U., Pihko, E., Eklund, K. M., Guttorm, T. K., Aro, M. and Lyytinen, H., "Brain responses to changes in speech sound durations differ between infants with and without familial risk for dyslexia", Dev. Neuropsychol., 22:407-422, 2002.

[17] Friedrich, M., Herold, B. and Friederici, A.D., "ERP correlates of processing native and non-native language word stress in infants with different language outcomes", Cortex, 45:662–676, 2009.

[18] Weber, C., Hahne, A., Friedrich, M., Friederici and A. D., „Discrimination of word stress in early infant perception: electrophysiological evidence", Cogn. Brain Res., 18:149-161, 2004

[19] Cutler, A. and Foss, D. J.,"On the role of sentence stress in sentence processing", Lang. Speech, 20: 1–10, 1977.

[20] Cutler, A. and Norris, D., "The role of strong syllables in segmentation for lexical access", Journal of J. Exp. Psychol.: Hum. Percept. Perform, 14(1):113–121, 1988.

[21] Siptár, P., Törkenczy, M., The phonology of Hungarian. Oxford University Press, 2007.

[22] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", retrieved from http://www.praat.org/, 2007.

[23] Rivera-Gaxiola, M. and Silva-Pereyra, J., Kuhl, P. K., "Brain potentials to native and non-native speech contrasts in 7- and 11-month-old American infants", Dev. Sci., 8:2, 162–172, 2005.

[24] Greenhouse, S. W. and Geisser, S., "On methods in the analysis of profile data", Psychometrika, 242:95-112, 1959.

[25] Thiessen, E. D. and Saffran, J. R., "When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants." Dev. Psychol., *39*(4): 706–716, 2003.

[26] Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J. and Nazzi, T., "Language specific prosodic preferences during the first half year of life: evidence from German and French infants." Infant Behav. Dev., 32:262–74, 2009.

[27] Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., Limissuri L. A. and Peperkamp, S., "Language-specific stress perception by 9-month-old French and Spanish infants." Developmental Sci., 12(6):914–919, 2009.

The Inadequacy of Rhythm Metrics to Quantify L2 Suprasegmental Characteristics

Lei He

Phonetics Laboratory, University of Zurich, Switzerland

lei.he@uzh.ch

Abstract

The study investigated the L2 speech rhythm of Chinese English speakers (L1 = Mandarin) using the metrics of ΔV, ΔC, %V, VarcoV, VarcoC, rPVI-C and nPVI-V. Five native speakers of American English and Mandarin were recruited to record five sentences in English. In addition, the Chinese speakers also recorded five Mandarin sentences. One-way ANOVAs were conducted to see if significant differences exist on each of the metrics among L1 English, L2 English and L1 Mandarin. Results show that the two L1's are categorically distinct on all metrics, conforming to the perceptually distinct rhythmicities of English and Mandarin. However, no significant differences were found between L1 and L2 English (which have different intuitive rhythmicities) on almost all the metrics, suggesting that the metrics are inadequate to capture the suprasegmental details that give the final make-up of speech rhythm. Finally, new directions of speech rhythm research and new applications of the rhythm metrics are sketched.

Index Terms: rhythm metrics, inadequacy, L2 English

1. Introduction

Most human communities have certain speech styles that are constrained to fit an external or imposed periodic intervals or beats, manifesting rhythmic patterns [1, 2]. This music-like feature makes poetry and nursery rhymes possible. Apart from this artificially created artistic feature of speech rhythm, languages, in their non-artistic forms, have repetitive patterns that are at least intuitively detectable. Moreover, rhythm is among the first acquired phonological features in first language acquisition. According to [3], "during the last trimester of intrauterine development, the fetus is known to be actively processing the sound of its mother's speech." After being filtered through the amniotic fluid (analogous to a low-pass filter), the fetus can only recognize the "melody and rhythm of the language" [3: 43]. Through the non-nutritious sucking technique, researchers discovered that neonates were able to distinguish rhythmically different languages from their mother tongue [4, 5, 6]. Perceptual experiments among adults and monkeys [7, 8, 9] also yielded similar results that languages of different rhythmicities are distinguishable, whereas rhythmically similar languages are not discernable.

Early researchers [10, 11] proposed two major classes of speech rhythm, i.e. "Morse-code" and "machine-gun" rhythm, which correspond to the widely used terminologies as "stress-timed" and "syllable-timed" respectively mentioned in [12: 54] as simple rhythm units. [13] adopted this dichotomy and claimed that languages either have isochronous feet (i.e. inter-stress intervals) or isochronous syllables (i.e. inter-syllable intervals). However, later investigations on the acoustic signals, nevertheless, failed to find exact isochrony in neither inter-stress nor inter-syllable intervals, for example [14], [15]. [16] even rejected "stress-" and "syllable-timing" as metalinguistic terms, due to the failure to find true isochrony instrumentally, and claimed that the rhythmic differences between languages were the result of phonologic rule idiosyncratic to different languages.

Departing from finding absolute syllabic or foot isochrony, researchers began to delve into the structural characteristics of languages and proposed that the two types of languages have varied degrees of vowel reduction and different complexities in syllable structures: Germanic languages such as English and German tend to reduce or centralized unstressed vowels and have more complicated syllable structures, whereas Romance languages such as French and Italian normally do not have obvious vowel reductions and have less syllable weights [14, 17].

[8] quantified this idea by calculating the durational standard deviations of vocalic intervals (linear composition of adjacent vowels) and consonantal intervals (linear composition of adjacent consonants) in an utterance (ΔC and ΔV respectively), and the proportion of vocalic duration out of the whole utterance (%V). Instead of measuring the global variability of interval durations, [18] and [19] averaged the durational differences between consecutive vocalic or consonantal intervals, and called their metrics the pairwise variability indices (PVIs). Moreover, the calculation of vocalic PVI is normalized (nPVI-V) to account for tempo changes, and the raw PVI (rPVI-C) is retained to calculate consonantal PVI. [20] also normalized the speech rate by taking the ratio between ΔC (or ΔV) and the mean duration of the intervals being analyzed (VarcoC and VarcoV). These rhythm metrics have fair success in categorizing canonical "stress-timed" and "Syllable-timed" languages (Germanic languages vs. Romance languages), for example [8], [19], [20], and [21]. Moreover, the metrics have been applied in L2 prosody [22, 23], pathological speech [24], and musicology [25, 26] as well.

However, whether these metrics are robust measures of speech rhythm is strongly debated [45, 46]. Based on different elicitation methods and materials, [48] analyzed the speech rhythm of six languages and concluded that the metrics scores were easily influenced by elicitation methods and materials, and therefore, unsafe to classify languages. Also, the metrics do not have much success in distinguishing intuitively very different L2 speech from the L1 in terms of rhythm [23]. This study aims to partially replicate [23] with different speakers to further examine the robustness the metrics on L2 speech.

2. Method

2.1. Informants

Five native speakers of American English and five native speakers of Mandarin Chinese participated in the study. The native English speech data were originally part of the pathology-free data set in [24], and were made accessible to the author after passing a web-based course "Protecting Human Research Participants" with a certificate issued by The

National Institute of Health Office of Extramural Research (USA). The Mandarin speakers (all Beijing natives) were third-year English majors in a Chinese university, and therefore were deemed post-intermediate or advance English learners. The average age of the speakers were 20 at the time of their participations, and the mean onset age of English learning is 13. They received a small remuneration upon the completion of the recording. Both groups read and recorded five English sentences; besides, the Chinese group also read and recorded five Mandarin sentences. Therefore, both between-subjects and within-subjects comparisons between L1 English, L2 English and L1 Mandarin can be made.

2.2. Materials

The English sentences were the ones used in [21, 22, 24] and the average length is 16.2 syllables per sentence. Mandarin sentences were created to reflect natural syllabic distributions in daily usage, i.e. less used syllables were avoided to frequent the sentences. Similar to the English sentences, distribution of stress and unstressed syllables was uncontrolled for [21]. Glides (/w/ and /j/) and liquids (/l/) were avoided because the boundary between an approximant and a vowel is hard to discern on the spectrogram [21]. The average length is 16.6 syllables à Mandarin sentence. The annex lists all the sentences.

2.3. Apparatus and procedures

The Chinese speakers were recorded individually in a quiet room. Before the recording started, they were given adequate time to familiarize the reading materials. They were required to read sentence by sentence at normal speed. In case of stuttering, they were asked to read the problematic sentence again until totally at ease with that particular sentence. In addition, they were encouraged to reduce the number of unnecessary pauses; however, they could pause at the end of a prosodic phrase, which is normal in daily speech. They were required to read Mandarin sentences first and English sentences next. All recordings were made by the Microtrack 24/96 solid state recorder with the Audio Technica 8531 headset microphone (Sampling rate = 48 kHz, bit-depth = 16). The sound files were later transferred to the computer hard disk for further analysis.

2.4. Segmentation and measurements

The author identified and labeled the vocalic and consonantal intervals by visual inspection of waveforms and wideband spectrograms displayed in Praat [27] with the assistance of audio signals. All speech data were segmented according to the segmentation protocol set forth in [21]. The durations of vocalic and consonantal intervals were measured using a Praat script. The metrics scores were calculated on the Excel spreadsheet, and statistical testing was done using R [28].

3. Data analysis and results

3.1. Descriptive statistics and data normality

Means and standard errors of the metrics scores across L1 English, L2 English and L1 Mandarin are presented in Table 1. Both Kolmogorov-Smirnov test and Shapiro-Wilk test were employed to assess data normality. Results of both tests all indicated that the data are normally distributed (all p's > 0.05, two-tailed, see Table 2 for test statistics), meeting the normality assumption of parametric statistics.

3.2. Inferential statistics

One-way ANOVAs were conducted on all the metrics scores, and the main effect of language was found on ΔC ($F(2, 12) = 46.03$, $p < 0.0001$, adjusted $R^2 = 0.8655$), ΔV ($F(2, 12) = 11.61$, $p < 0.005$, adjusted $R^2 = 0.6026$), %V ($F(2, 12) = 15.54$, $p < 0.0005$, adjusted $R^2 = 0.6751$), VarcoC ($F(2, 12) = 11.65$, $p < 0.005$, adjusted $R^2 = 0.6033$), VarcoV ($F(2, 12) = 15.32$, $p < 0.0005$, adjusted $R^2 = 0.6716$), rPVI-C ($F(2, 12) = 41.56$, $p < 0.0001$, adjusted $R^2 = 0.8528$), and nPVI-V ($F(2, 12) = 36.05$, $p < 0.0001$, adjusted $R^2 = 0.8335$). The effect sizes (R^2) are large according to [29]'s criterion.

Tukey HSD post hoc multiple comparisons (please also see Table 1 for reference) indicated that L1 Mandarin is significantly greater than L1 English on all the metrics except %V (significance levels range from $p < 0.05$ to $p < 0.0001$), and is significantly lower than L1 English on %V ($p < 0.0005$).

Similarly, L1 Mandarin is significantly greater than L2 English on all the metrics except %V (significance levels range from $p < 0.01$ to $p < 0.0001$). On %V, L1 Mandarin is

Table 1. *Means (std. errors) of the metrics. The wave line enclosed part are not statistically different (α = 0.05)*

	ΔV	ΔC	%V	VarcoV	VarcoC	rPVI-C	nPVI-V
L1 English	46.1004 (1.40705)	56.9879 (1.87970)	41.5554 (.83937)	59.3842 (4.17688)	48.9780 (2.16890)	67.5853 (.56329)	66.7095 (2.41580)
L2 English	52.6955 (3.12098)	59.0983 (1.91801)	44.6327 (.60266)	51.7630 (1.89642)	47.3238 (2.07840)	69.1565 (3.33706)	57.4366 (1.00518)
L1 Mandarin	36.8405 (2.16055)	34.5581 (2.20080)	51.5982 (2.01049)	35.6646 (2.77137)	35.2209 (2.34678)	41.7336 (2.37952)	39.5189 (3.00890)

Table 2. *Results of the Kolmogorov-Smirnov and Shapiro-Wilk tests (two-tailed).*

		ΔC	ΔV	%V	VarcoC	VarcoV	rPVI	nPVI
Kolmogorov-Smirnov Test	D	0.830	0.379	0.792	0.538	0.558	1.022	0.707
	p	0.497	0.999	0.557	0.935	0.914	0.247	0.700
Shapiro-Wilk Test	W	0.9702	0.9660	0.9433	0.9200	0.9469	0.9324	0.9461
	p	0.8616	0.7946	0.4254	0.1926	0.4765	0.2965	0.4657

significantly lower than L2 English ($p < 0.01$).

No significant differences were found on almost all the metrics except nPVI-V (all p's > 0.1) between L1 and L2 English. Nevertheless, L1 English is significantly greater than L2 English ($p < 0.05$) on nPVI-V.

To sum up, the rhythm metrics have fair success in distinguishing canonically "stress-timed" L1 English from "syllable-timed" L1 Mandarin. However, they were insensitive to the differences between L1 and L2 English, a result quite similar to [23].

4. Discussion

4.1. Metrics scores of L1 English and L1 Mandarin

As the results suggested, L1 English is significantly higher than L1 Mandarin on ΔV, VarcoV and nPVI-V. This conforms to the fact that English have higher degrees of vowel reductions in unstressed syllables. Besides, English has phonemic distinctions between tense and lax vowels. The concomitant length differences also contribute to the higher variability in vocalic interval durations. Likewise, the proportion of vocalic duration out of the whole utterance duration is significantly lower in English than in Mandarin, also because of the occurrence of reduced vowels.

Moreover, L1 English is significantly higher than L1 Mandarin on all the consonantal metrics (ΔC, VarcoC and rPVI-C), showing a greater durational variability in consonantal intervals. Such higher variability reflects the more complicated syllable structure of English. An English syllable can be as light as V, or as heavy as CCCVCCCC; whereas even the most complicated Mandarin syllable has a simpler structure of CGVN or CGVG (N refers to the nasal; G refers to the glide, which is often acoustically realized as part of a diphthong) [31]. Such results as shown by the vocalic and consonantal metrics scores have successfully distinguished between English and Mandarin, two typical languages showing "stress-" and "syllable-timing" rhythm, agreeing with previous studies, such as [19] and [23].

4.2. Insensitivity of rhythm metrics on L2 English and critiques of the rhythm metrics

Although rhythm metrics have fair success categorizing typical languages, it fails to measure the difference between L1 and L2 English as the results of this study indicate. Intuitively, L1 English and L2 English by Mandarin speakers are rhythmically different, and [23] even claimed that Chinese L2 English was impressionistically "syllable-timed". However, L1 and L2 English are not significantly different on all the metrics except nPVI-V.

Such results suggest that the participants in the study have achieved a high level of English learning, and have already acquired such phonological aspects as vowel reductions, weak forms, and syllable structures. It would not be difficult to imagine that if our Mandarin-speaking informants were beginners of English learning, the metrics scores would have been closer to those of Mandarin, because the L1 would have still taken a substantial proportion in the interlanguage system (see [32]'s Ontogeny and Phylogeny Model of L2 phonological development that sketches the chronological trajectories of L1, L2 and language universals in the interlanguage).

Insofar as syllable structure is concerned, inexperienced learners whose L1 has simpler syllable structure always epenthesize a vowel to break down a consonant cluster or delete one or more consonants to slim down a syllable onset or coda to fit the complex L2 syllable into a legitimate one of the L1 [32]. For example, [33] discovered that the epenthesis of the schwa was common among L1 Mandarin speakers' English production (e.g., /vɪg/ → [vɪ.gə]) to conform to the syllable structure of Mandarin. This way, longer consonantal intervals are truncated, resulting in lower durational variability of the consonantal intervals. Furthermore, [34] found that experienced and inexperienced L1 speakers of Mandarin and other languages differed in their production of lax/tense vowels in that experienced learners produced more accurate distinctions between pairs of vowels like /i:/ and /ɪ/. Since the segmental length difference is often a concomitant of lax/tense distinction, inexperienced learner's speech would manifest less variability in vocalic intervals, resulting in vocalic metrics scores more similar to those of Mandarin.

That experienced L2 English learners have acquired the syllable structure and segmental length difference can easily hoax the metrics that rely solely on interval duration variability. Therefore, L2 English is classified as similar to the L1 variety, although it sounds rhythmically different from L1 English. Hence, the metrics are not sensitive enough to capture such suprasegmental characteristics of L2 English at all stages of interlanguage development, at least for Mandarin speakers as shown in this and [23]'s studies.

At the methodological level, differences in interval duration variability are not the whole story of speech rhythm, thus using the metrics as the litmus test of speech rhythm overlooks many aspects in the speech signal. Rhythm (and not just speech rhythm) is characterized by the occurrence of prominent elements at regular or semi-regular intervals. In human speech, potential cues to prominence include f0, intensity, spectral quality and duration, and languages may be different in the selection of the cues. For instance, [35] discovered that f0, intensity and duration all play a role in cueing prominence as perceived by native English speakers; however, only f0 is functional in cueing prominence in Mandarin. In a preliminary attempt to examine the difference of other cues between L1 English, L1 Mandarin and L2 English, the author analyzed the same speech data in the present study in terms of syllabic intensity (measured as dB_{SPL}) variability, and found that the difference between L1 Mandarin and L2 English was due to chance alone; however, both L2 English, L1 Mandarin were significantly different from L1 English [36], suggesting that other prosodic aspects should be included in speech rhythm research.

4.3. New directions of speech rhythm research

Apart from interval durations and intensity variability, f0 is proved effective to signal prominence or has the effect of changing the perceived duration [37, 38, 39, 40, 41]. [42] discovered that native speakers of Swiss German and Swiss French/Metropolitan French differed in the weighting of pitch cues and durational cues in perceived rhythm. [43] incorporated the language-specific weighting values of pitch and duration into combined pitch-duration PVI, and found more similar scores than otherwise would be if calculated by traditional PVIs. This suggests that perceived rhythm may not be that divergent across-linguistically if the calculation is

scaled by language-specific weightings of different acoustic cues.

Moving away from the duration-based approach to speech rhythm, [44] adopted an amplitude-based approach that examines the amplitude modulation in the speech envelope, which is modeled as a nested hierarchy with tiers representing different prosodic units, such as feet and syllables, and the hierarchy captures different metrical patterns in nursery rhymes as different phase-locked patterns between foot amplitude modulations and syllable amplitude modulations, suggesting a methodological innovation in speech rhythm research.

4.4. New applications of rhythm metrics

Although rhythm metrics are strongly debated in speech rhythm research [45, 46], they are potentially useful in the forensic milieu, because the metrics scores manifest high individual idiosyncrasies [46, 47, 48]. With explicit emphasis on forensic applications, a series of research done at the Phonetics Laboratory of Zurich University proved that rhythm metrics are useful in speaker identification [49, 50, 51, 52].

5. Conclusion

The study investigated the robustness of rhythm metrics among L1 English, L2 English and L1 Mandarin. The results indicated that L1 English and L2 English were not significantly different on almost all the metrics, although they are impressionistically dissimilar. Such results conform to [23]'s findings. The results indicate that rhythm metrics are not adequate to quantify L2 suprasegmental characteristics and speech rhythm in general. For further research, a larger sample size including L2 learners of Mandarin who speak English as L1 is also desirable. Finally, new directions of speech rhythm research and new applications of rhythm metrics were briefly introduced.

6. Annex

6.1. English sentences

1) The supermarket chain shut down because of poor management.
2) Much more money must be donated to make this department succeed.
3) In this famous coffee shop they serve the best doughnuts in town.
4) The chairman decided to pave over the shopping center garden.
5) The standards committee met this afternoon in an open meeting.

6.2. Mandarin sentences

Standard Romanization [30], phonetic transcriptions, and English translations are shown:

1) *Dàjiě jīntiān zǎochén gēn māma qù zhèjiā chāoshì mǎi jiǎozi.* /ta tɕiə tɕin tʰian tsau ʈʂʰən kən mamə tɕʰy ʈʂɤ tsia ʈʂʰau ʂɻ mai tɕiao tsɥ/
   'My sister went to the supermarket with my mom this morning to buy some dumplings.'
2) *Tā hǎoxiǎng tīng dàjiā chàng nàbù diànshìjù de zhǔtí qǔ.* /tʰa xau ɕiaŋ tʰiŋ ta tɕia ʈʂʰaŋ na pu tian ʂɻ tsy tə ʈʂu tʰi tɕy/
   'He wants to listen to the theme song of that TV show sung by everybody.'
3) *Fùjìn zhèjiā kāfēitīng mài quánshì zuìhǎo de zhīshì dàngāo.* /fu tɕin ʈʂɤ tɕia kʰa fei tʰiŋ mai tɕʰyɛn ʂɻ tsui xau tə ʈʂɻ ʂɻ tan kau/
   'The coffee shop nearby serves the best cheesecakes in town.'
4) *Xiàozhǎng juédìng jiāng xuéxiào de zúqíuchǎng chóngxīn fānxīu.* /ɕiao ʈʂaŋ tsyɛ tiŋ tɕiaŋ ɕyɛ ɕiau tsu tɕʰiəu ʈʂʰaŋ ʈʂʰuŋ ɕin fan ɕiu/
   'The schoolmaster decided to refurbish the school pitch.'
5) *Tā gēn tóngxué shuōhǎo jīntiān zǎochén zài Kěndéjī ménkǒu jiànmiàn.* /tʰa kən tʰuŋ ɕyɛ ʂuo xau tɕin tʰian tsau ʈʂən tsai kʰən tɤ tɕi mən kʰou tɕiɛn mian/
   'She and her classmates decided to meet at the KFC franchise this morning.'

Please note that the non-IPA symbols [ɻ] and [ŋ] represent the rhotacized and non-rhotacized non-open central unrounded apical vowels in Mandarin [53].

7. Acknowledgements

I thank Dr. Satsuki Nakai in the School of Philosophy, Psychology and Language Sciences at The University of Edinburgh for her suggestions on the study and for having helped me with the acquisition of the native American English speech data. Thanks also go to Prof. Wu Hongyun in the School of Foreign Languages at Renmin University of China for providing me with working space during data collection, and for helping me find Chinese participants.

8. References

[1] List, G., "The boundaries of speech and song". Ethnomusicology, 1: 1-16, 1963.

[2] Merriam, A. P., The Anthropology of Music. Northwestern Univ. Press, 1964.

[3] Karmiloff, K. and Karmiloff-Smith, A., Pathways to Language: From Fetus to Adolescent. Harvard Univ. Press, 2001.

[4] Nazzi, T., Bertoncini, J. and Mehler, J., "Language discrimination by newborns: Towards an understanding of the role of rhythm", J. Exp. Psychol. Hum. Percept. Perform., 24: 756-766, 1998.

[5] Nazzi, T., Jusczyk, P. W. and Johnson, E. K., "Language discrimination by English-learning 5-month-olds: Effect of rhythm and familiarity", J. Mem. Lang., 43: 1-19, 2000.

[6] Bosch, L. and Sebastián-Gallés, N., "The role of prosody in infants'native language discrimination abilities: The case of two phonologically close language", in EUROSPEECH-1997, 231-234, 1997.

[7] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: A study based on speech resynthesis", J. acoust. Soc. Am., 105: 512-521, 1999.

[8] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73: 265-292, 1999.

[9] Ramus, F., Hauser, M. D., Miller, C., Morris, D. and Mehler, J. "Language discrimination by human newborns and by cotton-top Tamarin monkeys", Science, 288: 349-351, 2000.

[10] Classé, A., The Rhythm of English Prose. Blackwell, 1939.

[11] Lloyd James, A., Speech Signals in Telephony. Sir Isaac Pitman & Sons, 1940.

[12] Pike, K., The Intonation of American English. Univ. of Michigan Press, 1945.

[13] Abercrombie, D., Elements of General Phonetics. Edinburgh Univ. Press, 1967.

[14] Dauer, R., "Stress-timing and syllable-timing reanalyzed", J. Phonet., 11: 51-62, 1983.

[15] Bertrán, A. P., "Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages", Lang. Design, 2: 103-131, 1999.

[16] Nespor, I., "On the rhythm parameter in phonology", in I. Roca [Ed], Logical Issues in Language Acquisition, 157-195, Foris, 1990.

[17] Dauer, R. M., "Phonetic and phonological components of language rhythm", in ICPhS-11, Tallinn, Estonia, 447-450, 1987.

[18] Low, E. L., Grabe, E. and Nolan, F., "Quantitative characterization of speech rhythm: Syllable-timing in Singapore English", Lang. Speech, 43: 377-401, 2000.

[19] Grabe, E. and Low, E. L., "Durational variability in speech and rhythm class hypothesis", in N. Warner and C. Gussenhoven [Eds], Papers in Laboratory Phonology 7, 515-543, Mouton de Gruyter, 2002.

[20] Dellwo, V., "Rhythm and speech rate: A variation coefficient for deltaC", in P. Karnowski and I. Szigeti [Eds], Language and Language Processing, 231-241, Peter Lang, 2006.

[21] White, L. and Mattys, S. L., "Calibrating rhythm: First language and second language studies", J. Phonet., 35: 501-522, 2007.

[22] White, L. and Mattys, S. L., "Rhythm typology and variation in first and second languages", in P. Prieto, J. Mascaró and M.-J. Solé [Eds], Segmental and Prosodic Issues in Romance Phonology, 237-257. John Benjamins, 2007.

[23] Mok, P. and Dellwo, V., "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English", in Speech Prosody 2008, Campinas, Brazil, 423-426, 2008.

[24] Liss, J. M., White, L., Mattys, S., Lansford, K., Lotto, A. J., Spitzer, S. M. and Caviness, J. N., "Quantifying speech rhythm: Abnormalities in the dysarthrias", J. Speech Lang. Hear. R., 52: 1334-1352, 2009.

[25] Raju, M., Asu, E. L. and Ross, J., "Comparison of rhythm in musical scores and performances as measured with the pairwise variability index", Musicæ Scientiæ, 14: 51-71, 2010.

[26] Patel, A. D. and Daniele, J. R., "An empirical comparison of rhythm in language and music", Cognition, 87: B35-B45, 2003.

[27] Boersma, P. and Weenink, D., "Praat, a system for doing phonetics by computer", Glot Int., 5: 341-345, 2001.

[28] R Development Core Team. R: A language and environment for statistical computing. http://www.R-project.org/, 2011.

[29] Cohen, J., "A power primer", Psychol. Bulletin, 112: 155-159, 1992.

[30] ISO 7098: 1991, Romanization of Chinese, 1991.

[31] Duanmu, S., The Phonology of Standard Chinese, 2e. Oxford Univ. Press, 2007.

[32] Major, R. C., Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology. Erlbaum, 2001.

[33] Broselow, E., Chen, S.-I. and Wang, C., "The emergence of the unmarked in second language phonology". Studies in Sec. Lang. Acquis. 20: 261-280, 1998.

[34] Flege, J. E., Bohn, O.-S. and Jang, S., "Effects of experience on non-native speakers' production and perception of English vowels". J. Phonet. 25: 437-470, 1997.

[35] Wang, Q., "L2 Stress Perception: The reliance on different acoustic cues", in Speech Prosody 2008, Campinas, Brazil, 635-638, 2008.

[36] He, L., "Syllabic intensity variations as quantification of speech rhythm: Evidence from both L1 and L2". in Speech Prosody 2012, Shanghai, China, 466-469, 2012.

[37] Lehiste, I., "Influence of fundamental frequency pattern on the perception of duration", J. Phonet., 4: 113-117, 1976.

[38] Cumming, R. E., "Should rhythm metrics take account of fundamental frequency?", Cam. Occ. Pap. in Ling. 4: 1-16, 2008.

[39] Kohler, K. J. "The perception of prominence patterns". Phonetica 65: 257-269, 2008.

[40] Niebuhr, O., "Fundamental frequency-based rhythm effects on the perception of local syllable prominence". *Phonetica,* 66: 95-112, 2009.

[41] Cumming, R. E., "The effect of dynamic fundamental frequency on the perception of duration", J. Phonet. 39: 375-387, 2011.

[42] Cumming, R. E., "The language-specific interdependence of tonal and durational cues in perceived rhythmicality", Phonetica, 68: 1-25, 2011.

[43] Cumming, R. E., "Perceptually informed quantification of speech rhythm in pairwise variability indices", Phonetica, 68: 256-277, 2011.

[44] Leong, V., Prosodic Rhythm in the Speech Amplitude Envelope, Doctoral thesis, University of Cambridge, UK, 2012.

[45] Arvaniti, A., "Rhythm, Timing and the Timing of Rhythm", Phonetica, 66: 46-63, 2009.

[46] Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S., "How stable are acoustic metrics of contrastive speech rhythm?" J. acoust. Soc. Am. 127: 1559-1569, 2010.

[47] Yoon, T. J., "Capturing inter-speaker invariance using statistical measures of speech rhythm", in: Speech Prosody 2010, Chicago, USA, 2010.

[48] Arvaniti, A., "The usefulness of metrics in the quantification of speech rhythm", J. Phonet., 40: 351-373, 2012.

[49] Dellwo, V., Leemann, A. and Kolly, M-J., "Speaker idiosyncratic rhythmic features in the speech signal", in Interspeech 2012, Portland (OR), USA, 2012.

[50] Dellwo, V., Kolly, M.-J. and Leemann, A., "Speaker identification based on speech temporal information", Abstract presented at IAFPA 2012, Santander, Spain, 2012.

[51] Dellwo, V., Schmid, S., Leemann, A., Kolly, M.-J. and Mueller, M. "Speaker identification based on speech rhythm: the case of bilinguals". Abstract presented at PoRT2012, Glasgow, UK, 2012.

[52] Dellwo, V., Leemann, A. and Kolly, M-J., "Can speakers be identified auditorily based on suprasegmental temporal characteristics?" Abstract presented in Phonetik und Phonologie 9, Zürich, 2013.

[53] Pullum, G. and Ladusaw, W. A., Phonetic Symbol Guide. The Chicago Univ. Press, 1996.

Pitch range declination and reset in turn-taking organisation

Céline De Looze, Irena Yanushevskaya, John Kane and Ailbhe Ní Chasaide

Phonetics and Speech Laboratory
School of Linguistics, Speech and Communication Sciences, Trinity College Dublin, Ireland
[deloozec, yanushei, kanejo, anichsid]@tcd.ie

Abstract

This paper examines how pitch range declination and reset contribute to turn-taking organisation. This is part of a broader study of voice prosody, i.e., how pitch, voice quality and temporal features combine for various prosodic functions, both linguistic and paralinguistic. The present study first investigates the effect of the speech unit position in a turn on its pitch range. We also test the effect of the number of speech units in a turn as well as the turn duration on the turn-initial f_0 peak height at the beginning of the turn. Our results suggest a pitch range declination trend between the Initial and Median speech units of a turn but a violation of this declination for the Final units of the turn. They also demonstrate that the higher the number of speech units in a turn or the longer the turn, the higher the turn-initial f_0 peak height. We discuss our findings along the debate on Projection and Reaction theories and that of Hard vs. Soft pre-planning of speech production. We address how these findings may be useful to formulate a holistic model of prosody and to enhance human-machine interactions.

Index Terms: pitch range, declination, reset, pause, gap, turn-taking, reaction vs. projection theories, hard vs. soft pre-planning.

1. Introduction

Spoken interaction is a joint activity where all participants are involved in the co-construction of meaning and in the establishment and maintenance of social relationships. Turn-taking organisation is an instance of such coordination, for which participants agree on who speaks, when to talk, listen, hold and take turns [1].

Depending on the situational context, turn-taking (employed herein as speaker change) can be realised with a silent interval (or gap), no-gap-no-overlap (when the start of one speaker's turn perfectly coincides with the end of the other speaker's previous turn) or with a transition overlap (when the start of one speaker's turn overlaps with the other speaker's previous turn, excluding backchannels). In this organisation, turn-holding (employed herein as speaker hold) corresponds to several utterances of the same speaker within a turn, that are separated by a silent interval (or pause).

To manage turn-taking, speakers would produce, perceive and react to a set of signals (prosodic, pragmatic, syntactical, semantical, visual) (*Reaction Theory* e.g., [2, 3, 4, 5, 6]) or would anticipate or project the end of the turn from contextual and structural information (*Projection Theory*, e.g., [1, 7, 8]). Within the frame of Reaction Theory, it has been reported that, in many languages, a level pitch accent or a flat contour at the end of an utterance is indicative of a turn-holding while any other terminal contour (such as rises and falls) is indicative of turn-taking [4, 9, 10, 11, 12, 13].

Works in the study of prosody have often reported that, in read speech, fundamental frequency declines over the course of an utterance [14, 15, 16]. Declination is thought to be the by-product of some physiological processes (e.g. subglottal pressure [14, 17], activity of the laryngeal muscles [18], tracheal pull [15]) or 'controlled' by the speaker and may have some specific linguistic and paralinguistic functions. It may however be violated in certain instances of terminal rises associated with questions or hesitations [19].

Evidence of f_0 declination at supra-utterance levels (e.g. above the level of the Intonation Unit) has also been reported in many languages, with paragraph initial utterances of spoken texts having higher and wider pitch range than paragraph final utterances [20, 21, 22, 23, 24]. This pitch range declination over the paragraph (or f_0 supra-declination) may participate in signaling the organizational and hierarchical structure of the discourse (e.g. signaling topic changes). [24] reported, for Dutch, that the lower a text segment is embedded within the hierarchy of a text, the lower the pitch range. [23] observed that, in French, intonation units at the beginning of a paragraph have a wider pitch range than medial and final position intonation units, the difference being greater between the initial and medial units than between the medial and final units.

We hypothesize that, in interactional speech, f_0 declination can also be observed, both at the utterance and at the turn (supra-utterance) level, and that it plays an important role in turn-taking organisation. Speech units in conversational speech may be embedded within turn units, as speech units in read speech are embedded within paragraph units. As suggested by Couper-Kuhlen [25], describing data of conversational speech, a downward shift or decrease in f_0 across successive utterances by the same speaker may indicate that they belong to the same turn, while an upward shift or increase would signal turn changes. In particular, she suggests that : "*Beginning an intonation phrase relatively high in one's voice range allows room for subsequent intonation phrases to be positioned lower and thus affords the possibility of declination units, which can be used to structure a 'big package'. Because high onsets initiate pitch declination units, they can be thought of as projecting 'more to come' in this case, further intonation phrases within the declination unit. In this sense, they provide prospective prosodic cues to the 'big package' that is under way*" ([25]:43).

It has been further argued that the height of the f_0 peak at the beginning of an utterance is indicative of the utterance length, with longer utterances being marked by higher f_0 peaks [26, 27, 28]. Similarly, we hypothesize that, in interactional speech, the first unit of a speaker's turn may be marked by a pitch reset whose height depends on the number of speech units within the turn or on the turn duration. Note that the relation between initial f_0 peak height and utterance length is still controversial as other studies did not observe any [29, 30].

In this paper, we investigate (i) whether pitch range declination operates in interactional speech at the level of the turn and (ii) whether pitch reset at the beginning of a turn depends on the number of utterances within the turn or on the turn length.

We test two hypotheses:

- **Hypothesis H1**. Initial Inter-Pausal Units (IPUs) within a turn have a higher and wider pitch range than Median and Final IPUs. To test this, the effect of the IPU's position (Initial/ Median/ Final) in a turn on the IPU's pitch range is investigated.
- **Hypothesis H2**. The f_0 peak at the beginning of a turn is higher when the number of IPUs in the turn is larger or when the turn is longer. To test this, we investigate the effect of the number of IPUs in a turn and the turn duration on the turn's initial f_0 peak.

It should be noted that most prosodic research on turn-taking has tended to focus on the duration of silent intervals, and on the chunk of speech preceding these intervals. In examining f0 declination trends, we focus particularly on the initial portions of utterances, as well as the relationships among the utterances of a turn. This work complements parallel work on the final intonational contours in the same data [41] as well as ongoing research exploring voice quality and temporal features and their correlation with melodic characteristics.

2. Experiment

2.1. Data

The speech data was extracted from a corpus of task-oriented dyadic interactions in Irish English [31]. It consists of 6 gender-paired interactions (involving 6 female and 6 male speakers). The interactions are based on a shipwreck scenario game where participants are presented with 15 items and are given 10 minutes to rank them in order of usefulness to their survival. Recordings were carried out with participants in separate isolation booths using a professional Neumann microphone connected to an Apple Mac-based Digidesign Pro-Tools Mbox2 recording system. The audio signal was digitised at 96 kHz/24 Bit and recorded using Pro-Tools software as two separate audio streams. Audio was then downsampled to 16 kHz/8 Bit.

2.2. Annotation and measurement

2.2.1. Annotation

The data was annotated in terms of speech units and silences automatically, similarly to [32]. A binary voice activity detection (VAD) was carried out on both speaker channels for each dyadic interaction, using the VAD algorithm proposed in [33].

Pauses, gaps, no-gap-no-overlaps (NGNO) and transition overlaps (TOV) were determined from the binary VAD on both speaker channels. A schematic output of the annotation is shown in Fig. 1. A minimum duration for pauses was set to 100 ms (threshold set empirically to avoid speech events like plosives to be annotated as pauses), which means that every speech unit separated by a pause of less than 100 ms were chunked together into a single speech unit. Note that decisions on such threshold settings may significantly affect the resulting duration distributions [34]. This automatic procedure resulted in 176 gaps, 121 TOV and 498 pauses.

Herein, TOV were excluded from the analyses. Speech units separated by a pause are referred to as Inter-Pausal Units (IPU). A turn is defined as a speech unit composed of one or several IPUs. The terms 'turn-taking' and 'speaker change' are used interchangeably. No annotation of backchannels (BC) have been included, which means that BC have been automatically annotated as speaker changes.

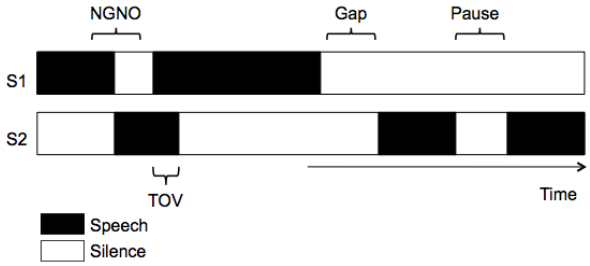


Figure 1: Schematic of a dyadic conversation between speaker 1 (S1) and speaker 2 (S2), illustrating occurrences of pauses, gaps, no-gap-no-overlaps (NGNO) and transition overlaps (TOV).

In the data, turns are composed of one to thirteen IPUs (mean=2.5, sd=2.12). Most common turns contain one IPU (40% of the data). To test Hypothesis H1, we excluded turns composed of one IPU as pitch range declination across several IPUs cannot be investigated in such turns. We also excluded turns composed of more than six IPUs as they are underrepresented in the data (less than 3%). From the resulting subset, IPUs were labelled Initial, Median and Final according to their position in a turn. In total, 292 IPU were labelled Initial, 243 Median and 311 Final. To test Hypothesis H2, the subset of selected data consisted of turns composed of one to six IPUs.

2.2.2. Measurement

Pitch range corresponds to the tonal space actually used in speech and is commonly described along two dimensions : its level (height) and span (range) [19]. Pitch range declination corresponds to the f_0 downward shift or decrease in pitch range across successive utterances within a turn and can be compared to the concepts "paragraph declination" [35], "supra declination behavior" [22] or "superordinate f_0 declination" [36] used in the literature for read speech.

Different acoustic measurements have been used to estimate pitch range level and span. Descriptive measurements such as f_0 mean and median based on f_0 values extracted every 0.01s on voiced segments or based on f_0 inflection points (or tonal targets, e.g. peaks and valleys) have been employed to account for utterance f_0 level. Measurements such as the difference between f_0 maximum and minimum values of utterances, f_0 95th and 5th percentiles, f_0 90th and 10th percentiles of utterances as well as f_0 standard deviation have been used to account for utterance f_0 span.

In this work, for each IPU and turn, standard descriptive measurements (f_0 maximum, minimum, median and the difference between f_0 maximum and minimum), representative, respectively, of an IPU's and a turn's initial f_0 peak, final valley, pitch range level and span, were computed (cf. Table 1). All these measurements are given on a logarithmic scale, the octave scale (i.e. log2(Hertz)), which is equivalent to the semitone scale. To account from cross-speaker differences, the data was normalized using z-scores. Pitch measurements were extracted using the phonetic software Praat [37]. To avoid possible pitch tracking errors at the pitch curve extrema, and enable their auto-

Features	Measurement	Abbreviation
initial f_0 peak	maximum f_0	$f_0$max
final f_0 valley	minimum f_0	$f_0$min
pitch level	median f_0	$f_0$med
pitch span	$f_0$max$-f_0$min	$f_0$maxmin

Table 1: Features computed for each turn (T) and Inter-Pausal Unit (IPU). Initial f_0 peaks are extracted at the beginning of each T and IPU, final f_0 valley at the end.

matic extraction, pitch floor and pitch ceiling, when creating a Pitch Object, were automatically adjusted to the speaker's pitch range (cf. [38] for more details).

2.3. Statistical analyses

A series of statistical analyses were carried out to test the effect of the independent variables (i.e. IPU position and number of IPUs in a turn) on the dependent variables (i.e. pitch range measurements). The analyses included a series of linear mixed models [39]. In our models, the position of the IPU (IPU-POS), the number of IPUs in a turn (N-IPU) and the duration of a turn (TDUR) were treated as fixed factors. Speaker was included as a random factor to take into account inter-speaker variability. The p-values were calculated using the method of Monte Carlo sampling by Markov chain (pMCMC = Monte Carlo Markov Chain [40]). In all our models, significance is set at a pMCMC $\alpha < 0.01$.

3. Results

3.1. H1 - Declination at the turn level

The effect of the speech unit position (IPU-POS) - 3 levels, Initial/ Median/ Final - on the IPU's pitch range is investigated. We assume that Initial IPUs have higher and wider pitch range than Median and Final IPUs. For the dependent variables $f_0$min, $f_0$max, $f_0$med and $f_0$maxmin, IPU-POS is tested as fixed factor and SPEAKER as random factor.

Results reveal a significant effect of IPU-POS on $f_0$min, $f_0$max, $f_0$med and $f_0$maxmin, being respectively higher and wider for Initial-IPUs than for Median-IPUs (respectively : t= -2.102, pMCMC$<$ 0.01; t= -1.921, pMCMC$<$ 0.01; t= -1.129, pMCMC$<$ 0.01;t= -1.316, pMCMC$<$ 0.01). This suggests a pitch range declination trend (i.e. lowering and narrowing of the pitch range) between the Initial and Median IPUs of the turn. The declination is however violated at the end of the turn, as it may be confounded with instances of rises associated with certain question forms or hesitations.

Two examples of pitch range declination trend observed in the data are given in Fig. 2 and 3. In Fig. 2, the turn is characterized by a declination trend over the three IPUs, the final IPU having a lower and narrower pitch range than the preceding IPUs. In Fig. 3, the turn is characterized by a declination trend over the first two IPUs which is violated on the final IPU, the latter having a higher pitch range than the preceding IPU.

3.2. H2 - Turn's initial f_0 peak as a function of the number of IPUs in a turn and turn duration

We investigate the effect of the number of IPUs (NIPU) in a turn and the turn duration (TDUR in log) on the turn's initial f_0 peak. We assume that the f_0 peak at the beginning of a turn is higher when the number of IPUs in the turn is larger or when the turn is longer. The models include NIPU and TDUR as fixed factor and SPEAKER as random factor.

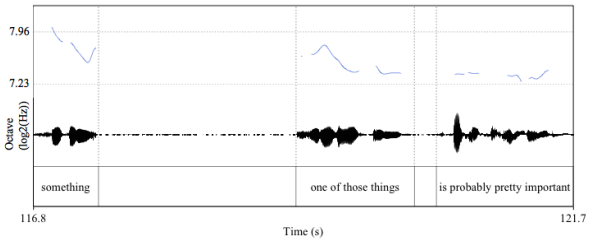


Figure 2: Example of a turn composed of 3 IPUs, uttered by a female speaker.

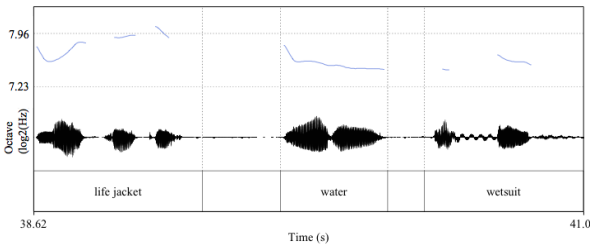


Figure 3: Example of a turn composed of 3 IPUs, uttered by a female speaker.

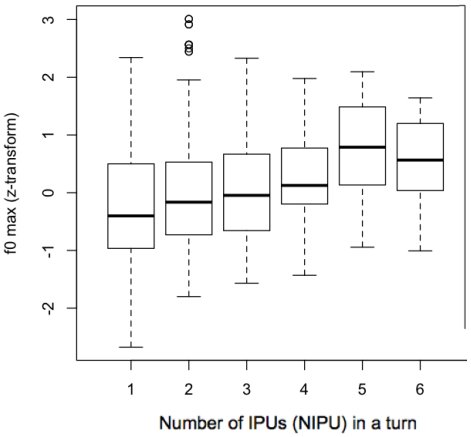


Figure 4: Turns' initial peaks (or $f_0$max) according to the number of IPUs in a turn.

Results reveal a significant effect of NIPU (pMCMC$<$ 0.01) and TDUR on $f_0$max (t=12.208, pMCMC$<$ 0.01). The higher the number of IPUs in a turn and the longer the turn, the higher the initial f_0 peak of the turn. Figure 4 indicates how the height of the f_0 peak increases with the number of IPUs in a turn. These findings suggest that the initial f_0 peak of a turn may be a salient cue in projecting the end of a turn. They corroborate earlier studies that observed a strong relation between an utterance's f_0 peak height and the utterance duration.

4. Discussion and conclusions

In this paper, we have investigated how pitch range declination and reset contribute to turn-taking organisation. We have

first tested the hypothesis H1 that Inter-Pausal Units in dialogue speech are embedded into turns as utterances in read speech are embedded into paragraphs, in such a way that Initial IPUs are higher and wider in pitch range than Median and Final IPUs. Our results suggest a declination trend (lowering and narrowing of the pitch range) between the Initial and Median IPUs of the turn, but not over the final IPU.

In a parallel study, using the same data, we have shown that 49% of units preceding a change of turn are declaratives, 35% questions and 10% backchannels and that these communicative types are associated with a falling tune (in 67% of the cases) or a low-rise tune (in 29% of the cases) [41]. The proportion of rises is the lowest in Declaratives (10%), it is higher in Incomplete Declaratives and WH questions (17%), and is the highest in Backchannels (24%). In Incomplete Questions, the pitch is predominantly rising (only 12% of samples have falling pitch). In Yes/No Questions, both falling and rising pitch pattern is used, with a slight preference for rises (about 54% in total).

Taking into account the effect of the utterance communicative type (e.g. declarative, incomplete, Wh-question, Yes/No question, hesitation, backchannel) on pitch modifications will allow us to better understand its singular role in turn-taking organisation. Pitch variations convey multiple functions in speech, e.g. signaling questions vs. statement, focus, topic changes and turn-taking, as well as in the signaling of intentions, attitudes and affect. The many linguistic functions of f0 changes were not controlled in the present experiment and may explain the variability encountered in the data. Future work will investigate the role of pitch variation at different levels, and will further attempt to link these to other prosodic variables, voice quality and temporal structure.

[23] observed that, in read speech, the difference in pitch range between the Initial and Medial units of a paragraph is greater than between the medial and final units. The present data show similar results. It would be interesting to investigate whether this pitch range 'break' between the initial and median units is indicative of a turn length, therefore may be used to signal turn change.

We have then tested the Hypothesis H2 that the f_0 peak at the beginning of a turn is higher when the number of IPUs in the turn is larger or when the turn is longer. Our results show that the higher the number of speech units in a turn and the longer the turn, the higher the initial f_0 peak height. This corroborates earlier findings on the relation between the initial f_0 peak height and the duration of an utterance [26, 27, 28].

These findings generally raise the debate of Hard vs. Soft pre-planning of speech production. On the one hand, it is proposed that speakers would be able to plan f_0 contours at a phrase level by adjusting the f_0 height at the beginning of the utterance to the utterance length. A higher initial f_0 may suggest a look-ahead or preplanning mechanism, by which utterance initial f_0 values are raised proportionate to utterance length [29]. On the other hand, it is suggested that speakers may proceed at a more local level, accent by accent. A lower f_0 at the end of the utterance may mean that adjustment is made on-the-fly. Our preliminary results suggest that speakers may plan their turn, adjusting its f_0 initial peak according to the turn length.

Overall, our findings suggest that pitch at the beginning of a turn and the break between the Initial and Median IPUs of a turn may contribute to turn-taking organisation. This means that not only syntactic and pragmatic information but prosody as well, appears to be used in projecting a speaker change.

We believe that both Reaction and Prediction theories can account for the underlying functioning of turn-taking organisation. As explained in [42], "*Redundancy is a well-studied and recurring principle of human language in use on virtually every level, and it is likely that a phenomenon as important as the taking of turns is orchestrated by a number of redundant control methods*" ([42]:566). In this view, we propose that speakers may anticipate the end of a turn based on the f_0 peak height at the beginning of the turn (as well as other signals) and may react to the lately uttered signals, adapting on-the-fly, by readjusting predictions if needed.

These findings could be directly applied to the modeling of human-machine interactions. A lot of work has been lately dedicated in improving the flow of conversation between a human and a computer or virtual agent. Standard methods have used a fixed duration threshold for the computer to begin speaking after the human interlocutor stops [43]. This strategy however does not really mirror what is usually done by humans. They, indeed, rather than wait for a silence to come, rely on syntactic, prosodic, pragmatic as well as visual cues to take the turn. Some studies have therefore investigated the use of these cues (prosodic and syntactic mainly) just before a silence to predict a speaker's hold or change [44].

In a parallel study [41], using the same data, we have shown that the combined discriminative power of functional and intonation labels (derived from speech-chunks immediately preceding pause and gap intervals) allows for differentiating turn-taking from turn-holding (mean classification error of 15%). In the present study, our results suggest that prosodic information at the beginning of a turn may also be a relevant cue to manage the conversation flow. The height of the initial f_0 peak of a turn could be used by a system to predict the end of the turn and the signals at the end of the turn (such as a final rise or fall vs. a flat tone) may be used to readjust prediction. This will be particularly addressed in our future classification experiments.

5. Acknowledgments

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET). We would like to thank Dr. Brian Vaughan (Dublin Institute of Technology) for providing the DIT Emotional Speech Corpus.

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simple systematic for the organization of turn-taking in conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[3] V. Yngve, "On getting a word in edgewise," in *Chicago Linguistics Society, 6th Meeting*, 1970, pp. 567–578.

[4] S. Duncan, "Some signals and rules for taking speaking turns in conversations.," *Journal of Personality and Social Psychology*, vol. 23, no. 2, pp. 283–292, 1972.

[5] A. Cutler and M Pearson, "On the analysis of prosodic turn-taking cues," *Intonation in Discourse*, pp. 139–156, 1986.

[6] C. Ford, B. Fox, and S. Thompson, "Practices in the construction of turns: The itc revisited," *Pragmatics*, vol. 6:3, pp. 427–454, 1996.

[7] E. Schegloff, "Discourse as an interactional achievement: Some use of 'uh-huh' and other things that come between sentences," *Georgetown University Round Table on Languages and Linguistics, Analyzing discourse: Text and talk*, pp. 71–93, 1982.

[8] J. De Ruiter, H. Mitterer, and N. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, pp. 515–535, 2006.

[9] J. Local and J. Kelly, "Projection and silences: Notes on phonetic and conversational structure," *Human Studies*, vol. 9, no. 2, pp. 185–204, 1986.

[10] C. Ford and S. Thompson, "Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns," *Studies in Interactional Sociolinguistics*, vol. 13, pp. 134–184, 1996.

[11] M. Selting, "On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation," *Pragmatics*, vol. 6, pp. 371–388, 1996.

[12] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and Speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[13] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics*, vol. 31, no. 2, pp. 251–276, 2003.

[14] P. Lieberman, "Intonation, perception, and language," *MIT Research Monograph*, 1967.

[15] S. Maeda, *A characterization of American English intonation*, Massachusetts Institute of Technology, 1976.

[16] J. t'Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: An Experimental Phonetic Approach to Speech Melody*, Cambridge: Cambridge University Press, 1990.

[17] R. Collier, "Physiological correlates of intonation patterns," *The Journal of the Acoustical Society of America*, vol. 58, pp. 249, 1975.

[18] J. Ohala, "Respiratory activity in speech," in *Speech Production and Speech Modelling*, pp. 23–53. Springer, 1990.

[19] D. Ladd, *Intonational Phonology*, Cambridge University Press, 2008.

[20] I. Lehiste, "Perception of sentence and paragraph boundaries," *Frontiers of speech communication research*, pp. 191–201, 1979.

[21] G. Bruce, "Textual aspects of prosody in swedish," *Phonetica*, vol. 39, no. 4-5, pp. 274–287, 1982.

[22] A. Sluijter and J. Terken, "Beyond sentence prosody: Paragraph intonation in dutch," *Phonetica*, vol. 50, no. 3, pp. 180–188, 1993.

[23] P Nicolas and D. Hirst, "Symbolic coding of higher-level characteristics of fundamental frequency curves," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[24] H. Ouden, L. Noordman, and J. Terken, "Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports," *Speech Communication*, vol. 51, no. 2, pp. 116–129, 2009.

[25] E. Couper-Kuhlen, "Prosody and sequence organization in English conversation," *Sound Patterns in Interaction. Amsterdam: John Benjamins*, pp. 335–376, 2004.

[26] K. Snider, "Tone and utterance length in Chumburung: An instrumental study," *28th Colloquium on African Languages and Linguistics, Leiden*, 1998.

[27] E. Couper-Kuhlen, "Interactional prosody: High onsets in reason-for-the-call turns," *Language in Society*, vol. 30, no. 1, pp. 29–53, 2001.

[28] A. Rialland, "Anticipatory raising in downstep realization: Evidence for preplanning in tone production," *Cross-linguistics Studies of Tonal Phenomenon: Tonogenesis, Typology, and Related Topics*, pp. 301–322, 2001.

[29] B. Connell, "Tone, utterance length and f_0 scaling," in *International symposium on tonal aspects of languages: With emphasis on tone languages*, 2004.

[30] P. Prieto, M. D'Imperio, G. Elordieta, S. Frota, M. Vigário, et al., "Evidence for 'soft' preplanning in tonal production: Initial scaling in Romance," in *Proceedings of Speech Prosody*, 2006, pp. 803–806.

[31] B. Vaughan, *Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings*, Ph.D. thesis, Dublin Institute of Technology (DIT), 2011.

[32] M. Heldner, J. Edlund, and J. Hirschberg, "Pitch similarity in the vicinity of backchannels," in *Proceedings of Interspeech 2010*, 2010, pp. 1–4.

[33] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[34] M. Wlodarczak and P. Wagner, "Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps," *Proceedings of Interspeech, Lyon, France*, pp. 1434–1437, 2013.

[35] J. Garrido, J.and Llisterri, C. Mota, and A. Ríos, "Prosodic differences in reading style: isolated vs. contextualized sentences," in *Third European Conference on Speech Communication and Technology*, 1993.

[36] N. Gronnum Thorsen, "Intonation and text in standard danish," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1205–1216, 1985.

[37] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2006.

[38] C. De Looze, *Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais contemporain*, Ph.D. thesis, Université de Provence, 2010.

[39] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar, "Linear and nonlinear mixed effects models," *R package version*, vol. 3, pp. 57, 2007.

[40] R. Baayen, *Analyzing linguistic data*, vol. 505, Cambridge University Press Cambridge, UK, 2008.

[41] I. Yanushevskaya, J. Kane, C. De Looze, and A. Ní Chasaide, "The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps," *Proceedings of Speech Prosody 2014*, In press.

[42] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[43] A. Raux and M. Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2008, pp. 1–10.

[44] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis

Mónica Domínguez[1], *Mireia Farrús*[1], *Alicia Burga*[1], *Leo Wanner*[2,1]

[1]TALN Group, N-RAS Research Centre
Department of Information and Communication Technologies
Universitat Pompeu Fabra
[2]Catalan Institute for Research and Advanced Studies (ICREA)
{monica.dominguez|mireia.farrus|alicia.burga|leo.wanner}@upf.edu

Abstract

This paper deals with the adaptation of AuToBI annotation for speech synthesis purposes. AuToBI is a tool that automatically determines and classifies the standard ToBI labels for American English. AuToBI annotation is performed word-by-word. However, for speech synthesis applications that use various layers of linguistic annotation (syntax, semantic information and prosody structures) and, in particular, for the detection of the correlation between the information structure and prosody, a labeling of intonation patterns at the intonational phrase level is essential. We present a rule-based procedure for initial AuToBI annotation and its adaptation a phrase-based annotation, avoiding thus a post-processing stage of the extracted labels. To validate our proposal, the outcome of the procedure is compared with manual annotation and with patterns prognosticated by information structure–prosody correlation argued for by main stream theories.

Index Terms: prosody, annotation, ToBI, AuToBI, thematicity, theme, rheme, speech synthesis.

1. Introduction

Prosodic features, such as rhythm, intonation, and stress are instrumental for the naturalness of speech and play thus an important role in the context of the "semantics–syntax–intonation" language interface in all speech-oriented Natural Language Processing (NLP) applications, especially in speech synthesis. The importance of prosody in NLP led linguists and speech technologists establish annotation standards for labeling prosodic events. One of them is ToBI (Tone and Break Indices) [1], a widely used convention thanks to its easy adaptation as a markup language for open-source speech synthesizers such as Festival [2].

The decade following the introduction of the ToBI convention, speech technology experienced an increasing interest in automated prosody labeling, mainly to avoid the time-consuming procedure of manual annotation.[1] As a consequence, a rather exhaustive number of works focused their interest on the automatic detection and annotation of prosodic events in speech; see, among others, [4, 5, 6, 7, 8]. One of the most well-known of them is AuToBI [9] for automatically detecting and classifying ToBI labels for American English. However, AuToBI labels prosody word by word, while what is required for NLP applications is segmentation at the phrase level that is based on a simplified converging model. Word-by-word segmentation is far too detailed to facilitate the connection between the other layers of annotation of the abovementioned "semantics–syntax–intonation" interface, especially when dealing with information structure [10].

In this paper, we discuss a rule-based procedure for the adaptation of AuToBI's word-by-word output to the needs of expressive speech synthesis, with the goal to be able to automatically establish a link between the information structure of an utterance and its prosody structure. The procedure groups AuToBi's word labels into intonational phrases (IPs) and proposes a single intonation pattern for each IP on the grounds of a set of criteria based upon the more detailed word-by-word labeling. Note, however, that we do not aim to address the general problem of phonologic/acoustic recognition of intermediate intonational phrases ('level 3' in ToBI terminology) that still pose a challenge for the state of the art; we merely aim to fit the needs for our research on the "semantics–syntax–intonation" interface.

The corpus that we label in our experiments is also manually annotated with the basic categories of the information structure theme and rheme (referred to as *thematicity*), which allows us to establish the correspondence between the prosodic patterns and themacity structures. This correspondence is used to validate our proposal of automatic prosodic pattern annotation and to contrast our work with the classical work of Steedman [10], which states that theme tends to be associated to the patterns L+H* and LH% (a clear increasing Low-High pattern), while rheme tends to be associated to the patterns H*L and H*LL% (clearly decreasing High-Low)—although both theme and rheme may be associated with other patterns as well.

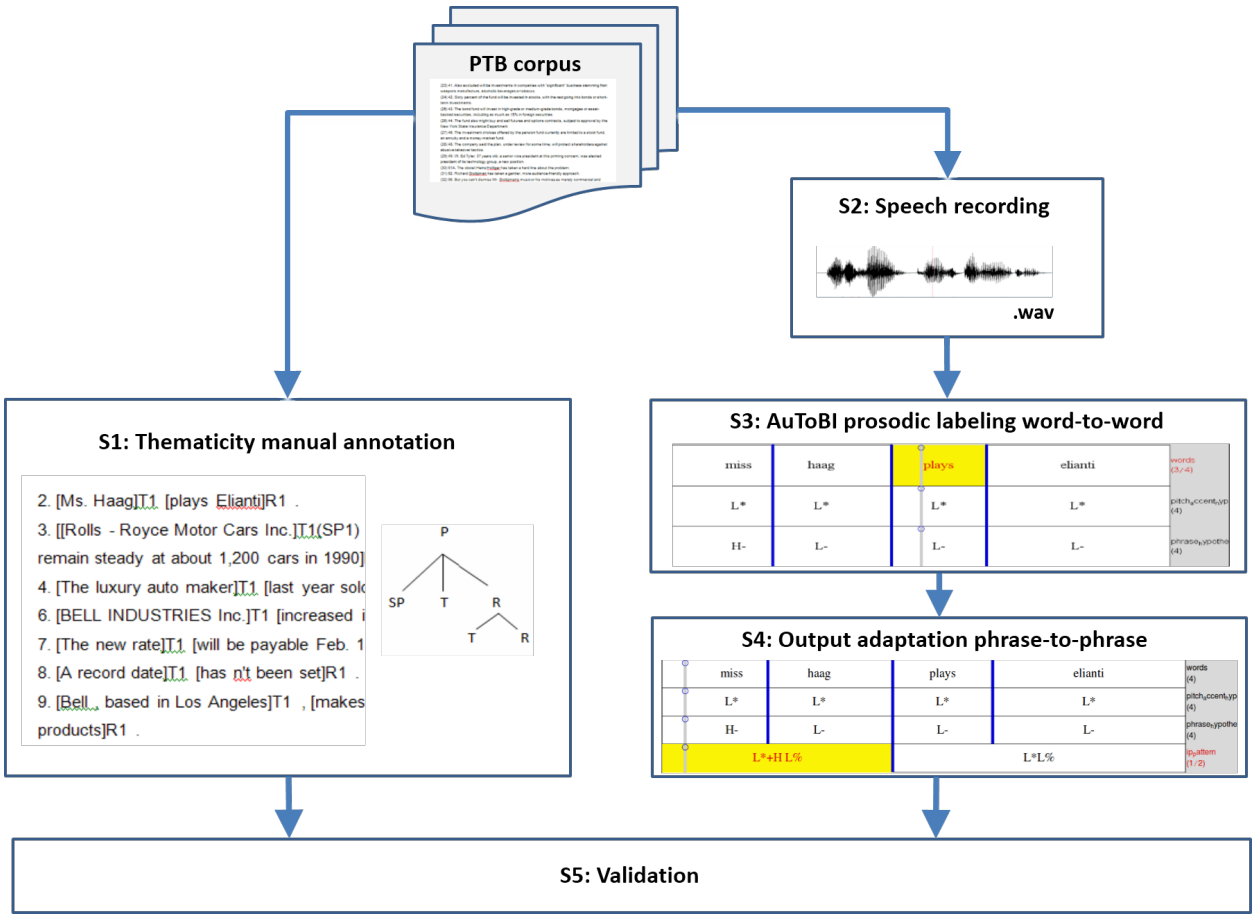
The paper is structured as follows. Section 2 describes the complete procedure of automatic annotation of phrase-based prosodic patterns, which involves both the AuToBI system and its adaptation to speech synthesis applications. Section 3 presents the annotation results and its validation through themacity structures, before Section 4, finally, summarizes the conclusions we draw from our preliminary work.

2. Annotating the Information–Prosody Interface

Our annotation procedure consists of five different stages, as shown in Figure 1: (1) thematicity annotation, (2) corpus recording, (3) AuToBI annotation, (4) output adaptation, and (5) validation of the results using manual reference annotations and the outcome of stage (1). In the first stage (S1), a reference corpus is annotated with the information structure (as pointed

[1]Syrdal et al. [3] estimated that experienced labelers could need between 100 and 200 times of the real time speech episode to annotate it.

Figure 1: *The stages of the proposed prosodic pattern annotation procedure and its validation.*



out above, we focus on the thematicity categories theme and rheme). In the second stage (S2), the reading of the corpus (or, as in our case, of a subset of the corpus) by a native speaker of American English is recorded. In the third stage (S3), the recorded speech is automatically labeled with the AuToBI tool. In the fourth (adaptation) stage (S4), the AuToBI word-by-word labels are transformed into IP pattern labels in accordance with our criteria. A final stage (S5) is used to assess the obtained patterns by comparing them with manual annotations and validate them with Steedman's theory [10] on the correlation between prosody and theme/rheme structures.

Next, stages S1 to S4 are described in more detail; the validation stage S5 is presented in a separate section that follows.

2.1. Thematicity annotation stage (S1)

The annotation of thematicity is assumed to be carried out manually over a plain text containing the consecutive sentences. In our experiments, this has been done in a series of blocks of about 40-50 sentences of a fragment of the Wallstreet Journal corpus extracted from the Penn Treebank [11], in accordance with the hierarchical thematicity structure of the Meaning-Text-Theory (MTT) [12].[2] In Figure 1, square brackets mark each communicative span (cf., e.g., '[. . .]T') and parentheses annotate embedded thematicity (cf., e.g., '[. . .]T(R)'). The annotators took into account the context (i.e., the previous sentence), but assumed an interpretation in which none of the elements is focalized or emphasized. The annotation was done by two groups of annotators (two in each group), who discussed in plenum their corresponding annotations to achieve a consensus and to refine the annotation guidelines.

For the validation of the IP pattern annotation procedure, the MTT-oriented annotation has been simplified to match Steedman's theme/rheme structures.

2.2. Speech recording stage (S2)

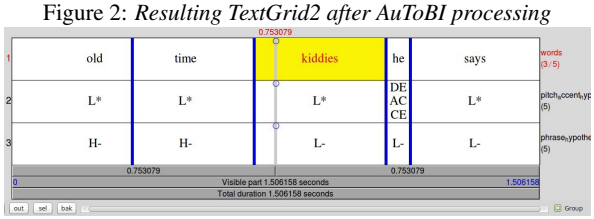
Within the speech recording stage, a subset of the corpus annotated in S1 with thematicity has been recorded. In our experiments, a non-expert native speaker of American English (not involved in the study) was instructed to read a selection of 109 sentences. The sentences contained varied information structure patterns, such that they were prosodically interesting for our study. The recording was done under professional conditions. The sentences were analyzed in order to create a reduction model from the AuToBI output to the IP level to be applied to the entire corpus.

[2]For details about the annotation criteria, see [13].

2.3. Automatic prosodic annotation stage (S3)

This stage consisted in segmenting audio files into words as required by AuToBI. This has been done to automatically process AuToBI's labeling task using Praat [14] and thus generating a TextGrid file for each audio file. Results were saved as TextGrid2 (see Figure 2), which has three interval tiers: the manually segmented word tier and two interval tiers generated automatically by AuToBI, one for the pitch accents and the second for the boundary tones.

Figure 2: *Resulting TextGrid2 after AuToBI processing*



2.4. AuToBI output adaptation stage (S4)

In spite of the fact that AuToBI meant a great step forward in the systematization of prosodic labeling, it has some major constraints for our descriptive approach, as has already been mentioned above.[3] Consequently, the information from AuToBI needs to be manipulated to meet our description requirements for intonational phrases within the information structure framework. For this purpose, we established a limited and manageable inventory of intonation patterns at the phrase level based upon the ToBI annotation convention [16]. We are labeling one main pitch accent (PA) and the boundary tone in each IP. Furthermore, while in the standard ToBI convention [17] four tiers of data are foreseen, namely a tone tier, an orthographic tier, a break tier and a miscellaneous tier, we are only making use of the tone tier, as reduction of detail is prioritized. The collection of patterns we currently use comprehends the items and their possible combinations summarized in Table 1.

Table 1: Patterns set by Pierrehumbert and Hirschberg [16].

pitch accents	L*, H*, L*+H, L+H*, H*+L, H+L*
boundary tones	L%, H%

This limited collection of twelve intonation patterns certainly does not cover all the possible natural tonal realizations of utterances, but it is expected to meet our initial requirements for a model to predict prosody events in speech synthesis applications. The integration of several layers of analysis is assumed to aid to solve the challenge of the prediction of prosody events in speech synthesis applications in that it caters for main pitch accents within the intonational phrase and relevant boundary tones at a clause level.

Our automatic adaptation stage can be envisaged as a loop of three steps over all sentences of the annotated corpus. The steps are: (1) Initial step, (2) Reduction step, and (3) Prerevision step.

[3] Some work has been done by Rosenberg [15] on the incorporation of intonational phrase boundaries into syntactic parsing for automatic summarization, but there is still a great deal of work to be done in this direction.

1. **Initial step.** This step consists in matching the output Textgrid from AuToBI to the sentence annotated in terms of theme/rheme. The result is a txt file (see Figure 3) that contains the following fields:

   - Id number of sentence
   - Chain of words
   - Communicative label
   - Chain of prosodic labels for those words

Figure 3: *Resulting txt matching AuToBI to thematicity labels*

```
Id.S.   Words                   AuToBI

0196    old time kiddies        R1      L*H-L*H-L*L-
0196    he says                 SP1     DEACCENTEDL-L*L-
0196    he                      T1(SP1) DEACCENTEDL-
0196    says                    R1(SP1) L*L-
```

2. **Reduction step.** The greatest part of the prosodic analysis is carried out during this step of the process. The strings of patterns from Step 2 are envisaged from the perspective of the intonational phrase in the pursuit of establishing not only the possible reduction models, but also the communicative and prosodic criteria to segment long utterances into smaller units. These units can help to draw a suitable intonational curve for speech synthesis purposes. As AuToBI does not predict bitonals, our reduction step seeks to predict possible bitonal PAs. The following automatic processing is performed on each pitch accent plus boundary tone (PABT) sequence:

   (a) Total deletion of deaccented items or word chains with a low BT (DL%). These intonation patterns match deaccented words which are disregarded in our IP characterization.

   (b) Substitution of deaccented items with a high BT (DH%) by a bitonal marker H+. High BTs in general may provide information on adjacent word stresses which are relevant in the detection of bitonals when they are followed by a main stress. A sequence of various H+ markers is reduced to a single H+ as it belongs to a sequence of deaccented words. Thus, the resulting single H+ matches a main stress and predicts a bitonal PA.

   (c) Word chains labeled as L*L% in a row can be disregarded for the IP contour definition. Three word-chains with such a label can be reduced to one L*L% IP label as only one word in such a chain will be more salient within the IP.

   (d) Initial L*H%L*L% has been reduced to L*+HL%. In this case, a high BT is turned into a bitonal.

   (e) 3-word combinations of L*H% and L*L% are turned to bitonals with either low or high BTs depending on the pattern chain. For instance, L*H% L*L% L*H% gives H+L*H%.

   The results from this label reduction process are saved into a txt file (see Figure 4) that contains the following fields:

   - Id number of sentence
   - Chain of words

- Communicative label
- Number of words
- Number of IPs
- Proposed ToBI label for each IP

Figure 4: *Resulting txt after reduction model processing*

Id.S.	Words	IS	N.W.	N.IP	PrePattern
0196	old time kiddies	R1	3	1	H+L*L%
0196	he says	SP1	2	1	L*L%
0196	he	T1(SP1)	1	0	0
0196	says	R1(SP1)	1	1	L*L%

3. **Pre-revision step.** After getting a proposed IP label, a Praat file needs to be created in order to revise all the material that has been automatically generated. Therefore, TextGrid3 file merges the existing tiers from TextGrid2 plus three more, namely:

   (a) clauses divided into intonational phrases and with their corresponding communicative labels,

   (b) same intonational phrases containing the proposed ToBI pattern, and

   (c) word divisions as in tier 1 for AuToBI input that will serve to place a pitch accent (PA) into the main stressed word within the IP and BT to be able to detect intermediate IP easily regardless pauses or silence dependency.

   See Figure 5 for illustration.

   Once the Textgrid3 file is generated, the manual process of revisor's validation of the proposed patterns takes place. The manual changes are saved as TextGrid4 (see Figure 6).

Figure 5: *Resulting TextGrid3 including processed tiers*

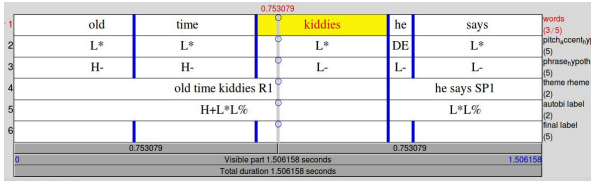
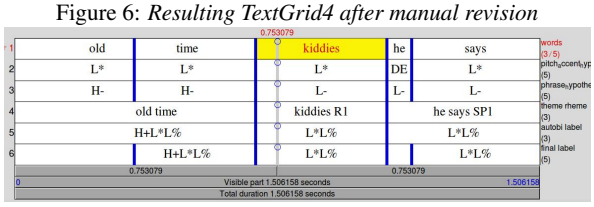


Figure 6: *Resulting TextGrid4 after manual revision*

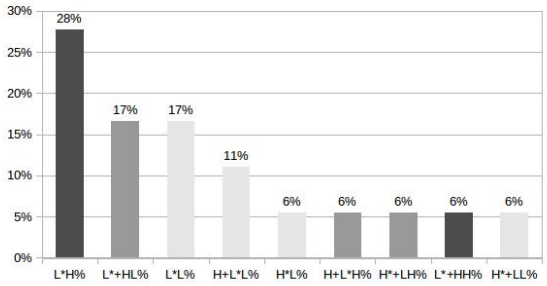


3. Validation stage (S5)

As mentioned above, the validation stage serves to assess the prosodic patterns obtained during the adaptation stage in order to evaluate the efficiency of our model. For this purpose, we first compared the results from our automatic reduction model at the intonational phrase level to a manual annotation. The comparison revealed that the model matches exactly the whole pattern in 58% of the total number of IPs. This includes number of IPs division and exact ToBI pattern assigned. There is a 18% of partially matched patterns (whose match corresponds in all cases to the BT). And the rest 24% of IPs does not match with the manual annotation.

Then, we compared the obtained prosodic patterns with those prognosticated by Steedman's [10] based on sentential theme/rheme structures. Figure 7 shows that themes tend to contain a rising intonation pattern as [10] claims, given that L*H%, L*+H H%, H*+L H% and H+L* H% (highlighted in dark gray) have a final rising intonation and L*+H L% contains a rising PA. These patterns add up to 63%, which proves that our model represents the general characterization made in theoretical approaches on this topic. Hence, the results can be regarded as reliable and we can conclude that apart from the obvious save in time in labeling effort, our reduction model is validated by existing theories on intonation applied to thematicity.

Figure 7: *Theme intonation patterns distribution (%)*



4. Conclusions

We presented a procedure for automatic prosodic pattern annotation that has been shown to be sufficiently reliable for experiments on the "semantics–syntax–intonation" language interface. In our validation, we have shown that our prosodic extraction is in concordance with Steedman's hypothesis for simple sentence structures. However, the great variety of intonation patterns that we found also proves that existing theories on thematicity characterization in prosodic terms, such as Steedman's [10], require a deeper insight from a qualitative perspective using real examples from different contexts, registers and speakers. For this reason, our procedure presented contributes, on the one hand, to the possibility of labeling large corpora with a substantial cut-off on manual revision efforts and minimizes, on the other hand, the risk of obtaining different labels as result of different annotators' subjective viewpoint, as long as all annotators are given a sistematic pattern and are asked to check whether there is an error with the initial output from AuToBI. Annotators can be trained to detect these errors and make more objective decisions when they spot an error than when they are labeling from scratch and therefore, have to make all decisions on their own.

5. Acknowledgements

Parts of this work have been funded by a grant from the European Commission under the contract number FP7-ICT-610411. The second author is partially funded by a grant from the Spanish Ministry of Economy and Competitivity in the framework of the Juan de la Cierva fellowship program (JCI-2012-12272).

6. References

[1] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: a standard for labelling English prosody". Proceedings of the ICSLP, vol. 2, 867-870, Sydney, Australia, 1992.

[2] Steedman, M., "Using APML to specify intonation". Magicster Project Deliverable 2.5. University of Edinburgh, 2005. Available at http://www.ltg.ed.ac.uk/magicster/deliverables/annex2.5/apml-howto.pdf

[3] Syrdal, A. K., Hirschberg, J., McGory, J. and Beckman, M., "Automatic ToBI prediction and alignment to speed manual labeling of prosody". Speech Communication, 33(1-2): 135–151, 2001.

[4] Noguchi, H., Kiriyama, K., Matsuda, H., Taniguchi, M., Den, Y. and Katagiri, Y., "Automatic labeling of Japanese prosody using j-toBI style description". Proceedings of the Eurospeech, 2259–2262, 1999.

[5] Lee, J.-S., Kim, B. and Lee, G. G., "Automatic corpus-based tone prediction using K-ToBI representation". Proceedings of the Conference on Empirical Methods in Natural Language Processing, 134–142, 2001.

[6] Ananthakrishnan, S. and Narayanan, S. S., "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model". Proceedings of the ICASSP, 269–272, Philadelphia, PA, 2005.

[7] Ananthakrishnan, S. and Narayanan, S. S., "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence". IEEE Transactions od Audio, Speech, and Language Processing, 16(1): 216–228, 2008.

[8] Rangarajan Sridhar, V. K., Bangalore, S. and Narayanan, S., "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework". IEEE Transactions on Audio, Speech, and Language Processing, 16(4): 797–811, 2008.

[9] Rosenberg, A., "AutoBI - a tool for automatic toBI annotation". Proceedings of Interspeech, 146–149, 2010.

[10] Steedman, M., "Information structure and the syntax-phonology interface", Linguistic Inquiry, 4(31):649–685, 2000.

[11] Marcus, M.P., Santorini, B. and Marcinkiewicz, M.A., "Building a Large Annotated Corpus of English: The Penn Treebank". Computational Linguistics, 19(2):313–330, 1993.

[12] Mel'čuk, I. A., "Communicative Organization in Natural Language: The semantic-communicative structure of sentences". Benjamins Academic Publishers, Amsterdam, 2001.

[13] Bohnet, B., Burga, A. and Wanner, L., "Towards the Annotation of Penn TreeBank with Information Structure". Proceedings of the Sixth International Joint Conference on Natural Language Processing, 1250–1256, Nagoya, Japan, 2013.

[14] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer" [Computer program]. Version 5.1.51, retrieved September 2013 from http://www.praat.org/.

[15] Maskey, S., Rosenberg, A. and Hirschberg, J., "Intonational Phrases for Speech Summarization". Interspeech, 2008.

[16] Pierrehumbert, J. and Hirschberg, J. , "The intonational structuring of discourse". Proceedings of the ACL, 136–144, New York, NY, 1986.

[17] Beckman, M. and Hirschberg, J., "The ToBI Annotation Conventions". The Ohio State University, OH, 1999. Available at http://www.ling.ohio-state.edu/ tobi/ame_tobi/annotation_conventions.html

Automatic Detection of Filled Pauses and Lengthenings in the Spontaneous Russian Speech

Vasilisa Verkhodanova[1], *Vladimir Shapranov*[2]

[1] SPIIRAS, 39, 14th line, St. Petersburg, Russia
[2] Betria Systems, Inc, 50, Building 11, Ligovskii Prospekt, St. Petersburg, Russia
verkhodanova@iias.spb.su, equidamoid@gmail.com

Abstract

During automatic speech processing a number of problems appear, and among them there are such as speech variation and different kinds of speech disfluences. In this article an algorithm for automatic detection of the most frequent of them (filled pauses and sound lengthenings) based on the analysis of their acoustical parameters is presented. The method of formant analysis was used to detect voiced hesitation phenomena and a method of band-filtering was used to detect unvoiced hesitation phenomena. For the experiments on filled pauses and lengthenings detection a specially collected corpus of spontaneous Russian map-task and appointment-task dialogs was used. The accuracy of voiced filled pauses and lengthening detection was 82%. And accuracy of detection of unvoiced fricative lengthening was 66%.

Index Terms: speech disfluencies, filled pauses, lengthenings, speech corpus, automatic speech processing, automatic speech recognition.

1. Introduction

A number of factors such as speech variation and different kinds of speech disfluences has a bad influence on automatic speech processing. Speech disfluencies are any of various breaks or irregularities that occur within the flow of otherwise fluent speech. These are filled pauses, sound lengthenings, self-repairs, etc. Another problem close to speech disfluencies are speech artifacts such as cough, laugh or sighs. The occurrence of these phenomena may be caused by exterior influence as well as by failures during speech act planning [1]. Hesitations are breaks in phonation that are often filled with certain sounds. Filled pauses are those hesitations that are filled with certain sounds, and the nature of sound lenghtenings is also hesitational. Such phenomena are semantic lacunas and their appearance means that speaker needs an additional time to formulate the next piece of utterance [2]. In oral communication filled pauses and lengthenings may play a valuable role such as helping a speaker to hold a conversational turn or expressing the speaker's thinking process of formulating the upcoming utterance fragment. Self-repairs appear when speakers want to change partly or entirely some piece of their utterances, and may be online: when speaker changes a piece of utterance immediately, or retrospective: speaker changes it post factum.

These phenomena are an obstacle for processing of spontaneous speech as well as its transcriptions, because speech recognition systems are usually trained on the structured data without speech disfluencies, what decreases speech recognition accuracy and leads to inaccurate transcriptions [3,4].

Nowadays there are two main types of methods of dealing with speech disfluencies: methods that process them by means of only acoustic parameters analysis, such as fundamental frequency transition and spectral envelope deformation [5,6] and methods that process them by means of combined language and acoustic modeling [7,8].

There are lots of works devoted to speech disfluencies modeling within the systems of automatic speech recognition [5,7,9]. Also there are approaches that deal with speech disfluencies at the stage of signal preprocessing [10], as well as speech disfluencies removal using speech transcriptions [9,11].

Thus, in [10] an algorithm, which defines and eliminates filled pauses and repetitions from the speech signal, is proposed. For detection of boundaries of filled pauses the following characteristics were applied: duration, pitch, spectral and formant characteristics. For extraction and further elimination of repetitions the proposed algorithm used duration and frequency of the repeated segments as well as the Euclidian distance between the logarithms of the Linear Predictive Coding (LPC) spectra of each pair of the voiced sections around a long pause. Also the fact that repetitions are usually accompanied by a pause was taken into account.

In [12] authors describe a method for automatic detection of filled pauses. They propose a method that detects filled pauses and word lengthening on the basis of two acoustical features: small F0 transition and small spectral envelope deformation, which are estimated by identifying the most predominant harmonic structure in the input. The method has been implemented and tested on a Japanese spontaneous speech corpus consisting of 100 utterances by five men and five women (10 utterances per subject). Each utterance contained at least one filled pause. Experimental results for a Japanese spoken dialogue corpus showed that the real-time filled-pause-detection system yielded a recall rate of 84.9% and a precision rate of 91.5%.

In [13] authors focus on the identification of disfluent sequences and their distinct structural regions, based on acoustic and prosodic features. For the experiments a speech corpus of university lectures in European Portuguese “Lectra” was used. The corpus contains records from seven 1-semester courses, where most of the classes are 60-90 minutes long, and consist of spontaneous speech mostly, and its current version contains about 32h of manual orthographic transcripts. Several machine learning methods have been applied, and the best results were achieved using Classification and Regression Trees (CART). The set of features which were most informative for cross-region identification encompasses word duration ratios, word confidence score, silent ratios, pitch, and energy slopes. The performance achieved for detecting words inside of disfluent

sequences was about 91% precision and 37% recall, when filled pauses and fragments were used as a feature. Presented results confirm that knowledge about filled pauses and fragments has a strong impact on the performance. Without it, the performance decayed to 66% precision and 20% recall.

There are number of publications aimed to rise speech disfluencies recognition quality by means of additional knowledge sources such as different language models. In [7] three types of speech disfluencies are considered: repetition, revisions (content replacement), restarts (or false starts). A part of Switchboard-I as well as its transcription (human transcriptions and ASR output) was taken for research. Normalized word and pause duration, pitch, jitter (undesirable phase and/or random frequency deviation of the transmitted signal), spectral tilt, and the ratio of the time, in which the vocal folds are open to the total length of the glottal cycle, were taken as the prosodic features. Also three types of language models were used: (1) hidden-event word-based language model that describes joint appearance of the key words and speech disfluencies in spontaneous speech; (2) hidden-event POS-based language model that uses statistics on part-of-speech (POS) to capture syntactically generalized patterns, such as the tendency to repeat prepositions; (3) repetition pattern language model for detection of repetitions.

For the application of disfluences detecting methods based on language modeling a large corpus of transcriptions is needed while for rule-based approaches there is no need for such corpus. Also rule-based approaches have an advantage of not relying on lexical information from a speech recognizer. For this research we decided to test the effectiveness of rule-based approach for detecting filled pauses and lengthenings in Russian spontaneous speech.

This paper is organized as follows: in the Section 2 the methodology for corpus recording and the collected corpus description are given. Section 3 is devoted to description of the method of filled pauses and lengthenings detection. In Section 4 the experimental results of hesitations and sound lengthening are presented.

2. Corpus of Russian Spontaneous Speech

Nowadays, for studying speech disfluencies corpora with Rich Transcription [11] are used. As example such corpus as Czech Broadcast Conversation MDE Transcripts [14] may be cited. This corpus consists of transcripts with metadata of the files in Czech Broadcast Conversation Speech Corpus [15], and its annotation contains such phenomena as background noises, filled pauses, laugh, smacks, etc [16].

For our purposes a corpus of spontaneous Russian speech was collected based on the task methodology: map-tasks and appointment-task. Thus, we have recorded speech that is informal and unrehearsed, and it is also the result of direct dialogue communication, what makes it spontaneous [17]. For example, in Edinburgh and Glasgow the HCRC corpus was collected, which consists only of map-task dialogs [18], and half of the another corpus, corpus of German speech Kiel, consists of appointment tasks [19].

Map task dialogs in the collected corpus represent a description of a route from start to finish, basing on the maps. Pair of participants had a map which had various landmarks drawn on it. One participant also had a route marked on their map. And the task was to describe the route to the other participant, who had to draw this route onto their own map. After fulfilling this task participants switched their roles and dialogue continued. For our investigation several pairs of maps of varied difficulty were created. As the criterion of difficulty the number of unmatched landmarks was used. An example of difficult maps is shown on the Figure 1. For dialogs based on appointment task, a pair of participants tried to find a common free time for: a) telephone talk (at least 15 minutes), b) meeting (1 hour) based on their individual schedules. Participants could not see maps or schedules of each other. Due to maps and schedules structure they had to ask questions, interrupt and discuss the route or possible free time. This resulted in speech disfluencies and artifacts appearance.

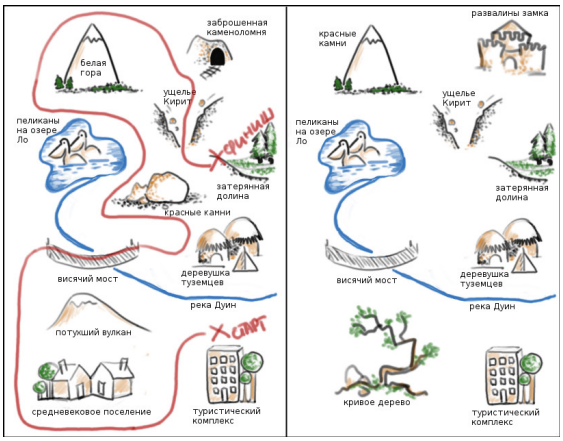


Figure 1: *An example of maps with the route (left) and without the route (right), used in map-task dialogs recording.*

The recorded corpus consists of 18 dialogs from 1.5 to 5 minutes. Recording was performed in the sound isolated room by means of two tablets PCs Samsung Galaxy Tab 2 with Smart Voice Recorder. Sample rate was 16kHz, bit rate - 256 Kbit/s. All the recordings were made in St. Petersburg in the end of 2012 - beginning of 2013. Participants were students: 6 women speakers and 6 men speakers from 17 to 23 years old with technical and humanitarian specialization.

Corpus was manually annotated in the Wave Assistant [18] on two levels: those disfluencies and artifacts that were characteristic for one speaker were marked on the first level, those that were characteristic for the other speaker - on the second level. During annotation 1042 phenomena such as filled pauses (for example pauses filled with [ ] and [ɐ] sounds), artifacts (as laugh, breath), self-repairs and false-starts as well as word-fillers were marked. The most frequent elements are shown on the Figure 2.

Figure 2: *A diagram of most frequent speech disfluencies and artifacts in the collected corpus, where ar.laugh - laugh, ar.brth – sighs and loud breath, h.a – hesitation [ɒ], h.e – hesitation [ə], i-i – lengthening of /i/, a-a – lengthening of /a/, "nu" and "vot": are common fillers in Russian.*

Sighs and loud breath, filled pauses [ ] and [m], self-rapairs and lengthening of sound /i/ appeared equally often in the speech of all 12 speakers. For speech of 11 speakers also lengthening of /a/ and filled pause [ɐ] were common. And almost everyone used such fillers as /vot/ ("there") and /nu/ ("well").

Due to the fact that certain disfluencies are communicatively significant and hardly can be distinguished from normal speech, on this stage of research we have confined ourselves to the most frequent elements of in speech disfluences – filled pauses and sound lengthenings.

3. Method of Filled Pauses and Lengthenings Detection

The basic idea of our method is to find acoustical features of filled pauses and sound lengthenings in speech signals by using spectrum analysis. Our method assumes that filled pauses and lengthenings contain a continuous voiced sound of an unvaried phoneme, due to this the neighboring instantaneous spectra are similar. For these phenomena such characteristics as unvaried value of pitch and duration of about 150-200ms are peculiar. This duration value is a reliable threshold for perception of speech pauses, because it is close to the value of mean syllable duration [21].

Taking into account only pitch change and duration it is possible to confuse sonorant sounds with filled pauses. For example, in such Russian word as "налево" /nalevo/ ("to the left") the pitch movement is almost horizontal as in filled pauses and lengthenings (Figure 3).

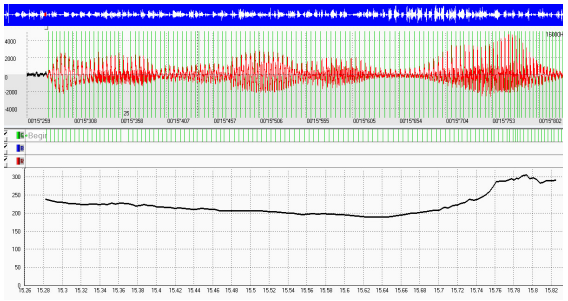


Figure 3: *A diagram of pitch movement for the word "налево" /nalevo/ ("to the left") with averaging interval of 50ms.*

As the measure of their similarity we have used a criterion of formant similarity between neighboring spectra. We also implemented the preliminary detection of lengthening of unvoiced fricatives.

In the following, we describe the main procedure of our method (Figure 4). First step was to calculate the Fourier transform to acquire a spectrogram with window length of 512 frames and step of 256 frames. This window length provides both reasonable spectral and temporal resolution.

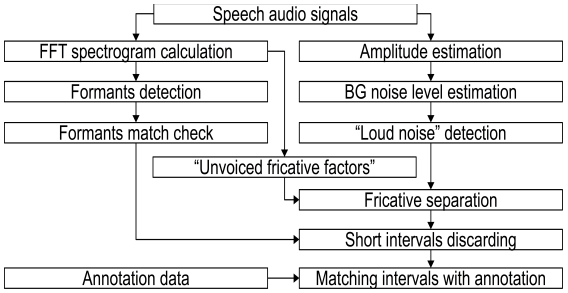


Figure 4: *Scheme of hesitations and breath detection method.*

The next stage was formants detection. First, spectrum was resampled to obtain exponential scale on frequency axis. This was done to acquire reasonable resolution in the middle- and high-frequency parts of the spectrum. Then we have searched for formants: the value of maximum and values of two surrounding samples were interpolated with quadratic curve and the position and value of the curve maximum was used as formant frequency and amplitude.

For formants match check we compared two neighboring spectra and estimated the coefficient of matching c for these two spectra (1). For every such pair of spectra the sum of formants' amplitudes multiplied by weight was estimated. Weight was calculated as a function of amplitude change and relative formant shift for every formant. Then this sum was divided by the sum of all amplitudes.

$$c = \frac{\sum_{match} A_n * F_{match}\left(A_n / A'_n, F_n / F'_n\right)}{\sum A_k} \quad (1)$$

where A_n, F_n - are amplitude and frequency in one neighboring spectrum and A'_n, F'_n - are amplitude and frequency in the other neighboring spectrum.

C reflects the sound invariableness in the current moment of time, if spectra are equal $c=1$, and if they are completely different, $c=0$. The diagram of *c* is shown on Figure 5 (the upper part). Those intervals where this function was above certain threshold for a long period of time are considered as filled pauses and lengthenings.

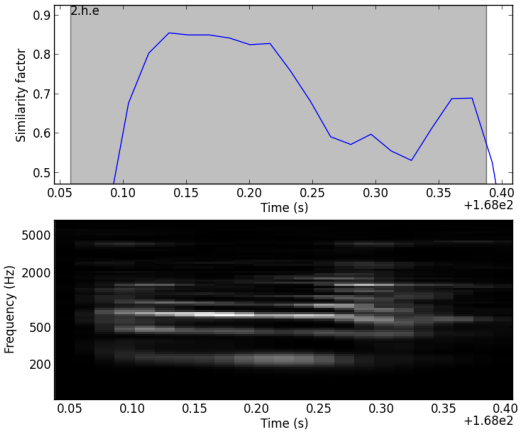


Figure 5: The diagram of s*imilarity function (above), with the gray background indicating mark in the annotation, and the resampled spectrogram (below) of the same signal part for filled pause /e/.*

To estimate an amplitude the signal was divided into overlapping frames, where the root mean square of samples in each frame is taken as the amplitude in correspondent moment of time.

To estimate a BG noise the signal was smoothed using the rectangular window with the length of 200ms, that is significantly less than characteristic length of silence intervals and greater than amplitude estimation window length. Minimum of this function was taken as background noise level.

The method based on the formant similarity described above doesn't perform well on the lengthenings of unvoiced fricatives. Due to the small amount of these elements (about only 1% of all annotated phenomena), almost all of them being sibilants lenghtenings, we relied on the fact that they are characterized by wide bands of certain frequencies ("fricative factors"). The situation of such bands for each unvoiced fricative sound is independent from the speaker. At this stage to detect unvoiced fricative lengthenings the following temporal series were computed: the ratio of the mean value of instantaneous spectrum samples in the band to the mean value of samples of the spectrum. Those intervals, where the series value exceeds a certain constant (more than 3), presumably contain the sound in question [22].

For fricatives separation the following actions were performed. For the found intervals values of "fricative factors" were examined by turns to detect among them those intervals that are corresponding to consonant lengthenings. The rest of the found elements were considered as breath.

Then the detected filled pause and lenghtening events were compared to the markup. For each event we looked for a mark that overlaped it, with the common part of these intervals being sufficiently large (the value of 0.4 was defined experimentally) (2):

$$L_{Ev\cup Mark} > 0.4\min\left(L_{Ev}, L_{Mark}\right) \qquad (2)$$

where the $L_{Ev\cup Mark}$ – is length of the common part, L_{Ev} – is the length of the event, and L_{Mark} – is the length of the mark. If the type of the mark matches the type of the event then the event was considered as match, otherwise it was considered as a false positive. All marks that were not matched during the events processing were treated as a false negative result.

4. Experimental Results

The filled pauses and sound lengthening algorithm based on the method described above was implemented and tested on a collected spontaneous Russian speech corpus. The training set consisted of 3 dialogs (4 speakers of different specialization) - two map-task and one appointment-task dialogue. The testing set was the other part of the corpus – 15 dialogs. The accuracy of voiced filled pauses and lengthenings detection was 82%. And accuracy of detection of unvoiced fricative lengthenings was 66%.

The main reasons for "misses" were the disorder of harmonic components in of hoarse voice and by laryngealized filled pauses and lengthening, the duration of which was not enough to overcome the threshold for correctly found elements. Another reason for misses was filled pauses consisting of two different sounds, such as /ae/. In such a case algorithm detected two lengthenings /a/ and /e/ ignoring the transition part, and both these lengthenings appeared to be too short to overcome the threshold. On the other hand, false alarms were mainly caused by lengthenings that were missing in the annotation and by noises and overlappings. For example the paper riffle sometimes is very similar to lengthening of a /s/ consonant and can be detected incorrectly.

5. Conclusions

This paper presents the method of filled pauses and sound lengthening detection by using the formant analysis. The experiments were based on the corpus of spontaneous Russian speech that was specially collected and manually annotated taking into account speech disfluencies and artifacts. The criterion of matching with the annotation marks was used as algorithm work estimation. The accuracy achieved for the voiced filled pauses and lengthenings detection was 82%. And the accuracy of the unvoiced fricative lengthening detection was 66%.

Further experiments will focus on more precise physical boundaries detection as well as on dealing with laryngealized sounds as well as on performing similar experiments with other Russian speech corpora within the other domain. Another stage of investigation will be devoted to context of filled pauses and lengthenings. This would help to detect more precisely their physical boundaries, that are of different nature, so there are such possible sounds as glottal stops in the beginning of filled pauses, transition parts between two sounds, etc. We also plan to apply our method to a Russian speech recognizer at a stage of signal preprocessing. Future work will also include an integration of the method with a speech dialogue system to make full use of the of filled pauses communicative functions.

6. References

[1] Podlesskaya, V.I., Kibrik, A.A., “Speech disfluencies and their reflection in discourse transcription”, VII International Conference on Cognitive Modelling in Linguistics Proc., 1: 194–204, 2004.

[2] Clark, H.H., Fox Tree, J.E., “Using uh and um in spontaneous speaking”, Cognition 84: 73–111, 2002.

[3] Verkhodanova, V.O., Karpov, A.A., “Speech disfluencies modeling in the automatic speech recognition systems”, The Bulletin of University of Tomsk, 363: 10–15, 2012 (in Rus.)

[4] Kipyatkova, I., Karpov, A., Verkhodanova, V., Zelezny, M., “Analysis of Long-distance Word Dependencies and Pronunciation Variability at Conversational Russian Speech Recognition”, Federated Conference on Computer Science and Information Systems Proc., 719–725, 2012.

[5] Masataka, G., Katunobu, I., Satoru, H., “A real-time filled pause detection system for spontaneous speech Recognition', 6th European Conference on Speech Communication and Technology Proc., 227–230, 1999.

[6] Veiga, A., Candeias, S., Lopes, C., Perdigao, F., “Characterization of hesitations using acoustic models”, 17th International Congress of Phonetic Sciences Proc., 2054–2057, 2011.

[7] Liu, Y., Shriberg, E., Stolcke, A., “Automatic Disfluency Identication in Conversational Speech Multiple Knowledge Sources”, 8th European Conference on Speech Communication and Technology Proc., 957–960, 2003.

[8] Liu, Y., Shriberg, E., Stolcke, A., et al., “Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies”, IEEE Transactions on Audio, Speech and Language Processing, 1(5): 1526–1540, 2006.

[9] Lease, M., Johnson, M., Charniak, E., “Recognizing disfluencies in conversational speech”, IEEE Transactions on Audio, Speech and Language Processing, 14(5): 1566–1573, 2006.

[10] Kaushik, M., Trinkle, M., Hashemi-Sakhtsari, A., “Automatic Detection and Removal of Disfluencies from Spontaneous Speech”,13th Australasian International Conference on Speech Science and Technology Proc., 98–101, 2010.

[11] Liu, Y., “Structural Event Detection for Rich Transcription of Speech”, PhD thesis, Purdue University and ICSI, Berkeley, 253 p., 2004.

[12] Masataka, G., Katunobu, I., Satoru, H, “A Real-time Filled Pause Detection System for Spontaneous Speech Recognition”, 6th European Conference on Speech Communication and Technology Proc., 227–230, 1999.

[13] Medeiros, R.B., Moniz, G.S., Batista, M.M., Trancoso, I., Nunes, L., “Disfluency Detection Based on Prosodic Features for University Lectures”, 14th Annual Conference of the International Speech Communication Association, 2629 – 2633, 2013.

[14] Corpus “Czech Broadcast Conversation MDE Transcripts”, LDC. Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T20, accessed 5 Oct 2013.

[15] Corpus “Czech Broadcast Conversation Speech”, LDC. Online: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009S02, accessed 5 Oct 2013.

[16] Kolar, J., Svec, J., Strassel, S., et al.,”Czech Spontaneous Speech Corpus with Structural Metadata”, 9th European Conference on Speech Communication and Technology Proc.,1165–1168, 2005.

[17] Zemskaya, E.A., “Russian spoken speech: linguistic analysis and the problems of learning”, Moscow, 1979. (in Rus.)

[18] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G.M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R., “The HCRC Map Task Corpus”, Language and Speech, 34: 351–366, 1991.

[19] Kohler, K.J., “Labelled data bank of spoken standard German: the Kiel corpus of read/spontaneous speech”,4th International Conference on Spoken Language Proc., 3: 1938–1941, 1996.

[20] Wave Assistant, the speech analyzer program by Speech Technology Center. Online: http://www.phonetics.pu.ru/wa/WA_S.EXE, accessed 5 Sep 2013.

[21] Krivnova, O.F., Chadrin, I.S., ”Pausing in the Natural and Synthesized Speech”, Conference on Theory and Practice of Speech Investigations Proc, 1999 (in Rus).

[22] Verkhodanova V., Shapranov V., “Automatic Detection of Speech Disfluencies in the Spontaneous Russian Speech”, in M. Zelezny et al. [Eds.], SPECOM 2013, LNAI 8113, 2013, pp 70-77, Springer International Publishing Switzerland, 2013.

Prosodic analysis of the speech of a child with cochlear implant

Aline Neves Pessoa-Almeida[1,2]; Alexsandro Meireles[2]; Sandra Madureira[1]; Zuleica Camargo[1]

[1]*Integrated Acoustic Analysis and Cognition Laboratory-LIAAC- LAEL-PUC-SP, Brazil*
[2]*Federal University of Espírito Santo, UFES –Brazil*

aline.pessoa@ufes.br; meirelesalex@gmail.com; madusali@pucsp.br; zcamargo@pucsp.br

Abstract

According to previous studies [1,6], acoustic and perceptual analysis can be considered useful clinical tools to investigate the speech characteristics of hearing impaired children (HIC). This study aimed at describing voice quality settings in speech samples from a HIC wearing cochlear implants. These samples were collected during speech therapy sessions in three moments: at the time the HIC was 5 years and 1 month old, and at 6 years and 1 month and at 7 years and 1 month. The perceptual analysis of the vocal quality was based on the Vocal Profile Analysis Scheme for Brazilian Portuguese (BP-VPAS - Camargo & Madureira, 2008). The recorded corpus was analyzed by means of the *ExpressionEvaluator script* (Barbosa, 2009) ran by *Praat* software v5.2.10. The measures, which were automatically extracted, comprised the fundamental frequency-f0, first f0 derivative, intensity, spectral slope and long-term mean spectrum. The correlations found between the acoustic and perceptual data are worth considering in rehabilitation programmes.

Index terms: Vocal quality; Auditory Perception; Acoustic Analysis; Cochlear Implant;

1. Introduction/Background:

This research focuses on the perception and production of voice quality settings and dynamic speech aspects [1-8] and it concerns the speech therapy and audiology settings since the analysis takes into account speech productions of a HIC and cochlear implant user (CI), collected during speech therapy sessions [6].

As indicated in a previous study [6], authors reinforce the demand for evaluating not only standardized speech tasks, but also semi-spontaneous speech data collected in the therapeutic environment.

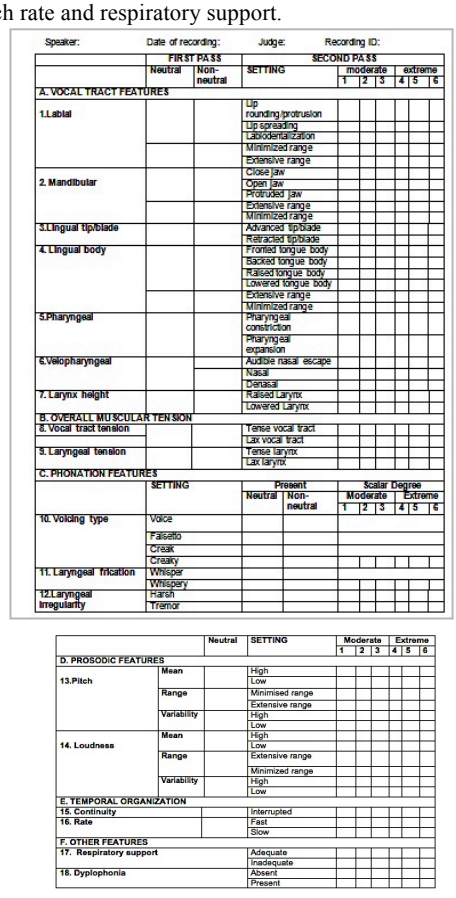
Perceptual and acoustic analysis of voice quality settings, prosodic features and of temporal organization aspects are useful clinical tools to investigate the speech characteristics of HIC [5,6,8].

The description of vocal quality settings (supralaryngeal, laryngeal and tension) and voice dynamics related aspects such as pitch, loudness, continuity and rate can provide useful means for clinic evaluation and intervention on HIC and CI individuals .

Speech segmental characteristics may undergo changes under the influence of voice quality settings. These interactions between segmental and prosodic levels [7,8] must be considered in relation to the physiological, acoustic, auditory and cognitive mechanisms involved in the production and perception of speech [8-10,16].

The analysis of the perceptual and acoustic correlates of vocal quality and voice dynamics makes it possible to identify changes in speech production which demonstrate the oral language development in children who use CI. [21-23]. The application of the Vocal Profile Analysis Scheme for Brazilian Portuguese (BP-VPAS) [9], based on VPAS 2007, (Figure 1)[10], enabled the perceptive description of two kinds of prosodic aspects: vocal quality settings and voice dynamics elements.

The vocal quality settings are taken as the result from the combined actions of the larynx and the supralaryngeal vocal tract [10-16]. Furthermore, such description aims at revealing the long-term tendencies that characterize vocal quality settings, which can be regarded as products of the respiratory, laryngeal/phonatory, supralaryngeal/articulatory systems, as well as muscular tension conditions. For the voice dynamics evaluation the BP-VPAS, the model provides the possibility of evaluating pitch and loudness parameters, the use of pauses, speech rate and respiratory support.



Speaker: Date of recording: Judge: Recording ID:

		FIRST PASS Neutral	FIRST PASS Non-neutral	SECOND PASS SETTING	SECOND PASS moderate 1	2	3	extreme 4	5	6
A. VOCAL TRACT FEATURES										
1.Labial				Lip rounding/protrusion						
				Lip spreading						
				Labiodentalization						
				Minimized range						
				Extensive range						
2. Mandibular				Close jaw						
				Open jaw						
				Protruded jaw						
				Extensive range						
				Minimized range						
3.Lingual tip/blade				Advanced tip/blade						
				Retracted tip/blade						
4. Lingual body				Fronted tongue body						
				Backed tongue body						
				Raised tongue body						
				Lowered tongue body						
				Extensive range						
				Minimized range						
5.Pharyngeal				Pharyngeal constriction						
				Pharyngeal expansion						
6.Velopharyngeal				Audible nasal escape						
				Nasal						
				Denasal						
7. Larynx height				Raised Larynx						
				Lowered Larynx						
B. OVERALL MUSCULAR TENSION										
8. Vocal tract tension				Tense vocal tract						
				Lax vocal tract						
9. Laryngeal tension				Tense larynx						
				Lax larynx						

		Present Neutral	Present Non-neutral	Scalar Degree Moderate 1	2	3	Extreme 4	5	6
C. PHONATION FEATURES									
10. Voicing type	Voice								
	Falsetto								
	Creak								
	Creaky								
11. Laryngeal frication	Whisper								
	Whispery								
12.Laryngeal irregularity	Harsh								
	Tremor								

		Neutral	SETTING	Moderate 1	2	3	Extreme 4	5	6
D. PROSODIC FEATURES									
13.Pitch	Mean		High						
			Low						
	Range		Minimised range						
			Extensive range						
	Variability		High						
			Low						
14. Loudness	Mean		High						
			Low						
	Range		Extensive range						
			Minimized range						
	Variability		High						
			Low						
E. TEMPORAL ORGANIZATION									
15. Continuity			Interrupted						
16. Rate			Fast						
			Slow						
F. OTHER FEATURES									
17. Respiratory support			Adequate						
			Inadequate						
18. Dyplophonia			Absent						
			Present						

Figure 1: Vocal Profile Analysis Scheme (VPAS) [13]

From the acoustic point of view, vocal quality and voice dynamics have been analyzed according to the following parameters: fundamental frequency (f0), first f0 derivate, intensity, spectral slope, and long-term average spectrum [2-17-20].

It is very difficult to analyze speech development in hearing impaired children due to methodological difficulties and the number of variables involved: degree of hearing loss; age at the beginning of amplification; involvement of the hearing impaired children's families especially at the beginning of the therapeutical care concerning the use of hearing aids and participation of the family in the hearing rehabilitation process; and changes in articulatory and phonatory patterns.

In the speech of children, the phonation system is oscillating and, due to the non-linear relationships between the elements, it can feature patterns of great variability. A model which allows the description of articulatory, phonatory and tense settings can be applied to evaluate changes in the adopted patterns that define the maturation of the mechanism.

The phonetic model [12] of voice quality settings description used to analyze the speech samples in this study does not make a distinction between normal and pathological speech. The focus is on the articulatory, laryngeal and tense settings which interact with the speech segments affecting their quality. In this way, the inherent quality of speech segments is modified. If, for example, an [i] is produced with a lip rounding setting of voice quality, its inherent characteristic of lip spreading will be modified. Since the vocal quality settings are described according to the principles of susceptibility to the speech segments, that is, an oral speech segment, for example, will be considered more susceptible to a nasalized voice quality setting than a nasal one, it follows that settings are defined and described in relation to key speech segments.

Considering the variability of speech patterns and the complex interactions between perception and production, a phonetic model [12] which provides the description of long term muscular adjustments related to the production of voice quality settings enables the analysis and comparison of interactions between prosody and segments.

The language acquisition process is very complex and descriptions of speech characteristics based on normal parameters are not helpful since speech characteristics are not stable. From the earliest babbling—especially in children with hearing impairment— babbling may show evidences of learning the relationships between motor gestures and acoustic characteristics.

This study aimed at describing voice quality settings in speech samples from a HIC who wears cochlear implants. These samples were collected during speech therapy sessions in three moments: at the time the HIC was 5 years and 1 month old and at 6 years and 1month and at 7 years and 1 month.

2. Methods

Speech samples comprised audio recordings from semi-spontaneous speech of a HI and user of CI child within the chronological age range of 5 years and 1 month and 7 years and 1 month (Table 1). The recording of the *corpus* took place in a therapeutic context, in a speech therapy room. The unit of analysis is long- term, recurring features throughout speech production. The analysis includes all long-term trends of speech production featuring a speaker in particular: the moments used speech samples analyzed are 5-15 seconds.

The instruments used to record the samples were a unidirectional *Le son* lapel microphone and a *Sony* MD digital recorder model MZ- R70. The edition, treatment, and sample analysis processes were carried out at the Acoustic Analysis and Cognition Integrated Laboratory (LIAAC) at PUC-SP. The recordings were digitalized at the sample frequency of 22050 Hz and 16 bits with the wav extension, using the Sound Forge software (version 7.0).

Table 1- *Subject characterization*

Subject	Hearing aids usage (chronological age: years; month)	IC –age at the time of surgery (chronological age: years; month)	Audio-recording sessions (chronological age/ auditory age: years; month)
Y.	0;7	2;7/ unilateral CI	Moment 1 (5;1/ 2;4), Moment 2 (6;1/ 3;4), Moment 3 (7;1/4;4)

The perceptual analysis was carried out with the use of the VPAS-PB [9,10] by two experienced judges.

The acoustic analysis was developed by using the *Expression Evaluator script* [17,19] running in the software *Praat*v5.2.10. The script generates f0 (median, interquartile semi-amplitude, 99,5% quartile and skewness), first f0 derivate (mean, standard deviation (SD) and skewness), as well as intensity (skewness), spectral slope (mean, SD and skewness) and LTAS (SD) measures [17,19].

The perceptual and acoustic results were statistically analyzed by means of XlStat software [16]. There were two elements under analysis: the perceptual judgments and acoustic measures results. Canonical correlation analysis between perceptual and acoustic data and the discriminant analysis of the acoustic correlates were performed taking into account the speech productions of the child at 5 years and 1 month, 6 years and 1 month and 7 years and 1 month.

It is important to highlight of methodological adequacy and consistency of the statistical analysis (canonical and discrimination analysis) adopted to consider the correlations between qualitative (perceptual evaluation of the settings) and quantitative (acoustic measures) in this study. This kind of analysis can be applied to semi-spontaneous and spontaneous speech excerpts and does not require the use of standardized speech samples

The issue of adopting relative measures instead of absolute measures allows correlation between perceptual evaluation and acoustic measures in the characterization of each speaker profile. This procedure does not require the labeling of vowel and consonantal segments, which may not be well delimited in certain children speech productions in the earlier stages of language development and in those productions considered altered for the age bracket. Even the speech of hearing children in the earlier stages of language development can not be described according to norm parameters.

This research was approved by the Ethics Committee at PUC-SP (#135/2009).

3. Results

The circular diagrams derived from the canonical correlation analysis between perceptual and acoustic data, considering moments 1, 2 and 3, are presented in Figures: 1, 2 and 3.

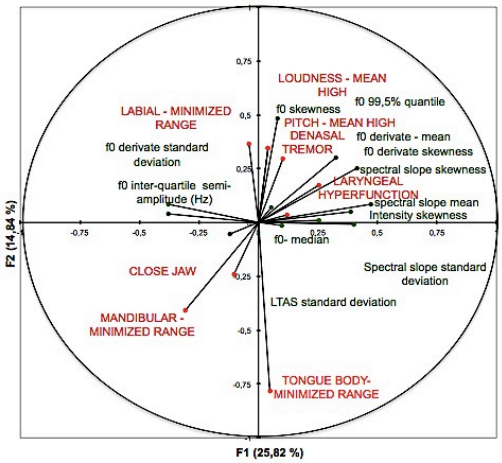


Figure 2: Circular diagrams from canonic correlation analysis: correlations between acoustic and perceptual data from a CI user in moment 1.

The correlations presented in Figure 2 comprised the most frequent vocal quality settings and voice dynamics parameters: f0-medianandusual *pitch* high (90,5%) and usual *loudness mean* high (90,5%,) f0-median associated with tremor (64,4%), f0-derivate skewness with labial minimized range (61%) and spectral slope - standard deviation associated with labial minimized range (50,5%).

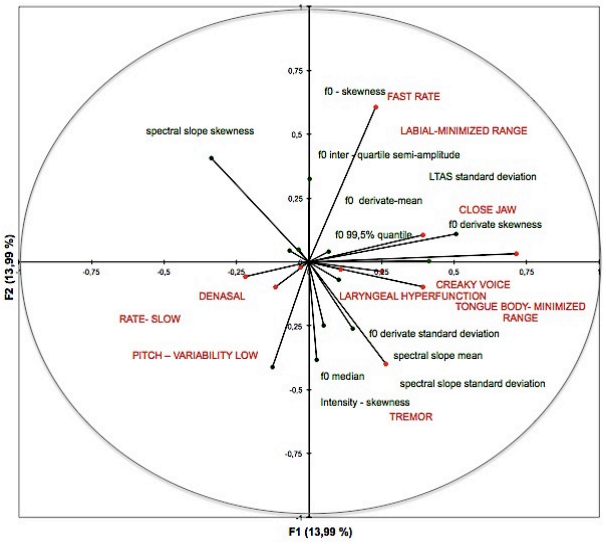


Figure 3: Circular diagrams from canonic correlation analysis: correlations between acoustic and perceptual data from a CI user in moment 2.

The correlations shown in Figure 3 concern the most frequent vocal quality settings and voice dynamics parameters: fast rate with f0-skewness (46,2%), creaky voice associated with f0-interquartile semi-amplitude (45,1%), slow rate with f0-interquartile semi-amplitude (41,5%) and tense larynx associated with intensity skewness(41,4%)

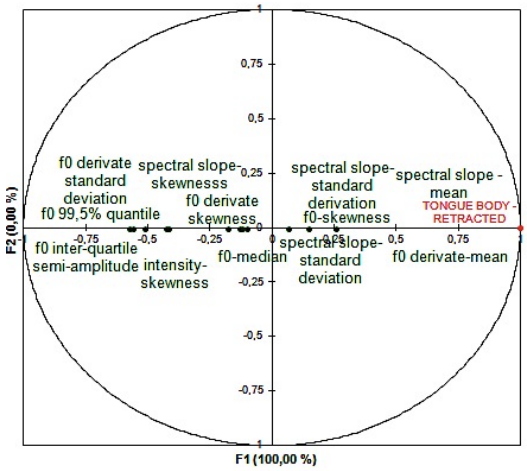


Figure 4: Circular diagrams from canonic correlation analysis: correlations between acoustic and perceptual data from a CI user in moment 3.

The correlations shown in Figure 4 revealed the most frequent vocal quality settings and voice dynamics parameters: f0-standard deviation derivative and minimized range of tongue body (93,9%) and harsh voice (93,9%) associated with falsetto and f0-median (91,3%).

4. Discussion and Conclusions

The results presented in terms of perceptual and acoustic data from a cochlear implant user during a period of speech therapy can be interpreted as being derived from mobilizations and adaptations of the articulators to achieve specific articulatory targets. Both language development issues and speech therapy strategies are factors which influence these speech maneuvers. These tendencies reveal the interaction between articulatory mechanisms and laryngeal (phonatory) events in language acquisition. In the articulatory domain, lingual, jaw and velopharyngeal settings were related to loudness, pitch and speech rate elements. In the phonatory domain, the tension (laryngeal hyperfunction) and laryngeal (harsh voice and whisper) settings groups were found to be very productive.

The association of laryngeal hyperfunction to high habitual pitch may be compatible with minimized range settings of jaw and tongue [22]. Such combinations are commonly described as yielding mechanisms for vocal tract and laryngeal hyperfunction, especially if conditions related to the developmental stages of the vocal apparatus are considered.

Taking into account the perceptual and acoustic data distribution in canonical correlation analysis (figures 3, 4 and 5), the interaction between some supralaryngeal mobilizations, especially tongue body and spectral slope and LTAS measures, is highlighted[10,14]. In the analyzed samples, these data can be interpreted as revealing some vocal loading in order to achieve some articulatory targets and, again, reinforce the complex interactions between supralaryngeal settings, voice dynamics elements and spectral measures in language acquisition for HI children [2,5-6, 15]

The findings also reinforce the complex interactions between pitch control and laryngeal (harsh voice and whisper) and muscular tension settings (laryngeal hyperfunction) [4, 6, 22-25], reflected in f0 acoustic measures. The values concerning habitual f0 at moment 1 (auditory age: 2y4m) were higher than those at moment 2 (auditory age: 3y4m), whereas at moment 3 (auditory age: 4y4m) these values increased 406,4 Hz. For this speaker, the influence of acoustic data of spectral slope and f0-skewness and intensity-skewness, in relation to laryngeal hyperfunction associated with aperiodicity are noteworthy. *Pitch* and *loudness* variability decreased. The phonetic literature refers to extreme and abrupt pitch variations not only for AASI users, but also for CI users [8,22,25,28]. Such findings indicate the influence of laryngeal hyperfunction and aperiodicity on pitch extension and variability and also reinforce the complex interactions of f0 control mechanisms in HIC, particularly in terms of the association of voice quality settings -vocal tract settings (minimized range – tongue, jaw and lip), laryngeal (tremor) and vocal tract hyperfunction. The findings concerning f0 variations in the listener population speech indicate that speech intelligibility and phonological discrimination can be affected if vocal aperiodicity characteristics are present. In relation to the vocal dynamics elements, the slow rate at moment 1 turned into a fast one at the moment 3, when the analysis detected a higher influence of the retracted tongue body setting of vocal quality. From the acoustic point of view, differentiation of evolutionary stages occurred mainly due to the distribution of f0 and spectral slope measures, which are supported by descriptions of perceptual basis. The measurements of median f0 increased over time (auditory age from 2y4m to 4y4m), as expected, because body and laryngeal development would be lowering. However, the retracted tongue body setting can change raised larynx setting and interfere with the characteristic vibration of the vocal folds. In addition, the resonance imbalances widely reported in the literature [1, 2, 6, 8, 15, 22, 23, 25, 26] may be related to vocal tract adjustments reported in this speaker (body of tongue, jaw and velopharyngeal mobility). The collected data likewise strengthen aspects of reduced movement of articulators. The collection revealed by the three

samples shows the evolution of meaningful maintenance vocal tract adjustments. The discriminant analysis of the findings using the VPAS-PB, in terms of moments of the recordings (moments 1, 2 and 3), revealed segregation of the variables at the rate of 74.16%. When individually analyzed, total segregation was higher for emission moment 3, with 100%, whereas moment 2 presented 75%, and moment 1, 66.67%. The most influential factors for this differentiation referred to the combinations of minimized range of jaw (grade 1), laryngeal hyperfunction (grade 1), decreased pitch variability (grade 1), and retracted tongue body (grade 1), represented by 80,18% combined with minimized range of tongue body (grade 2), jaw (grade 2) settings and the absence of laryngeal hyperfunction, at the rate of 19,82%. The discriminant analysis of the acoustic measures by *ExpressionEvaluator Script*, in terms of moments of the recordings (1, 2, and 3), revealed total segregation of the variables at 64.52%. Considering each moment in therapy, total segregation was higher for moment 2 emissions, by 80%, whereas moment 3 presented 72,22%, and 1 moment, 37.50%. The most influential factors for such differentiation referred to the combinations of f0-derivate SD, f0-mediana and interquartile (99,5%), represented by 88,21%, combined measures off0-semi-amplitude interquartile (11,79%). The f0 values-mean shows smoother variation at moments 1 and 2 than at the third moment, in which abrupt variations were found. In moment 3, f0-semi-amplitude interquartile arises - which reflects greater change compared to times 1 and 2. At moment 3, the f0 measure at 99.5% quartile showed higher values than at the other two moments. Similar characteristics in moments 2 and 3 were found, leading us to consider that a greater improvement took place between the first and second moments. The second and the third moments were characterized by more accurate productions of speech sounds. These aspects show the dynamic nature of speech and the complex interactions that take place between segmental and prosodic elements. At moment 3, a better delimitation of prosodic groups coincided with a period of gradual refinement of the ability in terms of articulatory productions [7]. It is worth pointing out that, besides the diagnosis and early intervention being important for the prognostics [3,4,21], specific rehabilitation procedures concerning the oral sensorial-motor system, voice and speech, seem to be crucial for a good verbal-oral language development and for the acoustic feedback [1,3,6]. Such information leads to the possibility to detail the articulatory maneuvers adopted by the children in developmental language process, and enhances the reliance of therapy, which indicates probable strategies in trying to attain the acoustic-articulatory targets in speech production. The findings reinforce some correlations between the acoustic and perceptual data, which are relevant to be considered in rehabilitation processes.

5. Acknowledgements

We acknowledge Plínio Barbosa from UNICAMP, for the revised version of the *ExpressionEvaluator Script*. This study was sponsored by a FAPESP grant (2009/10644-7).

6. References:

[1] Pessoa, A.N.; Pereira, L.K.; Camargo, Z.A.; Madureira, S.; Novaes, B.C.A.C. An analysis of voice quality and voice dynamics in the speech production of a cochlear implant user. In: 13th Meeting of the International Clinical Linguistics and Phonetics Association - ICPLA, Oslo-Norway, June 23-26, 2010, p-286

[2] Stuchi RF; Nascimento LT; Bevilacqua MC and Brito Neto RV. "Linguagem oral de crianças com cinco anos de uso do implante coclear". Pró-Fono 2007 Abr-Jun;19(2):167-76

[3] Boothroyd, A.Auditory development of the hearing child. Scandinavian Audiology, 26(Suppl. 46) , 9-16. 1997.

[4] Flexer, C. Facilitating Hearing and Listening in Young Children (2nd ed.). San Diego, CA: Singular Publishing Group.1999.

[5] Madureira S; Barzaghi L and Mendes B. "Voicing contrasts and the deaf: production and perception issues". In: Windsor F; Kelly ML; Hewlett N. (Org.). Investigation in Clinical Phonetics andLinguistcs. 1:417-28, 2002

[6] Pessoa, A.N; Novaes, BCAC; Madureira, S; Camargo, Z. Perceptual and acoustic correlates of a speech in a bilateral cochlear implant user. In: Abstract Book Speech Prosody 2012, 6th International Conference, Qiuwu Ma, Hongwei Ding and Daniel Hirst (eds.), Tongji University Press, Shanghai, China, May 22- 25,ISBN 978-7-5608-4869-3, v2, p51-54

[7] Albano E, Barbosa P, Gama-Rossi A, Madureira S, andSilva A. "A interface fonética-fonologia e a interação prosódia-segmentos". In: Estudos Linguísticos XXVII - Anais do XLV Seminário do Grupo de Estudos Linguísticos do Estado de São Paulo-GEL'97. Campinas, p.135-43, 1997.

[8] Cukier S; Camargo Z. "Abordagem da qualidade vocal em um falante com deficiência auditiva: aspectos acústicos relevantes do sinal de fala". Revista CEFAC, Jan-Mar. 7(1): 93-101, 2005.

[9] Camargo ZA; Madureira S. "Avaliação vocal sob a perspectiva fonética: investigação preliminar". São Paulo: Distúrbios da Comunicação, Abr. 20(1): 77-96, 2008.

[10] Camargo Z; Madureira S. "Dimensões perceptivas das alterações de qualidade vocal e suas correlações aos planos da acústica e da fisiologia". DELTA - PUCSP,25(2): 285-317, 2009.

[11] Laver J, Wirz SL, Mackenzie-Beck J and Hiller SM. "A perceptual protocol for the analysis of vocal profiles". Edinburgh University Department of Linguistics Work in Progress,14: 139-155, 1981.

[12] Laver J. "The phonetic description of voice quality". Cambridge: Cambridge University Press, 1980.

[13] Laver, J.; Mackenzie-Beck, J. "Vocal Profile Analysis scheme-VPAS". Edinburgh: QMUC, Speech Science Research Centre; 2007.

[14] Hammaberg B.; Gauffin J. "Perceptual and acoustics characteristics of quality differences in pathological voices as related to physiological aspects". In: Fujimura O, Hirano M. Vocal foldphysiology. San Diego: Singular. 283-303, 1995.

[15] Abberton E. "Voice Quality of deaf speakers". In: Kent RD, Ball MJ. Voice Quality Measurement. San Diego: Singular. 22: 449-59, 2000.

[16] Rusilo LC, Madureira S.; Camargo Z. "Evaluating Speech samples for the Voice Profile Analysis Scheme for Brazilian Portuguese (BP-VPAS)". In: Proceedings of the 4rd ISCA Workshop ExLing May 25-27; Paris, p.51, 2011.

[17] Barbosa PA. "Incursões em torno do ritmo da fala". Campinas: Pontes/FAPESP, 2006.

[18] Barbosa PA. "From Syntax to acoustic duration: a dynamical model of speech rhythm production". Oxford: Speech Communication, 2007 Sept. 49(9): 725-42.

[19] Barbosa PA. "Detecting changes in speech expressiveness in participants of a radio program In: Proceedings of Interspeech". Brighton. p. 2155-58, 2009.

[20] Hirst D. "The analysis by synthesis of speech melody: from data to models". Journal of Speech Sciences,1(1): 55-83, 2011.

[21] Yoshinaga-Itano C. "From Screening to Early Identification and Intervention: Discovering Predictors to Successful Outcomes for Children With Significant Hearing Loss". J Deaf Stud Deaf Educ,Winter; 8(1): 11-30, 2003.

[22] Wirz S. "The voice of the Deaf". In: Fawcus M (Edit). Voice Disorders and their Management. Croom Helm 1986.

[23] Tobey EA, Geers AE, Brenner CB, Altuna D. andGabbert G. "Factors associated with development of speech production skills in children implanted by age five". Ear & Hearing, Feb; 24(1): 36–45, 2003.

[24] Lattin, J; Carrol, D J D, and Green, P E. "Análise de dados multivariados". São Paulo: Cegrage Learning, 2011

[25] Xu L, Zhou N, Chen X, Li, and Schultz, Z. Vocal singing by prelingually-deafened children with cochlear implant. Hearing Research, Jun. 255: 129–34, 2009.

[26] Baudonck, ED; Dhooge, I. andLierde, KV."Objective vocal quality in children using cochlear implants: a multiparameter approach". J Voice, vol. 25, n 6, 2011, p. 683-691, 2011.

[27] Benninguer, MS. "Quality of the Voice Literature: What is There and What is Missing". JournalofVoice. Nov; 25(6): 647-52, 2011.

[28] Lee, KY, Tong, MC, Van Hasselt, CA. The tone production performance of children receiving cochlear implants at different ages. Ear Hear, v.28, n.2 Suppl, p.34S-37S, 2007.

Structural and Prosodic Correlates of Prominence in Free Word Order Language Discourse

Tatiana Luchkina [1], *Jennifer S. Cole*[1]

[1] Department of Linguistics, University of Illinois at Urbana-Champaign, USA
luchkin1@illinois.edu, jscole@ling.illinois.edu

Abstract

Production and perception experiments with native speakers of Russian, a free word order language, show that prosody and change in word order are used to mark discourse-prominent constituents. Concurrent application of these cues to prominence is possible, as evident from distinctively higher f0 and intensity maxima, and duration values associated with ex-situ words, as well as their higher visibility in discourse. Distinctive acoustic-prosodic realization of ex-situ words may cue their relatively high informational load and discourse prominence, as well as (redundantly) signal that the word is left- or right- dislocated.

Index Terms: Information structure, prosodic prominence, word order

1. Introduction

Independently of the modality of presentation, human processing of discourse involves identifying the information structure status of discourse elements. Information structure, and the related notion of accessibility, have been offered a variety of interpretations in the linguistic and psychological literature [1,2,3]. One approach adopts a tripartite distinction of discourse entities into the most accessible (or given, active), least accessible (or novel, inactive) and inferable (semi-active) [4]. Discourse-novel entities are described as more prominent than the discourse-given ones. New information (also known as focus or rheme), is opposed to discourse-given information, (topic or theme) as categories of information structure (IS) and are differentiated with regards to their visibility in discourse or *discourse prominence* [5].

Prosody, morphology, and the structural organization of information provide different means of encoding the discourse status of a word [6,7]. The prosodic encoding of discourse-prominence is a well-studied, psychologically real property of discourse production and perception [8,9,10]. It involves perceptually salient changes in voice quality, duration and intensity, as well as changes in f0. The degree of discourse prominence or relative discourse accessibility expressed by prosodic means can be further affected by structural means, i.e., strategic positioning of the discourse-prominent word in a syntactic phrase or clause. Structural prominence is especially suited for the so-called free word order languages, where the syntactic function of a word is marked overtly by means of morphological case. In such languages, the ordering of sentential constituents can be used to encode their relative discourse prominence, and signal their information structure category.

2. Two routes to prominence

Cross-linguistic studies show that languages use *either* prosodic marking or constituent ordering to encode discourse prominence and IS. Based on a comparative study of Italian and Turkish (relatively free word order languages), and English (fixed word order language), Donati & Nespor [11] propose that languages with rigid word order allow prosodic marking of discourse information ('focus') at different locations in the sentence, while languages with flexible word order do it through word order, and, consequently, exhibit less variation in the location of prosodic prominence. This model, with dual routes for the encoding of discourse information, predicts that it will be relatively uncommon that a language uses both word order and prosodic marking simultaneously to signal novel or important information in discourse. With respect to this [11] suggested to categorize languages into prominence dislocating (e.g., English), i.e., those which utilize prosodic cues to mark prominence, and constituent dislocating (e.g., Turkish or Italian), i.e., those which primarily use word order, or structural cues.

Consider how discourse prominence is expressed in English. While rightmost accent placement is the default location of the prominent constituent in English (1), the phenomenon of metrical reversal or stress shift [12, 13] can displace phrasal prominence leftward to signal the IS marked category of contrastive focus (2):

(1) Joel bought a green PORSCHE.
(2) Joel bought a GREEN porsche.

Italian illustrates a different prominence marking strategy [11], where speakers may choose to move the discourse-prominent word to a syntactic location that is systematically associated with its IS category or discourse-prominent status. Thus, while SVO is the canonical word order in Italian (*Mario arrive* 'Mario arrives'), the reverse ordering of S and V confers prominence to the subject (*Arrivo Mario*).

This sort of overt movement is often said to be prosodically motivated, which means that a word undergoes overt movement in order to be associated with the main phrasal prominence position, where it is perceptually salient.

While the proposal by Donati & Nespor parsimoniously accounts for the cross-linguistic distribution of the two distinct prominence encoding mechanisms, structural and prosodic, this work draws attention to the growing body of empirical evidence that word order is an optional resource for encoding IS categories and discourse-prominence in free word order languages, along with prosodic means [14,15,16,17, among others]. Languages which are known to display IS-triggered movement (Spanish, Greek, Russian, Georgian, and Italian, among others) are also known to use prosodic means to mark prominence in-situ, i.e., on constituents that are discourse-prominent, but which have not been moved to a syntactically designated prominent position. For such languages, one interesting question is whether the concurrent application of structural and/or prosodic means is purposeful, i.e., used to encode different categories of information in discourse.

Luchkina and Cole [18] report a preliminary investigation of this issue for Russian, a highly free word order language which simultaneously exhibits structural and prosodic prominence as means of marking information that is novel or

particularly salient. The semantically neutral, default word order in Russian is SVO, and as in other free word order languages, in Russian, a word can appear in its canonical position (in-situ), fronted, or post-posed. The ordering of the constituents in a sentence marks IS and not grammatical function. In the following example from [18], continuations (a) and (b) are both possible for the sentence in (4), but are felicitous under different discourse conditions: given the context provided in (4), the word Ivan, critical to the understanding of who does the cooking, may be located in the rightmost position, where it is structurally prominent (as in b), or may occur pre-verbally as in (a) and receive (optional) prosodic prominence.

(4) Tri druga, Ivan, Petr, i Andrey, nahsli novjiy retsept pizzj.
Three friends, Ivan Petr and Andrey, found a new pizza recipe.
a. (Smotri!) IVAN gotovit pizzu.
(look) Ivan-SUBJ cooks pizza-OBJ

b. (Smotri!) Pizzu gotovit Ivan.
(look) pizza-OBJ cooks Ivan-SUBJ

Luchkina & Cole present an analysis of prominence ratings provided by linguistically naïve native speakers of Russian and find that salient acoustic-prosodic, as well as structural cues to discourse-prominence are used jointly in judgments of the prominence status of a word. The prominence ratings reported in [18] are based on the production of one (model) speaker, which raises the question of whether utterances produced by other Russian speakers would yield similar perceptual ratings.

In this paper we extend our earlier work and explore how prosody and word order function independently and in combination to mark IS in a free word order language like Russian. The proposed experimental design seeks to determine (1) whether in discourse, the sentential position of a word affects it perceived prominence; (2) and whether cross-application of prosodic and structural cues translates into a yet greater degree of discourse prominence.

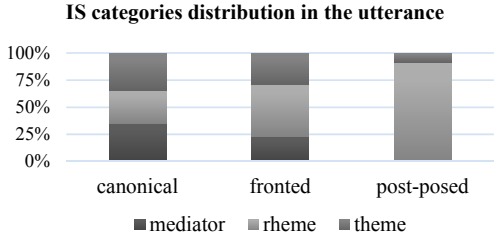
We test the hypothesis that in free word order discourse, an ex-situ sentential position acts as an independent cue to prominence, and if so, whether prominence may be further reinforced with acoustic-prosodic features associated with such position. To this end, we analyze word order and IS properties of two authentic Russian narratives for acoustic evidence of prosodic marking in relation to word order and IS. We (1) use the word-level prominence ratings obtained in a perceptual prominence rating task to gauge the native speakers' sensitivity to prosodic and structural means of encoding discourse-prominent information and (2) match the two classes of prominence correlates, structural and acoustic-prosodic, with the perceived prominence scores to test the ability of each of these sources of discourse-relevant information to predict the word IS category in the narratives chosen for this study. We analyze the experimental results with respect to the interaction between word order, prosodic marking, and perceived prominence.

3. Experiment 1: Production task

3.1. Materials

Both discourse samples used in this study come from two published narratives, which display a range of word order and prosodic features. With an average sentence length of 5.2 content words (SD=1.77, 230 content words), approximately 30% of the sentences in the narratives deviate from the canonical SVO order. Adopting the discourse annotation framework introduced in [3], the distribution of the following IS categories[1] was assessed: 'THEME' (word carrying discourse-given information, 2-12 mentions per narrative), 'RHEME' (words carrying discourse-new information, 1st mention in the narrative), 'MEDIATOR' (words carrying inferable information, 1st mention in the narrative), and 'CONNECTOR' (function words). The sentential position of each word in the narratives was marked as in-situ or ex-situ (specifically, 'fronted' or 'post-posed', relative to SVO order).

Figure 1: *Observed distribution of IS categories in the analyzed discourses*



The observed distribution of IS categories differed by sentential position (Pearson $\chi^2(6)=85.91$, $p<0.001$). Figure 1 illustrates that discourse-novel information is the most 'mobile' information category in the corpus, with more ex-situ locations than any other IS category, and clearly dominates the post-posed (utterance-final) position. Such non-random distribution of discourse information provides evidence that word order variability in the narratives under analysis may be indicative of IS category.

3.2. Acoustic-prosodic features pre-processing & production data overview

For the purposes of the perceived prominence rating task, both narratives were read orally by a female speaker of Russian (henceforth, the model speaker), age 27. Read productions from 14 speakers (ages 21-52, 8 females) were collected for the purposes of extended acoustic-prosodic analysis of the study materials. The acoustic-prosodic measures of f0 and intensity mean, minima, and maxima, f0 range, velocity, and excursion size, as well as vowel duration were taken from every syllable of each IS-coded content word in the corpus. Additionally, distance in milliseconds from the tonal center of gravity of each stressed vowel to the vowel midpoint was measured. All measurements were extracted automatically in Praat [20]. The values of max f0 and max intensity were taken from the center region of the vowel in order minimize the influence of the adjacent segments at the voice onsets and inter-segmental transitions. Each f0 output was transformed to semitone values relative to a fixed value of 100 Hz[2]. Prior to

[1] The IS category of each content word was annotated by one of the authors (TL) and another native Russian speaker. Inter-rater agreement (linearly weighted Kappa) between the annotators, across texts was satisfactory: $\varkappa=0.89$, SE=0.03, $\alpha=0.05$.

[2] The semitone scale was chosen to reduce male-female acoustic differences and ensure that +1 standard deviation constituted a perceptually equivalent interval as -1 standard deviation.

being submitted to regression analyses, all acoustic-prosodic measures were centered using mean-centered coding [21].
Extracted acoustic measures were examined as correlates of prosodic prominence, and analyzed for their relationship to the IS and sentential position of a word.
The model speaker's production data were subject to a preliminary analysis to determine how well cross-linguistically attested prosodic correlates of prominence (intensity and f0 maxima, and duration of the stressed vowel) can be predicted from the word's IS category and position in the utterance. A series of linear regression analyses each featuring one acoustic parameter of interest as the dependent variable revealed that both the ex-situ sentence position and IS category RHEME are reliably associated with greater intensity and f0 maxima, as well as longer duration of the stressed vowel (see Table1).

Table 1: *Significant predictors of vowel intensity, duration, and f0, with respect to the carrier word*[3]

(max)intensity	(max)f0	vowel duration
ex-situ (t=3.16, p<0.005)	ex-situ, fronted (t=6.36, p<0.001)	ex-situ (t=2.57, p=0.01)
RHEME (t=2.32, p=0.02	RHEME (t=2.84, p=0.005)	RHEME (t=2.94, p<0.01)

Significant patterns of co-variation were observed between some of the acoustic measures: e.g., a significant rise in max f0 was reliably associated with concurrent increases in intensity (t=24.77, p<0.001) and duration (t=5.35, p<0.001). Analysis of the model speaker's production data established that the acoustic parameters of f0, intensity, and duration contribute perceptual salience to the words carrying discourse-novel information and/or occurring ex-situ. Before we validate this finding with production data from multiple speakers, we test the psychological reality of the acoustic-prosodic and structural cues to prominence, i.e., gauge their ability to affect reader's or listener's perception of a word as prominent during discourse comprehension.

4. Experiment 2: Prominence rating task

Structural and acoustic-prosodic cues to prominence were determined in reading and auditory comprehension tasks performed by linguistically naïve native speakers of Russian (N=49 (reading modality), N=27 (auditory modality)). The task included thirty-nine clause-size excerpts from the narratives. Each clause, or target segment, was presented along with the preceding context. The mode of presentation of the target sentence was either written text or audio recording of the model speaker's production. Respondents read the entire portion of the text preceding the target segment, read or listened to the target segment and identified discourse-prominent word(s) in the target segment by associating them with one level of the binary feature "+/- prominent". Following [22], no formal definition of prominence was given. Participants were instructed to mark only those words that 'were the focus of their attention' in the utterance, based on the preceding context. Any number of content words could be marked as prominent.

4.1. Results

Assessing consistency of the responses: Responses to the prominence rating task were assessed for intra- and inter-rater agreement. The kappa coefficients assessing consistency of the intra-rater rating behavior fall within the range 32.4 – 88.1 (mean=70.4, SE=2.21) and translate into moderate to very high agreement. The inter-rater agreement coefficients translate into fair, though highly significant agreement levels: Fleiss' kappa=0.26 (p<0.001) for the written and 0.36 (p<0.001) for the auditory modality.
Overall picture of the perceptual prominence ratings: Following [9], each word in the narratives was assigned two discourse prominence scores (one per test modality). Prominence scores were obtained by dividing the total number of times a word was chosen as salient by the total number of participants who responded to the relevant test question. Additionally, a global prominence score was computed for each word to represent its perceived prominence across the presentation modalities. In order to gauge the respondents' sensitivity to word order changes in the narratives, a two-way analysis of variance assessed prominence of the IS categories represented in the corpus under (a) canonical and (b) non-canonical word order. With the global prominence score used as the dependent variable, the ANOVA crossed the factors IS and Word Order (see Table 2). Results revealed no significant main effect of Word Order (F=0.137, p>0.05), a significant main effect of IS (F=47.82, p<0.001) and a significant interaction between these two factors (F=3.36, p<0.01).

Figure 2: *Global perceived prominence scores (y-axis) of the 4 IS categories (x-axis) as determined by the combined results of the silent reading and listening prominence rating tasks.*

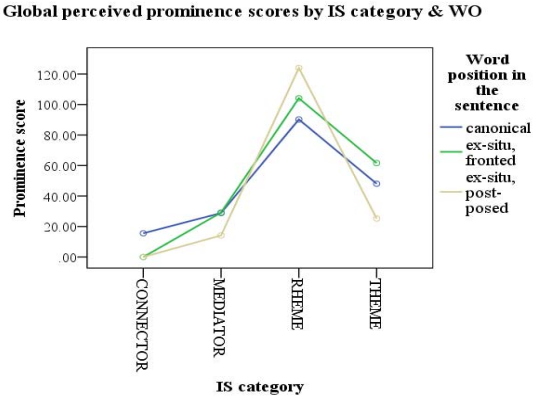


Figure 2 illustrates that IS categories with highest prominence ratings, THEME (discourse-given information, mean prominence score =49.6, SE=6.0) and RHEME (discourse-novel information, mean prominence score =106.0, SE=3.5), demonstrate a meaningful dissociation in the prominence scores as a function of the word position in the sentence. Pairwise comparisons reveal that both *in-situ* (canonical) and *fronted words* carrying *discourse-given* information are viewed as significantly more discourse-prominent than those post-posed relative to their canonical position (mean

[3] Factors '*speaker*' and '*word*' (not shown in Table1) were included in the model as random effects.

difference=22.83, p=0.05 for in-situ words, and mean difference=36.4, p<0.05 for fronted words). The dissociation goes the opposite way for the ex-situ words carrying *discourse-new* information (RHEME): *post-posed* discourse-novel words receive significantly higher discourse prominence scores than those appearing the in-situ or fronted, relative to their canonical position (mean difference=33.7, p<0.001 for in-situ words, and mean difference= 19.9, p=0.05 for fronted words).

Correlates of perceived prominence: Modality-specific prominence scores were modelled with linear regression analyses. Predictors for auditory prominence perception were extracted from the recording of the model speaker's reading performance and included the acoustic-prosodic measures of standardized f0 and intensity maxima, f0 range and excursion size, and vowel duration taken from the stressed syllable of each IS-coded content word.

In silent reading of the Russian corpus, words located in ex-situ position, specifically, post-posed (t=7.17, p<0.001) or fronted (t=5.78, p<0.001), relative to the canonical position, as well as words carrying discourse-novel information (t=2.69, p<0.01) were associated with higher prominence scores.

In the auditory modality, these factors were complemented with acoustic predictors duration (t=2.23, p<0.05), intensity (t=2.27, p=0.05), and f0 range (t=2.78, p<0.01).

4.2. Introducing production data from multiple speakers

While results of the auditory prominence rating task present compelling evidence for structural and prosodic prominence being utilized by the listeners during discourse comprehension, they hinge on the reading performance of the model speaker. To test whether the perceived prominence scores from the auditory modality are consistent with the IS properties of the narratives encoded via word order and multi-speaker prosody, acoustic-prosodic measurements of the model speaker's reading performance were augmented with those from 14 speakers (who did not participate in the prominence rating task). A multinomial logistic mixed effects regression was fit to the data to model the IS category of a word (discourse-given or THEME, discourse-novel or RHEME, inferable or MEDIATOR) using its auditory modality prominence score, sentence position, and the acoustic measures for that word extracted from the productions of the 15 speakers[4]. Results of the logistic regression analysis confirm that structural and prosodic correlates of discourse prominence successfully predict the IS category of content words. Specifically, discourse-novel information, which is the IS category that received the highest perceived prominence scores (z=6.69, p<0.001), bears a reliable association with ex-situ sentential positions (z=1.77, p=0.08 for fronted RHEMEs; z=1.97, p<0.5 for post-posed RHEMEs) and distinctive acoustic-prosodic realization. The latter is supported by a number of parameters from oral productions by 15 Russian speakers, including f0 range (z=1.99, p<0.05) and maxima (z=3.23, p=0.001), lower mean f0 values (z=-3.18, p=0.001), as well as greater duration (z=2.10, p<0.05) and mean intensity (z=2.28, p<0.05) associated with the stressed vowel of RHEMEs. *Discourse-inferable* words (MEDIATOR), on the other hand, are characterized by lower perceived prominence scores (z=-4,84, p<0.001) and smaller duration (z=-4.30, p<0.001) and mean f0 values (z=-2.12, p<0.05).

[4] The regression model included two random effects, *segment* and *speaker*. Each level of the variable *segment* corresponds to one unique vowel from which the acoustic-prosodic measurements were extracted.

5. Discussion

The goal of this work is to parameterize perceived prominence in a free word order language like Russian and understand which factors guide naïve readers' or listeners' perception of a word as prominent in a discourse or narrative. To this end, we offer an empirical test of whether variation in word order, along with the more established acoustic-prosodic ways of marking discourse-prominent information, can serve as a means of encoding the information status of a word and, by doing so, mediate its perceived prominence.

Analyses of the production and perceptual prominence ratings data presented in this work successfully capture the close interrelatedness of the prosodic and structural cues to prominence in the corpus of two published narratives. Results demonstrate that independent of the modality of presentation, words that carry discourse-novel information are perceived as more prominent.

In a free word order language such as Russian, information status is encoded via two routes, prosodic and structural. In the auditory modality, listeners treat the acoustic-prosodic realization of a word as a cue to its discourse status. This is evident from the finding that distinctive f0 qualities, greater intensity and vowel duration reliably trigger perception of a word as prominent. About 30% of the utterances in the mini corpus deviate from canonical word order, which means that words carrying novel or given information may occur ex-situ, i.e., be fronted or post-posed, relative to the canonical SVO position. Results of the prominence rating task show that apart from the acoustic effects of prosody, an ex-situ position of a word also contributes to its perception as prominent.

Analysis of the syntactic and acoustic-prosodic characteristics of perceived prominence reveals that Russian allows cross-application of different cues to prominence within the same utterance. Ex-situ positions are associated with (1) a higher prominence score and (2) in the auditory modality, distinctive perceptual qualities. Such distinctive acoustic-prosodic realization of non-canonically positioned words may not only cue their relatively high informational load and discourse prominence, but, in a language that exhibits focus fronting and right-edge dislocation for IS purposes [16,25], may also (redundantly) signal that the word is left- or right- dislocated.

6. Conclusion

This study contributes to the understanding of discourse-prominence in a free word order language. Results of the production and perception experiments performed by linguistically naïve native speakers of Russian reveal that concurrent application of prosodic and structural cues does not preclude either cue class from being perceived as a signal to information that calls for special attention in discourse. Further work is necessary to determine whether cross-application of prominence cues is characteristic of all vs. select categories of discourse-prominent information and whether its effect is additive, i.e., leading to a word being associated with a yet greater degree of perceived prominence.

7. References

[1] Watson, D. G. (2010). The many roads to prominence: Understanding emphasis in conversation. In B. Ross (Ed.), The Psychology of Learning and Motivation, 52, 163-183. Elsevier.

[2] Chafe, W.L. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Li, C. (ed.), Subject and Topic. New York: Ac. Press, 25-55.

[3] Arnold, J.E., Losongco, A., Wasow, T., & Ginstrom, R. (2000). Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. Language, 76, 28-55.

[4] Calhoun, S., M. Nissim, M. Steedman, and J.M. Brenier. (2005). A framework for annotating information structure in discourse: Pie in the Sky. Proceedings of the workshop, ACL, 45–52.

[5] Katz, J., Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. Language, 87, 771-816.

[6] Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. Cognitive Psychology, 19, 498–550.

[7] Stolterfoht, B., Friederici, A., Alter, K., and A. Steube (2007). Processing focus structure and implicit prosody during reading: Differential ERP effects. Cognition 104, 565-590.

[8] Watson, D. G., Arnold, J. E., & Tanenhouse, M. K. (2008). Tic Tac TOE: Effective of predictability and importance on acoustic prominence in language production. Cognition 106(3), 1548-1557.

[9] Mo, Y., Cole, J., & Lee, E. (2008). Native listeners? Prominence and boundary perception. Speech Prosody 2008.

[10] Kaland, C., Krahmer, E. & Swerts, M. (2011). Contrastive intonation: speaker- or listener-driven? In W.S. Lee & E. Zee (Eds.), Proceedings of the 17th International Conference of Phonetic Sciences, 1006-1009. Hong Kong: City University of Hong Kong.

[11] Donati, C. & N. Nespor (2003). From Focus to Syntax. Lingua, 113-11, 1119-42.

[12] Neeleman, A., Reinhart, T. (1998). Scrambling and the PF interface. In: Butt, M., Geuder, W. (Eds.). The Projection of Arguments: Lexical and Compositional Factors. CSLI Publications, Stanford, CA, 309–353.

[13] Calhoun, S. (2010). The Centrality of Metrical Structure in Signaling Information Structure: A Probabilistic Perspective, *Language*, 86(1), 1-42.

[14] Skopeteas, S. & Fanselow, G. (2010). Focus in Georgian and the expression of contrast. Lingua, 120, 1370-1391.

[15] Sekerina, I. A. (1999). The Scrambling Complexity Hypothesis and Processing of Split Scrambling Constructions in Russian. Journal of Slavic Linguistics, 7.2, 218-265.

[16] Slioussar, N. (2011a). Processing of a free word order language: The Role of Syntax and Context. Journal of Psycholinguistic Research, 40:291-306.

[17] Arvaniti, A. & Adamou, E. (2011). Focus expression in Romani. In Proceedings of the 28th West Coast conference on Formal Linguistics, Somerville, MA: Cascadilla Proceedings Project.

[18] Luchkina, T. & Cole, J. S. (2013). Routes to Prominence in Free Word Order Language Discourse, in Mertens, P. & A.C. Simon (eds), Proceedings of the Prosody-Discourse Interface Conference 2013 (IDP-2013). Leuven, September 11-13, 2013, pp. 13-19.

[19] Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.60.

[21] Mitchell, M. (2012). Interpreting and visualizing regression models using Stata. Stata Press, College Station, TX.

[22] Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2011). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, v.1, 2011, p. 425-452.

[23] Neeleman, A., Titov, E. (2009). Focus, contrast, and stress in Russian. *Linguistic Inquiry* 40, 514–524.

16 Friday 2

Segmental Influences on the Perception of Pitch Accent Scaling in English

Jonathan Barnes[1], *Alejna Brugos*[1], *Nanette Veilleux*[2], *Stefanie Shattuck-Hufnagel*[3]

[1] Boston University, Boston, Massachusetts, USA
[2] Simmons College, Boston, Massachusetts, USA
[3] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
jabarnes@bu.edu, abrugos@bu.edu, veilleux@simmons.edu, sshuf@mit.edu

Abstract

In both tone and intonation systems, segmental context is known to influence production and perception of target F0 contours in various ways. Many languages, for example, prefer to realize critical F0 events during maximally sonorous intervals, either by varying the timing of pitch movements, or by virtue of distributional limitations on certain contour types. Current analytic practice, by contrast, routinely ignores segmental backdrop when estimating the perceptual efficacy of putative cues, such as F0 turning points, to tone scaling and timing patterns. Results of the perception study presented here argue that pitch accent scaling is best modeled using a weighted average of F0 sampled over a defined region of interest, and that individual sample weights are determined in part by the sonority of the segments from which they are taken. That is, samples from lower sonority segments contribute less to integrated scaling percepts than those from higher sonority segments. This model, called TCoG-F(requency), accounts for crosslinguistic tonal timing and distribution patterns in the literature, and underscores the danger of analyzing tonal phenomena completely apart from the segments that express them.

Index Terms: Intonation, Pitch perception, tone scaling, tonal timing, sonority, Tonal Center of Gravity.

1. Introduction

Since the dawn of the autosegmental era in tonal and intonational phonology [10, 19, 4, 22], we have grown accustomed to thinking of linguistic pitch specifications as existing apart from, or parallel to, the segmental skeleton of the spoken utterance. Specifications on the so-called tonal tier must then be associated, according to the dictates of the grammar, with appropriate Tone-Bearing Units in order to be realized phonetically. At the same time, however, it is well known that both perception and production of F0 can be influenced significantly by the segmental contexts in which contours are realized. For example, 'microprosodic' effects on the F0 contour stem from, e.g., voicing differences in syllable-initial consonants, differences in vowel height between otherwise comparable syllables, etc., and experimentalists routinely control for such effects [15, 17, 26].

Less widely appreciated, however, are the ramifications of a commonly-remarked tendency for languages to avoid realizing critical portions of F0 contours within lower-sonority regions of the segmental string. This tendency manifests itself in various ways: On the one hand, intonational phonologists have observed what appear to be systematic alterations to tonal timing patterns in order to ensure optimal expression of F0 contours in a given segmental context. For example, in a variety of languages, accentual High F0 targets occur relatively earlier in closed syllables than in open, and in syllables closed by obstruents than in those closed by sonorants [5, 18, 23, 24, 25]. On the other hand, languages can impose categorical distributional restrictions on the association of certain tone patterns with particular kinds of segment hosts. For example, cross-linguistically, contour tones tend to be restricted to syllables with longer, higher-sonority rhymes [11, 28, 29].

The latter pattern, typically observed in languages with lexical tone contrasts, has been explained as resulting from the comparatively greater salience of the percept of pitch during segments that are higher in intensity and richer in harmonic structure [8, 29]. Under this scenario, languages deploy their fullest array of tonal contrasts only in contexts where these contrasts are most likely to be accurately perceived. This explanation could account for the first cases mentioned above as well: If F0 peaks associated with intonational High pitch accents occur relatively earlier in closed syllables than in open, for example, we might attribute this to speakers' desire to realize critical pitch information (i.e. the bulk of elevated F0) within a more sonorous portion of the syllable (i.e. the nucleus rather than the coda).[1]

Against this backdrop of general awareness of the influence of segmental context on tone perception and production, there is also a somewhat paradoxical countervailing tendency to assume that putative tonal targets such as F0 turning points are of equal perceptual value regardless of the nature of their host segments. According to this practice, linguistically meaningful pitch accent scaling patterns are equated phonetically with measured F0 maxima, regardless of where in the segmental string those maxima fall; similarly, the temporal location of F0 turning points is estimated according to the visual salience of 'corners' in the F0 track, regardless of whether those corners fall in regions of the signal with high or low auditory salience. In an analogous vein, the subjective continuity of intonation contours, even through regions where F0 is heavily disrupted by intervals of voicelessness, has led some to assume that listeners have an ability to effectively 'restore' missing F0 intervals to the signal via interpolation, or extrapolation based on existing trajectories [12, 13, 21]. Accordingly, F0 stylization algorithms such as the Fujisaki model, MOMEL, or Tilt [9, 14, 27] create continuous F0 tracks based on gappy originals, in some cases even locating critical F0 target points within such 'filled in' intervals, when the shape of the interpolated pitch curve suggests it. It is the tension between this tendency, on the one hand, and the literature on avoidance of low-

[1] Another explanation that has been offered for these altered timing patterns is based on House's Spectral Stability Hypothesis [16, 7, 24], whereby pitch movements are more readily perceived as such when they are realized during regions of spectral stability. For reasons of space, we will not treat this hypothesis further here, other than to note that it relates less obviously to the distributional restrictions on lexical tones noted above.

sonority segmental hosts on the other, that inspired the experiments described in this paper.

A first step toward resolving these issues was taken recently by [1], who demonstrate that, at least for perceived F0 target scaling, the "perceptual completion" approach cannot be correct.[2] In that study, subjects made judgments regarding the scaling of synthetic English High pitch accents (L+H*) realized either as clear peak- and plateau-shaped F0 contours extending over fully voiced segmental intervals (e.g., in a context like '*DAY' might fit*), or as analogous contours in which the region corresponding to the nuclear pitch accent contained 'missing' or inferable peaks/plateaux (i.e. mirror-image rises and falls separated by the closures and releases of voiceless stops, as in a context like *'DATE' might fit*). Rather than either extrapolating or interpolating F0 across such voiceless intervals in a way that would register systematically on scaling judgments, subjects were seen to behave as though the missing intervals were absent altogether, ignoring the gaps, and judging relative pitch accent scaling exclusively on the basis of the F0 values actually present in the signal.

While that study gives an indication of how listeners treat F0 gaps created by voiceless stops, it remains unclear what listeners do with intervals in which F0 is in fact present, but with lower-amplitude or spectral impoverishment. In the current study we use similar investigative techniques to those of (1), applied specifically to the perception of measurable F0 over lower-sonority intervals. We hypothesize that vowel vs. silence are in fact two ends of a continuum of possible F0 carriers with differing degrees of salience, along which higher sonority segments such as liquids or nasals give way gradually to lower sonority ones, such as voiced fricatives or stops. We predict that these sonority-based differences in the robustness of perceived pitch will be manifested in listeners' scaling judgments. We situate the results of this study in a model of tone scaling perception we call TCoG-F (Tonal Center of Gravity in the Frequency dimension), where the perceived scaling of an F0 event (e.g., the elevated F0 associated with a High pitch accent) is modeled as a weighted average of F0 measured over a particular region of interest. In the calculation of this average, F0 samples taken from more sonorous regions are accorded heavier weights, and are thus predicted to extend relatively greater influence over perceived scaling.

2. Methods

The reasoning behind the current design is as follows: Consider a set of utterances, such as those depicted in Figure 1, bearing rise-fall-rise intonation contours, with plateau-shaped L+H* nuclear pitch accents on the first word of the sentence (*'X' might fit*, uttered perhaps in the context of the solution to a crossword puzzle clue), where F0 rises through the nuclear vowel, remains high and level through the coda nasal, and falls thereafter, before rising at the end to signal something like tentativeness. In all three examples, the syllable rhymes (and their constituent nuclei and codas in a. and c.) are identical in duration, as are the relevant segments of F0 contour. What differs is only the sonority of the portion of the syllable rhyme bearing the high, level portion of the accentual plateau for the three target words *Dane*, *day* and *Dave*. Assuming, as hypothesized above, that the perceived scaling of this pitch accent involves averaging over F0 samples taken during the entire rhyme of the accented syllable, we expect first that the perceived scaling of all three of these words will end up lower than the maximum F0 realized during the plateau-portion of the contour. However, to the extent that the perceptual contribution of any given F0 sample is weighted by a factor representing the sonority of the segment bearing it, we expect the perceived scaling of the pitch accents in these three utterances to differ from one another as well. Since the highest portion of the pitch accent in *day* (i.e. the plateau) occurs in a region of greater sonority than the analogous portion of the pitch accent in *Dane*, the high F0 samples for *day* will contribute more to the resulting average, so that *day* will sound higher to listeners than *Dane,* despite identical F0. Correspondingly, owing to the lower sonority of this region in *Dave,* we predict the pitch accent in 1c to sound lower than those in a and b, again despite the lack of difference in objective F0.

To test these predictions, we designed a set of experimental stimuli similar to those used by [1] but differing in several critical ways. As just described, all stimuli were instances of the English words *day, Dane,* and *Dave,* realized in the target position of the frame sentence *X might fit*. All target stimuli were realized with rise-fall-rise intonation contours (ToBI L+H* L-H%), with the nuclear pitch accent on the first word. However, F0 for these base utterances was resynthesized to create contours of two basic types: plateau-shaped pitch accents and sharp-peak-shaped pitch accents, in effect extrapolating the preceding rise and following fall to a single higher intersection point instead of a plateau. These are both depicted in Figure 1, and their deployment in our task is detailed in Section 2.2 below.

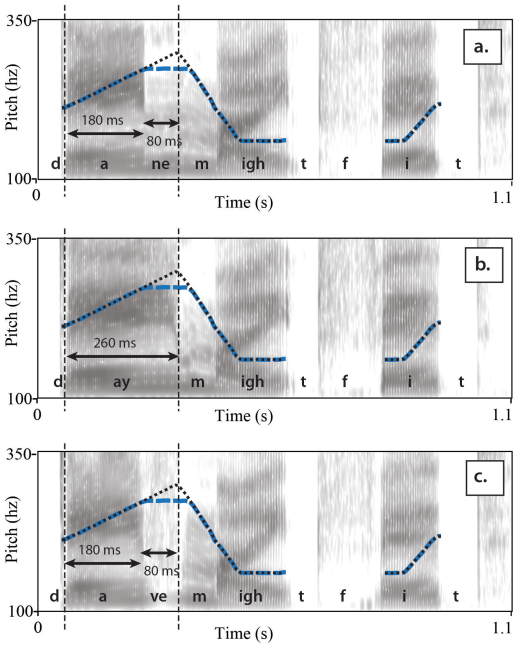


Figure 1: *Spectrograms with superimposed pitch tracks for plateaux (blue dashed lines) and peaks (black dotted lines) for* Dane *(a),* day *(b) and* Dave *(c).*

[2] At least in the most literal sense. It is still possible that non-F0 cues within voiceless regions contribute either to the 'subjective continuity' of the pitch contour, or to the perception of 'prominence' (a linguistic dimension sometimes cued in part by the higher-than/lower-than relations that underly linguistic pitch distinctions [20]. At the same time, is worth noting the extent to which relative pitch and prominence relations vary orthogonally (as in the expression of lexical tone contrasts, or the difference between downstepped and non-downstepped nuclear pitch accents).

2.1. Stimulus Creation

Target phrases were created from two sets of base recordings, one produced by a male native English speaker, the other by a female, and then resynthesized using Praat [3]. Synthesized segment durations for the female speaker, given in Fig. 1, were based on mean values over multiple utterances. For peaks, F0 rises were identical in duration (260 ms) and scaling (212-300 Hz, a 6 st rise) for this speaker for all stimulus types, and were followed by a 140 ms fall to 160 Hz. Plateau stimuli had a rise of the same slope as the peaks, but truncated after 180 ms (212-273 Hz), followed by a 101 ms plateau (at 273 Hz), and a 119 ms fall to 160 Hz (the same slope as in the peak stimuli, but starting from the end of the plateau). Duration and F0 values for stimuli based on the male speaker were comparable, though different in quantitative detail.

2.2. Experimental task

Our primary question pertains to the perceived relative scaling of nuclear L+H* pitch accents realized on syllables with differing rhyme types. However, to avoid the potential for confounds inherent in the direct pairwise comparison of syllables with differing segmental content, the relative scaling of these contours was investigated indirectly. That is, the bulk of experimental trials consisted of the pairing of a given target item (i.e. a target utterance with *day, Dane, or Dave,* with either a peak- or plateau-shaped pitch accent) with one out of a continuum of standard reference contours. These reference contours were segmentally identical to the target item, but F0 throughout the accented syllable was held steady at one of 7 levels, the highest at 300 Hz, descending thereafter in .5 semitone increments (Figure 2). After the fall from the accented syllable, F0 was identical for target items and standards.

If we assume that corresponding level standards sound identical in scaling regardless of syllable type, since sonority-related weighting variation cannot change perceived pitch for a flat-F0 contour, then any effects of F0 weighting differences on tokens with changing F0 should be manifest in subjects' perception of the relative scaling of target items and their respective level standards. (E.g., *day* might sound equal in pitch to its standard level 5, while *Dave* might reach only level 4.)

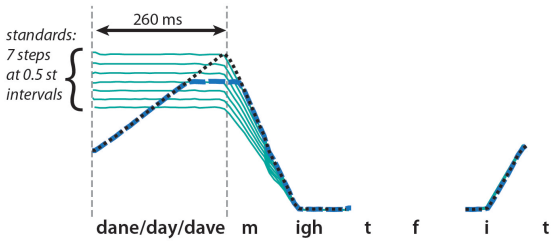


Figure 2: *Pitch tracks for standards (teal solid), plateaux (blue dashed) and peaks (black dotted).*

The task itself was 2AFC: 77 native speakers of American English were presented with pairs of contours (a target item and a level standard), and decided which contour's target word reached a higher pitch. After 6 consecutive correct responses in an initial block comparing standards separated by ≥ 3 steps, the experiment began, with each test item (2 accent shapes X 3 word types) compared to its continuum of 7 standards (2 reps of standards 1, 2 & 7, and 3 reps of standards 3, 4, 5, & 6 = 18) in 2 orders, for a total of 216 trials. Additionally, there were two more trial types interspersed. The first was represented by 36 trials pairing two level standards separated by either 2 or 3 continuum steps for *day* target types (18 comparisons x 2 orders). These trials served as a baseline measure of participants' accuracy in discriminating pitch levels. The final trial type involved 18 pairings consisting of each sharp-peak-shaped version of a given syllable type with its plateau-shaped counterpart (3 reps X 2 orders X 3 word types). These trials served as an additional test of the hypothesis that the phenomena under investigation here are in fact the result of lowered perceptual salience of F0 samples taken from utterance intervals of lesser sonority. In all such pairings, as with the level vs. level comparisons, there is in fact a "correct" answer: The sharp peak version of the pitch accent should sound higher than its plateau-shaped counterpart because it is, in fact, by any measure, higher.[3] On the other hand, since the entirety of the region of F0 difference between peaks and plateaux fell within the region of sonority difference across word types, if our hypothesis is correct, the scaling difference between the two should be relatively easy to detect for *day,* but progressively harder in the lower sonority rhymes of *Dane-* and *Dave*-type stimuli. All trial types were mixed together, and all 270 experimental trials were presented in random order, with breaks after every 50 trials. (See Figure 3 for a schematic summary of the 4 trial types.)

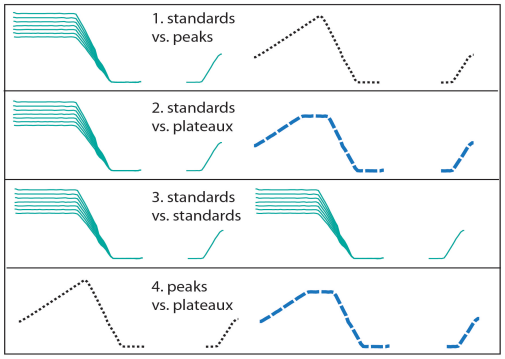


Figure 3: *Schematic showing the 4 types of experimental trials.*

2.3. Results and analysis

Data from 62 participants is included in the analysis. (Of 77 total, 15 of whom did not reach criterion for inclusion based on discrimination of level standards.) Fig. 4 displays results, pooled across subjects. Lines represent the percentage of trials in which *day, Dane,* or *Dave* was judged higher than each of its 7 level standards. Comparing target types, the percentage of 'higher-than' judgments for *Dave* clearly declines earlier in the continuum of level standards than does that of *Dane,* which in turn declines earlier than *day*. We infer from this that *day* sounds higher to listeners than *Dane,* and that *Dave* sounds lower. This is confirmed by a mixed-effects logistic regression analysis, using both standard level and target-syllable type, as well as accent shape (peak vs. plateau) as fixed factors, and participant as a random factor. The resulting model (N = 12,971, log-likelihood = -5573) shows a main effect of standard level (Est. = -0.934 (SE = 0.017), Wald Z = -55.73, p

[3] Mean F0 for the accentual high region, however measured and weighted, was higher for peaks than for plateaux.

< .001), and of word type, with *day* differing in a positive direction from *Dane* (Est. = 0.256, (*SE* = 0.079), Wald *Z* = 3.24, *p* = .001), and *Dave* differing from *Dane* in a negative direction (Est. = -0.504 (*SE* = 0.08), Wald *Z* = -6.29, *p* < .001). Importantly, in addition to a main effect of accent shape, with plateaux differing in a negative direction from peaks (Est. = -0.941 (*SE* = 0.082), Wald *Z* = -11.35, *p* < .001), there was also a significant interaction between accent shape and word type: the peak-plateau difference was less salient for *Dave* than for *Dane* (Est. = 0.276 (*SE* = 0.118), Wald *Z* = 2.34, *p* < .05) (*day* did not differ significantly from *Dane* in this respect).

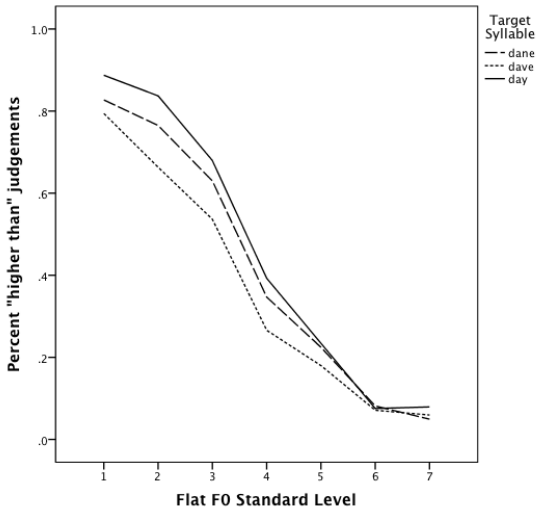


Figure 4: *Percent 'Higher-than' judgments for the three syllable types, as a function of the level standard against which they were compared.*

These results strongly bear out the predictions of the TCoG-F hypothesis detailed above. *Day* tokens sounded higher than *Dane,* which sounded higher than *Dave.* Critical in understanding this primary result is the significant interaction between accent shape and word type. The fact that the perceived scaling difference between peaks and plateaux was less pronounced when realized over the voiced fricative in *Dave* than in the nasal coda of *Dane* suggests that pitch percepts stemming from the former region are indeed less robust than those originating in the latter.

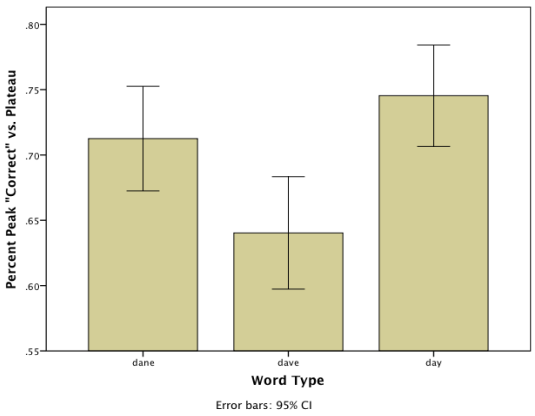


Figure 5: *Percent of peak vs. plateau trials in which peaks were correctly judged to sound higher than plateaux, for three target syllable types.*

This same conclusion is supported by trials in which subjects compared peak-shaped versions of a contour directly with their plateau analogues (Figure 5). Listener judgments of these comparisons were most accurate (i.e. listeners heard peaks as higher) for *day,* less so for *Dane*, and were least accurate for *Dave.* Another mixed-effects logistic regression (with word type as a fixed effect and participant as a random effect, *N* = 1462, log-likelihood = -849.4) shows that, while the *day-Dane* distinction was non-significant (Est. = 0.187 (*SE* = 0.15), Wald *Z* = 1.248, *p* = .21), the distinction between *Dave* and *Dane* was significant in the predicted direction (Est. = -0.358 (*SE* = 0.144), Wald *Z* = -2.494, *p* = .012). This lowered accuracy for scaling judgments where the sole difference between the two F0 contours lies within the lower sonority region of the coda is, again, just what TCoG-F would predict: Lower F0 sample weights during the crucial interval understate the objective difference between the two contours and make judgments more error-prone.

It is also worth noting that the connection between this result and the earlier one is a first glance not obvious: on the face of it, the fact that *Dave* trials sounded systematically lower than *Dane* or *day* to our listeners when paired with level standards bears no logical connection to the degree of accuracy listeners might exhibit when comparing one kind of *Dave, Dane,* or *day* contour to another in paired scaling judgments. The connection between the two results becomes clear only through the lens of TCoG-F.

3. Conclusions

While from a phonological point of view, the advent of autosegmentalism brought with it a great deal of progress and innovation, in the domain of phonetic realization, there remains a persistent danger that the core autosegmental insight, the separation of tonal and segmental phenomena onto distinct representational "tiers", may at times be taken too literally. The findings described here suggest that listeners' perception of F0 in speech signals is influenced by the nature of the consonant and vowel segments over which the F0 pattern occurs, even when voicing continues through those segments; that is, F0 values in regions controlled by more-sonorous segments (like vowels and nasals) are weighted more heavily than F0 values in regions controlled by less-sonorous segments (like voiced fricatives). This means that models based on a straightforward mapping between values of F0 peaks and valleys on the one hand, and perceived intonational targets, on the other, will need to be modified to take account of the influence of host segments. Taken together with earlier findings demonstrating the role of F0 contour shape on perceived tonal target alignment [2], these results support a model of intonation processing based on the Tonal Center of Gravity in both the time and frequency domains. We suggest furthermore that perceptual registration of individual cues in the speech signal is only the first step in the process of integrating multiple cues to form a single linguistically-relevant auditory percept. Future work will test this hypothesis in languages other than American English.

4. Acknowledgements

We gratefully acknowledge the support of NSF grants 1023853, 1023954, and 1023596.

5. References

[1] Barnes, J., A. Brugos, N. Veilleux & S. Shattuck-Hufnagel. 2011. Voiceless intervals and perceptual completion in Fo contours: Evidence from scaling perception in American English. *Proceedings of the 17th International Congress of Phonetic Sciences, August 2011, Hong Kong.*

[2] Barnes, J., N. Veilleux, A. Brugos, & S. Shattuck-Hufnagel. 2012. Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Journal of Laboratory Phonology* 3(2): 343-389.

[3] Boersma, P. & D. Weenink. *Praat: doing phonetics by computer*, accessed May 1, 2009. http://www.praat.org.

[4] Bruce, Gösta. 1977. *Swedish word accents in a sentence perspective*. (Travaux de l'Institut de Linguistique de Lund 12.) Lund, Sweden: CWK Gleerup.

[5] Caspers, J. & V. van Heuven. 1993. Effects of time pressure on the phonetic realization of the Dutch accent-lending pitch rise and fall. *Phonetica* 50: 161-171.

[6] d'Alessandro, C. & P. Mertens. 1995. Automatic pitch contour stylization using a model of tonal perception. *Computer Speech and Language* 9: 257-288.

[7] Dogil, G & A. Schweitzer. 2011. Quantal effects in the Temporal Alignment of Prosodic Events. *Proceedings of the 17th International Congress of Phonetic Sciences, August 2011, Hong Kong.*

[8] Flemming, E. 2008. The grammar of coarticulation. To appear in M. Embarki & C. Dodane (eds.), *La Coarticulation: Indices, Direction et Representation*.

[9] Fujisaki, H. & K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)* **5**(4): 233–241.

[10] Goldsmith, J. 1976. Autosegmental phonology. PhD Thesis, MIT.

[11] Gordon, M. 1999. Syllable weight: Phonetics, phonology, typology. PhD Thesis, UCLA.

[12] Hermes, D. 1998. Auditory and visual similarity of pitch contours. *Journal of Speech, Language, and Hearing Research* 41(1): 63-72.

[13] Hermes, D. 2006. Stylization of Pitch Contours. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (eds.), *Methods in Empirical Prosody Research,* Berlin-New York: de Gruyter, 29-62.

[14] Hirst, D. & R. Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de l'Institut de Phonétique d'Aix 15: 71–85, Univ. de Provence.

[15] Hombert, J. M., J. Ohala & W. Ewan. 1979. Phonetic explanations for the development of tones. *Language* 55: 37-58.

[16] House, D. 1990. *Tonal perception in speech*. Lund, Sweden: Lund University Press.

[17] Kohler, K. 1990. Macro and micro F0 in the synthesis of intonation. In J. Kingston & M. Beckman (eds), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech,* Cambridge: CUP, 115-138.

[18] Ladd, D. R., I. Mennen & A. Schepman. 2000. Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America* 107: 2685-2696.

[19] Leben, W. 1973. Suprasegmental phonology. PhD Thesis, MIT.

[20] Mixdorff, H & O. Niebuhr. 2013. The Influence of F0 Contour Continuity on Prominence Perception. *Proceedings of Interspeech 2013, Lyon, France.*

[21] Nooteboom, S. 1997. The prosody of speech: Melody and rhythm. In W. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Science,* Oxford: Blackwell, 640-673.

[22] Pierrehumbert, J. 1980. The Phonetics and Phonology of English Intonation. PhD Thesis, MIT.

[23] Prieto, P. 2009. Tonal alignment patterns in Catalan nuclear falls. *Lingua* 119 (6): 865-880.

[24] Prieto, P. & F. Torreira. 2007. The segmental anchoring hypothesis revisited. Syllable structure and speech rate effects on peak timing in Spanish. *Journal of Phonetics* 35(4): 473-500.

[25] Santen, J. van & J. Hirschberg. 1994. Segmental effects on timing and height of pitch contours. *Proc. ICSLP 94*, Yokohama, 719-722.

[26] Silverman, K. 1987. The structure and processing of fundamental frequency contours. PhD Thesis, University of Cambridge.

[27] Taylor, P. 1998. The Tilt intonation model. In R. Mannell & J. Robert-Ribes, (eds.), *Proc. ICSLP 98*, volume 4, 1383–1386.

[28] Zhang, J. 2001. The Effects of Duration and Sonority on Contour Tone Distribution— Typological Survey and Formal Analysis. PhD Thesis, UCLA.

[29] Zhang, J. 2004. The role of contrast-specific and language-specific phonetics in contour tone distribution. In B. Hayes, R. Kirchner & D. Steriade (eds.), *Phonetically-Based Phonology*, Cambridge: Cambridge University Press, 157-190.

Intonational phonology in Bengali and English infant-directed speech

Kristine M. Yu [1], *Sameer ud Dowla Khan* [2], *Megha Sundara* [3]

[1] Department of Linguistics, University of Massachusetts, Amherst, MA, USA
[2] Department of Linguistics, Reed College, Portland, OR, USA
[3] Department of Linguistics, University of California, Los Angeles, CA, USA
krisyu@linguist.umass.edu, skhan@reed.edu, megha.sundara@humnet.ucla.edu

Abstract

We examined the phonetics and phonology of intonation of infant-directed speech (IDS) and non-IDS in story-reading in two typologically-divergent languages, English and Bengali. In addition to finding an increase in f0 range and variability in IDS, replicating previous work on IDS prosody, we found novel evidence that f0 manipulations in IDS are constrained by intonational phonology. Speakers in both languages used an increased proportion of tonal elements with higher tonal targets and more turning points in IDS, within the language-specific intonational grammar. The tonal elements showing increased use in IDS also were associated with marking topic and focus. Thus, phonetic changes in IDS may in part be induced by speakers' choices of phonological tonal elements, which in turn may be connected with choices about marking discourse structure.

Index Terms: infant-directed speech, intonational phonology, speech style, information structure, Bengali, English

1. Introduction and background

It has long been noted that gradient phonetic variation in prosodic elements such as pitch, rhythm, and duration can carry information about interpersonal interaction and emotional state, cf. [1] for a review. However, work in intonational phonology ([1], [2], i.a.) has shown that prosodic variation is also structured according to the language-specific phonological grammar, which generates well-formed tunes constructed over discrete phonological categories, conveying linguistic meanings such as contrastive focus ([3], i.a.).

A classic example of gradient prosodic manipulation occurs in the speech style of infant-directed speech (IDS) or "motherese": relative to adult-directed speech, IDS has been shown to exhibit higher mean fundamental frequency (f0), higher maximum f0, an expanded pitch range, greater f0 variability, and shorter utterances across a variety of languages ([4], [5], [6], i.a.). With almost no exception, work on IDS prosody has exclusively involved phonetic description of pitch manipulations as described above. Moreover, these gradient phonetic manipulations have been described as serving attentional and emotional functions. For instance, high pitch and increased pitch variability such as bell- and sinusoidal-shaped pitch modulations have been suggested to maintain infant attention and elicit positive emotional rapport ([7], [8], [9], [10], [11], [12], [13]).

Without attention to intonational phonology, phonetic characterizations of IDS prosody leave open the possibility of pitch variation in IDS wholly motivated by attentional and emotional considerations, e.g. with sinusoidal ups and downs over a wide pitch range, irrespective of constraints of phonological grammar. To date, very little in the literature suggests otherwise, with the exception of three studies that explored phonological constraints on pitch variation in IDS. In the tone languages Thai [6] and Mandarin [14], it has been found that f0 cues to lexical tonal identity are preserved in IDS. Only one study thus far ([15]) addresses intonational phonological constraints on IDS; pitch range expansion in Japanese was shown to be restricted to boundary tones, a sub-component of the intonational grammar where pitch is not lexically contrastive. In all three cases, the pitch manipulations ascribed to IDS did not interfere with lexically-contrastive aspects of pitch; the current study extends this recent work to two languages that have no lexical contrasts in pitch but that nevertheless manipulate pitch linguistically as intonation.

Our study investigates the hypothesis that pitch variability in IDS can be constrained by intonational phonological grammar in English and Bengali, languages with typologically divergent prosodic systems. We compared how parents of infants and young children read a story standardized across the languages in adult-directed speech (ADS) vs. IDS. We annotated the readings with intonational transcriptions based on intonational models of lab speech-style ADS in the two languages.

While [15] examined how pitch variation changes in Japanese IDS are conditioned on particular tonal elements, our study examines how pitch variation changes in IDS are realized via *changes in the choice of tonal elements within intonational grammar*. Moreover, our phonological analysis of IDS intonation allows us to examine how prosodic choices in IDS may reflect attention to conveying linguistic meaning and structure, in addition to regulating attention and emotional rapport. To preview our results: based on comparing the distribution of discrete tonal elements within the two speech styles, we found that in both languages, IDS showed an increase in tonal elements with high tonal targets and multiple turning points. These distributional changes occurred within the constraints of each language's intonational grammar. Moreover, these changes also resulted in an increase in the use of tonal elements used to mark aspects of discourse structure.

1.1. English intonation and transcription

The most widely known model of the intonation of American English (henceforth, simply "English") is often referred to as MAE_ToBI, from the Mainstream American English Tones and Break Indices transcription system devised to annotate the intonation of this variety ([1], [2], [3]). In MAE_ToBI, pitch accents are borne by the primary stressed syllable of prominent words, where stress is contrastive and thus must encode lexical information. The pitch accent inventory includes default H* (high) and L* (low) tones as well as rising L+H* (early rise) and L*+H (late rise) tones and one falling H+!H* tone. The choice is related to attitude, focus status, and tonal environment ([3], [16], [17]). The choice of pitch accent is somewhat variable, although certain tune–meaning relationships have been proposed in the literature: L+H* and

L*+H often convey focus ([3]). Boundary tones occur at the right edge of Intonation Phrases (IPs) and intermediate phrases (ip). The IP boundary tone inventory includes L-L% (low falling), H-H% (high rising), L-H% (low rising), H-L% (high level), and !H-L% (downstepped high level), the choice of which conveys sentence type, finality, etc. ([3]).

1.2. Bengali intonation and transcription

The intonational phonological model of Bangladeshi Standard Bengali (henceforth, simply "Bengali") presented here is based on that of [18], [19]. The pitch accent inventory of Bengali includes default L* (low), H* (high), L*+H (rising), as well as f-marked (focus-marking) versions of the latter two, fH* (extra high) and L*+fH (rise to extra high), the choice of which conveys speaker attitude, focus status/type, and tonal environment ([18], [19]). Focus-marking tones are easily distinguished by their unique tonal interactions not seen in other pitch accents (i.e. violation of downtrend, immunity from overriding, triggering of post-focal tone compression); these phonetic and phonological criteria were used in the current study to identify focus-marking tones.

Each pitch accent projects a small prosodic unit called the Accentual Phrase (AP), which is marked on the right edge by a tone of the opposite target of the pitch accent: Ha (high target) and La (low target). These phrases group into ips, marked by sharp final contours: H- (sharp final rise) and L- (sharp final fall). As in English, ips group into IPs, marked by boundary tones that covey sentence type, information structure, and finality ([18], [19]). H% (high rising) and HL% (high falling) can serve as topicalizers, and both LH% (low rising) and HLH% (high dipping) can serve as markers of non-finality.

1.3. Comparison of English and Bengali intonation

Noteworthy similarities and differences between Bengali and English intonation are in (a) pitch contour regularity, (b) focus realization, and (c) boundary tone complexity.

One of the most clearly noticeable differences between English and Bengali is the regular repeating patterns seen in the Bengali pitch contour and the general lack of such regularity in English. This is largely an effect of the restricted pitch accent and AP boundary tone distribution of Bengali. While a content word in English is fairly free to bear any of five basic pitch accents (H*, L*, L+H*, L*+H, H+!H*) or none, Bengali content words almost always bear L* followed by a Ha, except in very specific cases: H* followed by La conveys surprise or sarcasm, and L*+H is only a secondary option phrase-finally, and the two f-marked pitch accents (fH*, L*+fH) are only used to convey narrow focus.

In both languages, focus is realized via pitch accent choice and post-focal tone compression (i.e. deaccenting). In Bengali, an f-marked tone (fH*, L*+fH) marks the focused word. These are sharply distinct from the non-focused pitch accent L*, and can even be distinguished from H* and L*+H as the latter two obey downtrend and do not trigger post-focal compression, unlike f-marked tones. In English, the focused element can bear L+H* and L*+H rather than H* in declaratives [3]. The distinction between tones associated with and without focus in Bengali is clearer than in English, where H* and L+H* appear to have overlapping allophonic variation [20], although they are functionally distinct [21].

Lastly, Bengali and English differ in the complexity of their boundary tones. While Bengali has three levels of tonally-marked prosodic structure (i.e. AP, ip, IP), English has only two (i.e. ip, IP). Bengali APs and ips typically end in a single underlying tone, but the ip tones L- and H- both inherently incorporate a sharp phrase-final rise or fall in pitch. English ip tones L- and H- do not have this sharp phrase-final rise or fall, and are instead realized as "phrase accents", i.e. generally low or generally high pitch across the phrase-final stretch. The five Bengali IP tones are very complex, including one, two, or three tonal targets (e.g. HLH%), all of them involving large changes in pitch. The five English IP tones, on the other hand, can have only one or two tonal targets (e.g. L-H%), and two of them involve a stretch of flat pitch (H-L%, !H-L%), which is at best rare in Bengali.

2. Methods

2.1. Data collection

The speech of nine (5 male, 5 female) native speakers of English and ten (5 male, 5 female) native speakers of Bengali was recorded in a quiet room in Los Angeles. Subjects were paid $5 for their participation, which typically took 20 min. To help ensure that subjects were comfortable and familiar with the IDS task, only parents were recruited. The English speakers had infants of 4.4±0.7 months of age at the time of recording, while there was no specific cutoff for the age of the Bengali speakers' children due to subject pool limitations. However, all Bengali subjects lived with their (grand)children <10 yrs of age, and six subjects were teachers at a Bengali-language weekend school for children.

Subjects were recorded while reading the "North Wind and Sun" fable, the Bengali version of which was taken from [22]. Storytelling was used as a context appropriate for both IDS and non-IDS default read speech. The particular text was chosen for comparison with other work on speech rhythm using the same text, as part of a larger study. Using translations of the same text across both languages ensured that the recordings from the two languages would not be affected by different semantic/pragmatic features triggered by reading different stories. Furthermore, unlike studies that examine spontaneous IDS, the current study kept the text constant to minimize changes in semantics, morphosyntax, and the underlying segments across conditions. This way we could observe how speakers' prosodic choices might change, given the same morphosyntactic and discourse structure.

Multiple recordings were made using a Shure SM10A head-mounted microphone plugged into a laptop computer via a preamplifier. The first of the two conditions was default reading ("non-IDS"), in which subjects were asked to "read at a comfortable pace". This task direction was designed to be comparable with other studies of speech rhythm using lab speech, as part of a larger study. The second condition was simulated infant-directed reading ("IDS"), in which the subject was asked to read the same passage as if speaking to their 4-5 month-old infant. (Bengali speakers were asked to imagine their child at that age.). This is an age range where IDS has fewer single-word utterances, is less dominated by soothing/comforting affect, and is intensely used for rapport and attention ([9], [6]). Thus, it is an age range where paralinguistic demands play a large role in driving prosodic

manipulations in IDS. Childlike illustrations and plush toys were used to help further encourage this register.[1]

2.2. Data analysis

For each condition, the three clearest repetitions with minimal or no disfluencies were analyzed, giving 2 conditions x 3 repetitions = 6 recordings per speaker. Text grids were generated in Praat ([23]) to allow annotation in MAE_ToBI tone labels ([24]) for English and B-ToBI tone labels ([18], [19]) for Bengali, by a transcriber trained in each system.

Both acoustic-phonetic measurements and categorical phonological measurements were taken during each sound file. These include: pitch range, tonal inventory, total number of pitch accents and boundary tones, and the frequency of use of each type of pitch accent and boundary tone. Currently, the results presented reflect the whole Bengali data set, but annotations for the English data set are incomplete and preliminary results presented here reflect one repetition from each speech style for each speaker. For the Bengali data, a random sample of the data has also been transcribed by a second transcriber to check for intertranscriber reliability.

Transcribers in both languages were trained only on non-IDS speech materials from outside of the current study, and were encouraged to propose new tones or take note of new tonal interactions when approaching the IDS and non-IDS recordings from the current study. This was done to discourage transcribers from forcing aspects of the non-IDS intonational phonological grammar onto what could hypothetically be a completely different grammar in IDS. All tones transcribed in IDS were also found in the non-IDS recordings in this study, including two tones in Bengali that had not yet been identified in previous work (and were not found in the training files): L+H*, M%.

To analyze changes between the distribution of pitch accent and boundary tone types between styles, proportions of each tone type were calculated, aggregated across repetitions. Separate mixed effects logistic models were built for individual tonal elements, with the proportion of an individual tonal element as the dependent variable, speech style and speaker sex as fixed effects, and random effects by speaker (cf. [25] for advantages of logistic models for proportional data). Results reported as significant were significant with Bonferroni corrections for multiple comparisons.

3. Results

3.1. Phonetic changes in IDS: both languages

F0 data was extracted at 10ms intervals from the recorded files using RAPT ([26]). Mean, minimum, and maximum f0 (in Hz), as well as f0 range and change in f0 variability were calculated over each phrase and averaged across phrases for each speaker, following [5]. Paired t-tests indicated significant increases in IDS for mean f0, maximum f0, f0 range, and f0 variability, all with $p < 0.01$, but no change in minimum f0, for both languages. These results are consistent with previous studies of IDS in many languages, supporting the validity of properties of the simulated IDS recorded here as bearing on properties of IDS in general.

3.2. Phonological changes in IDS: Bengali

While the number of pitch accents significantly decreased by 7% in IDS ($t(9)=3.1$, $p=0.01$), the proportion of focus-marking pitch accents significantly increased. Proportions of pitch accents between speech styles are shown below in Figure 1. The proportion of fH* accents significantly increased from an average of 0.9% across speakers to 5.0% in IDS ($\beta_{IDS} = 2.1$, $p = 2e\text{-}6$). The proportion of L*+fH accents significantly increased from 1.9% to 4.1%, with a larger increase for females ($\beta_{IDS \text{ nested in } SEX} = 1.0$, $p = 0.007$). The proportion of (default) L* accents significantly decreased, while the proportion of (non-default) H* accents significantly increased.

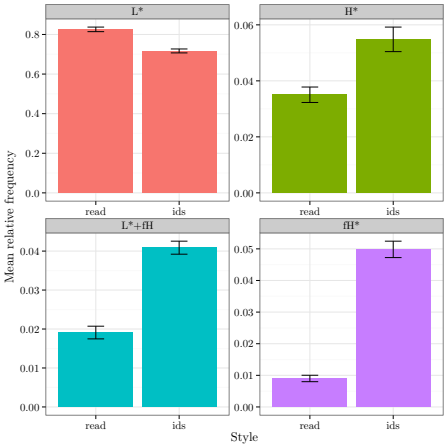


Fig. 1: Change in relative frequency of pitch accents between speech styles, averaged over speakers. Error bars show ±1SE.

In addition, the number of IPs significantly increased by 49% in IDS ($t(9)=5.5$, $p=3.8e\text{-}4$) and the number of ips by 21% ($t(9)=-3.5$, $p=0.006$), while the number of APs significantly decreased by 10% in IDS ($t(9)=-3.5$, $p=6.5e\text{-}3$). Among the AP tone types, there was no significant change in the proportion of La tones, but the proportion of Ha tones significantly decreased in IDS ($\beta_{IDS} = -0.64$, $p = 7e\text{-}8$), while the proportion of fHa tones significantly increased ($\beta_{IDS} = 0.79$, $p = 2e\text{-}8$). Changes in the proportion of IP tone types are shown in Figure 2. The proportion of HLH% tones significantly increased from an average of 8% to 17% in IDS ($\beta_{IDS} = 0.80$, $p = 1.9e\text{-}5$), as did the proportion of HL% tones, from an average of 2% to 7% ($\beta_{IDS} = 1.2$, $p = 3.1e\text{-}4$). In contrast, the proportion of L% and LH% tones significantly decreased ($\beta_{IDS} = -0.31$, $p = 7.2e\text{-}3$. $\beta_{IDS} = -0.58$, $p = 3.9e\text{-}6$).

3.3. Phonological changes in IDS: English

While the number of pitch accents did not significantly change between speech styles, ($t(8)=-1.8$, $p=0.11$), there was a significant increase in IDS in the proportion of rising pitch accents of L*+H from 0.7% to 4.9% on average across speakers, and of L+H* from 2.8% to 10.3% on average (L*+H: $\beta_{IDS} = 1.9$, $p = 0.003$; L+H*: $\beta_{IDS} = 1.44$, $p = 4.17e\text{-}5$). There was a concomitant insignificant trend for decreases in IDS in the proportion of L* and H* accents. In addition, speakers produced a significantly higher average of 32% more

[1] In pilot work when an infant was in the room, the infant would often fuss and cry, disturbing the audio recording, causing disfluencies in the subject's speech, and interrupting him/her at unpredictable intervals. As the study of intonational phonology in IDS is in its beginning stages, we decided to abstract away from these disturbances for the time being.

ips and 35% more IPs in IDS (ip: t(8) = 3.4, p = 0.01, IP: t(8) = 4.8, p = 0.001). There were no significant changes in the proportion of different boundary tone sequences, and the majority of boundary tone sequences were L-L% (85%).

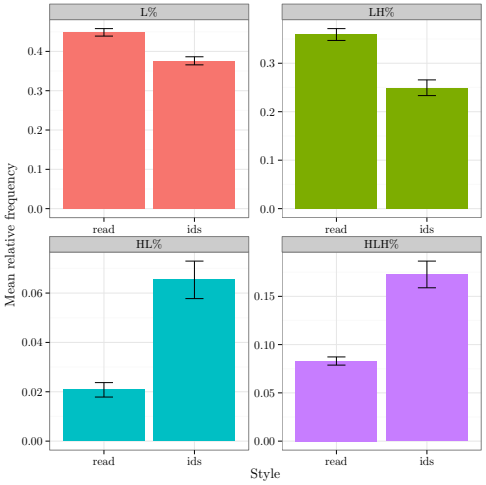


Figure 2: Change in relative frequency of IP tones between speech styles averaged over speakers. Error bars show ±1SE.

3.4. Summary of results

To summarize, both languages showed an overall acoustic-phonetic expansion of the pitch range and overall increase in f0 variability. Both languages also had phonological changes in intonation: increases in the use of non-default pitch accents (despite the lack of increase in pitch accents overall), increases in the number of ips and IPs, and for Bengali, increases in the use of some complex boundary tones.

4. Discussion

We found that IDS prosody in a storytelling context was characterized by gradient phonetic changes as well as changes in the choice of discrete tonal elements within the language-specific intonational grammar.

The overall expansion of f0 range, higher f0 maxima and increased f0 variability in IDS was realized within the constraints of the language-specific grammar. First, we found that pitch variation in IDS could be understood using phonological models of intonation: all tones we found in IDS were also found in non-IDS. Second, changes in the choice of tonal elements in IDS were changes that could induce the gradient phonetic changes of higher f0 maxima and increased f0 variability. Both languages showed an increase in pitch accents with higher tonal targets and more turning points in IDS (L+H* and L*+H in English; H*, fH* and L*+fH in Bengali); Bengali also had an increase in use of an AP tone with a higher tonal target, fHa. In addition, both languages showed increases in the number of higher-level boundary tones. Higher-level boundary tones such as IP tones are realized in an expanded pitch range relative to lower-level boundary tones. Moreover, in Bengali, the inventory of IP tones includes complex contours not available for lower-level boundary tones; indeed, there was a decrease in default L% tones but an increase in HL% and HLH% tones in IDS. Thus, while non-IDS in Bengali was characterized by a series of predictable contours built with L*, Ha, and L%, IDS replaced these with H*/fH* and complex boundary tones.

In addition, the change in choice of tonal elements in IDS appears to not only be influenced by the pitch patterns they induce, but also by their tone-meaning mappings. Almost all the pitch accents with increased use in IDS in both languages are ones associated with focus, and Bengali also showed an increase in use of the focus-marking fHa AP tone and the topicalization marker HL%. The lack of an increase in the number of pitch accents in both languages can also be understood as a consequence of focus-marking within intonational grammar: preliminary analyses suggest that the increased use of focus-marking pitch accents in both English and Bengali in IDS limited the increase in the number of pitch accents overall since the use of focus-marking pitch accents causes post-focal tonal compression and deaccenting in both Bengali and English. Preliminary inspection also suggests that speakers were not deploying tones associated with focus- and topic-marking in IDS in a manner divorced from the discourse structure, which is what one might expect if these tones were chosen simply due to the pitch patterns they induce. For instance, L+H* accents in English did not appear on pronouns, which were discourse-given. Thus, we speculate that the phonetic pitch changes in IDS may be in part a consequence of increased marking of discourse structure in IDS, as suggested in [10].

5. Conclusions

In this study, we examined changes in patterns of pitch variation in English and Bengali when speakers read a story in a non-infant directed style and an infant-directed style. We found that in both languages, there was an increase in f0 range via raising of f0 maxima as well as increased f0 variability in IDS, replicating previous cross-linguistic phonetic work on IDS prosody. In addition, we performed phonological analyses of intonation in the two speech styles, which revealed that the phonetic pitch manipulations in IDS we observed were constrained by language-specific intonational grammar. We observed that IDS and non-IDS intonation could be modeled with the same underlying tonal elements, but that the distribution of use of these elements differed between styles. In IDS, speakers increased the proportion of tonal elements with high tonal targets and more turning points, which induce higher f0 maxima and greater f0 variability. Thus, the phonetic changes in pitch in IDS may have been in part a consequence of speakers' choice of phonological tonal elements. Moreover, tonal elements with increased use in IDS were also associated with marking focus and topic in previous work. Preliminary analyses suggest that these tonal elements were not placed arbitrarily, but in ways consistent with them being used to mark discourse structure. Thus, phonetic changes in pitch in storytelling IDS may also be in part due to increased marking of discourse structure. In further work, we are continuing to analyze the role of discourse and syntactic structure in the prosodic choices made by the speakers, and we are performing phonetic analyses of f0 assessing the contribution of different tonal elements to increasing f0 range and variability.

6. Acknowledgements

Special thanks to Alejna Brugos, J'aime Panna Roemer and Megan Keough for help in recording and annotating the sound files.

7. References

[1] Ladd, D. Robert. *Intonational phonology*, Cambridge: Cambridge University Press, 1996.

[2] Pierrehumbert, J.B. "The phonology and phonetics of English intonation", Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1980.

[3] Pierrehumbert, J. and Hirschberg, J. "The meaning of intonational contours in the interpretation of discourse", in P. Cohen, J. Morgan, and M. Pollack [Ed], Intentions in Communication, 271-311, 1990.

[4] Fernald, A. and Simon, T. "Expanded intonation contours in mothers' speech to newborns," Dev. Psych., 20(1): 104-113, 1984.

[5] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., and Fukui, I. "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants", J. Child Lang., 16: 477-501, 1989.

[6] Kitamura, C., Thanavishuth, C., Burnham, D. and Luksaneeyanawin, S. "Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language", Inf. Behav. and Dev., 24(4): 372-392, 2002.

[7] Fernald, A. & Kuhl, P. "Acoustic determinants of infant preference for motherese speech", Inf. Behav. and Dev., 10: 279-293, 1987.

[8] Stern, D.N., Spieker, S., Barnett, R. K., and MacKain, K. "The prosody of maternal speech: infant age and context related changes," J. Child Lang., 10: 1-15, 1983.

[9] Stern, D.N., Spieker S., and MacKain, K. "Intonation contours as signals in maternal speech to prelinguistic infants", Dev. Psych., 18(5): 727-735, 1982.

[10] Fernald A., and Mazzie, C. "Prosody and focus in speech to infants and adults", Dev. Psych., 27(2): 209-221, 1991.

[11] Werker, J.F. and McLeod, P.J. "Infant preference for both male and female infant-directed talk: a developmental study of attentional and affective responsiveness", Can. J. Psychol., 43(2): 230—246, 1989.

[12] Fernald, A. "Intonation and communicative intent in mothers' speech to infants: is the melody the message?", Child Dev., 60: 1497-1510, 1989.

[13] Trainor, L.J., Austin, C.M., and Desjardins, R.N. "Is infant-directed speech prosody a result of the vocal expression of emotion", Psych. Sci., 11(3): 188-195, 2000.

[14] Liu, H., Tsao, F., and Kuhl, P.K. "Acoustic analysis of lexical tone in Mandarin infant-directed speech", Dev. Psych., 43(4): 912-917, 2007.

[15] Igarashi, Y., Nishikawa K., Tanaka K., and Mazuka R., "Phonological theory informs the analysis of intonational exaggeration in Japanese infant-directed speech", J. of the Acoustical Soc. of America, 134(2): 1283-1294, 2013.

[16] Dainora, A. "Does intonational meaning come from tones or tunes? Evidence against a compositional approach", Speech Prosody 2002, 2002.

[17] Dainora, A. "Modeling intonation in English: a probabilistic approach to phonological competence", in Goldstein, L., Whalen, D.H., Best, Catherine T. [Ed], Laboratory Phonology 8, Mouton de Gruyter, 2006.

[18] Khan, S.D. "Intonational phonology and focus prosody of Bengali", Ph.D. dissertation, University of California, Los Angeles, CA, 2008.

[19] Khan, S.D. "The intonational phonology of Bangladeshi Standard Bengali", in Jun, S.-A. [Ed], Prosodic Typology II: The Phonology of Intonation and Phrasing, Oxford Univ. Press, 2014 in press.

[20] Bartels, C. and Kingston, J. "Salient pitch cues in the perception of contrastive focus," in Bosch, P. and van der Sandt, R. [Ed], Focus and natural language processing, IBM Deutschland. 1994

[21] Ito, K., Bibyk, S.A., Wagner, L., and Speer, S.S. "Interpretation of contrastive pitch accent in six to eleven-year-old English-speaking children (and adults)", J. Child Lang., 41(1): 84-110, 2014.

[22] Khan, S.D. "Bengali (Bangladeshi Standard)", JIPA, 40(2), 221-225, 2010.

[23] Boersma, P. "Praat, a system for doing phonetics by computer", Glot International, 5(9/10): 341-345, 2001.

[24] Beckman, M. and Ayers, G. Guidelines for ToBI labeling, 1997, http://www.ling.ohio-state.edu//~tobi/ame_tobi/labelling_guide_v3.pdf.

[25] Jaeger, F. "Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models", J. Mem. & Lang., 59(4): 434-446.

[26] Talkin, D. "A robust algorithm for pitch tracking (RAPT)", in Kleijn, W. B. and Paliwal, K. K. [Ed], Speech coding and synthesis, Elsevier, 1995.

Hemispheric lateralization of sentence intonation in left handed subjects with typical and atypical language lateralization: an fMRI study

Eszter Varga[5], Zsuzsanna Schnell[7], Gábor Perlaki[1,2,3], Gergely Orsi[1,2,3], Mihály Aradi[2], Tibor Auer[4], Flóra John[1], Tamás Dóczi[3,6], Sámuel Komoly[1], Norbert Kovács[1], Attila Schwarcz[3,6], ,Tamás Tényi[5], Róbert Herold[5], József Janszky[1,3], Réka Horváth[1]

[1]Department of Neurology, University of Pécs, Pécs, Hungary
[2]Pécs Diagnostic Centre, Pécs, Hungary
[3]MTA-PTE Clinical Neuroscience MR Research Group, Pécs, Hungary
[4]MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom
[5]Department of Psychiatry and Psychotherapy, University of Pécs, Pécs, Hungary
[6]Department of Neurosurgery, University of Pecs, Pécs, Hungary
[7]Department of Linguistics and Institute of Psychology, University of Pécs, Hungary
eszter.varga@kk.pte.hu

Abstract

Prosody (as the melody of speech) is an important component of human social interactions. More specifically, linguistic prosody conveys meaning of speech through syllable, word, or sentence level stress and intonation. In the modern neuroimaging era the hemispheric representation of sentence intonation is widely investigated. Most of these studies suggest bilateral activations predominantly in the perisylvian language areas and in the subdominant homologues. However, there are some inconsistencies about the hemispheric representation and lateralization of linguistic prosody. These inconsistencies could be due to the lack of attention on the language lateralization of the subjects. The present study aims to investigate the hemispheric representation and lateralization of linguistic prosody with a sentence intonation task in two groups of left handed subjects with typical and atypical language lateralization. Functional MRI was used to test the assumption that - according to the functional lateralization hypothesis - the representation of sentence intonation is predominantly lateralized within the language dominant hemisphere and the lateralization of sentence intonation is associated with language lateralization in both groups.

Left handers were examined to create two groups of subjects with typical and atypical language lateralization. In all, 32 healthy subjects were evaluated with a standard verbal fluency task with fMRI in order to assess functional hemispheric language lateralization. In our final investigation the atypical group consisted of 8 subjects with right hemispheric language dominance (LI<-0.2) and the typical group also consisted of 8 subjects with left hemispheric language dominance (LI>0.2).

Sentence intonation task was utilized to test linguistic prosody skills with fMRI. 49 pairs of sentences (18 pairs of neutral-neutral sentences, 10 pairs of interrogative-interrogative sentences, and 1 pair of interrogative-neutral sentence) were presented with an event-related design. Sentences were matched in terms of syntactic structure, semantic complexity and length and all were affectively neutral. In the fMRI data analysis interrogative pairs were compared to neutral pairs.

One of the main findings of our study is that subjects with typical language lateralization activated the middle temporal gyrus (MTG) on the right side. The activation of the MTG in the right hemisphere is classically associated with the encoding of prosodic information. Furthermore, both groups recruited the frontal and temporal language areas predominantly in the language-dominant hemisphere. Moreover, between-group comparison showed significantly stronger activations in subjects with typical language lateralization only in left sided language areas: pars triangularis of the inferior frontal gyrus, the superior frontal gyrus and the inferior parietal lobule.

This finding is in accordance with the functional lateralization hypothesis of prosody, and suggests a correlation between linguistic prosody lateralization and language lateralization.

Index Terms: sentence intonation, language lateralization, left-handers, fMRI

Introduction

Prosody is an important component of everyday discourse and thus, of social interaction. It conveys different pieces of linguistic information at the word and sentence level (also known as linguistic prosody), and expresses information about the speaker's emotional state (called emotional prosody) [1,2]. More specifically, linguistic prosody conveys meaning through syllable, word, or sentence level stress and intonation. Currently, there are four general hypotheses about the hemispheric localization of speech prosody. (1) Acoustic lateralization hypothesis posits that both linguistic and emotional prosody are processed in the right hemisphere through the extraction of suprasegmental information [3]. (2) Functional lateralization hypothesis proposes that linguistic prosody, like word stress, sentence focus and sentence modus, engages neural mechanisms in the left hemisphere, while emotional prosodic information (such as emotions and personality) are encoded in the right hemisphere [2,4]. (3) Several findings claim that the lateralization of speech prosody and the division of labor of the two hemispheres depend on different acoustic cues, such as temporal vs. spectral cues, high frequency vs. low frequency cues and rapidly changing vs. slowly changing acoustic cues [5,6]. (4) However, some clinical studies posit that prosody processing is not lateralized at all, but is rather subserved by subcortical regions [7,8].

Relatively few fMRI studies have been published that specifically focus on prosody based language processing [6,9,10,11] and hence, there are still some inconsistencies about the hemispheric representation and lateralization of subjects. Most of these fMRI studies suggest bilateral

activations predominantly in the perisylvian language areas and in the subdominant homologues. The mentioned inconsistencies could be due to the fact that these studies did not take into consideration the language lateralization of the examined subjects.

The present study aims to investigate the hemispheric representation and lateralization of linguistic prosody with a sentence intonation task in two groups of left-handed subjects with typical and atypical language lateralization. Functional MRI was used to test the assumption – held by the functional lateralization hypothesis - that the representation of sentence intonation is predominantly lateralized within the language dominant hemisphere and thus the lateralization of sentence intonation is associated with language lateralization in both groups.

Methods

2.1. Verbal fluency task

Left-handers were examined to create two groups of subjects with typical and atypical language lateralization. In the modern neuroimaging era it is believed that among left handers the incidence of atypical language lateralization is higher (15-30%) than in right handers (4-6%) [12], hence, in the present study, 32 healthy, left-handed subjects were evaluated with a standard verbal fluency task with fMRI in order to assess functional hemispheric language lateralization. The paradigm included seven cycles of 30-second-long rests alternating with 30-second-long internal word generation tasks. During the active conditions, the subjects were asked to silently generate different words starting with a particular letter. During the rest periods, the subjects were instructed to stop the active task and relax. In our final investigation the atypical group consisted of 8 subjects with right hemispheric language dominance (LI (lateralization index)<-0.2), while the typical group, also formed by 8 subjects, was characterized by left hemispheric language dominance (LI>0.2).

2.2. Linguistic prosody task:

A sentence intonation task was utilized to test linguistic prosody skills with fMRI. 33 pairs of sentences (see examples) were presented by a professional actress with an appropriate intonation: 18 pairs of neutral-neutral sentences, 10 pairs of interrogative-interrogative sentences, and 5 pairs of interrogative-neutral sentences. As we can see in the examples below, in Hungarian each sentence pair was based on the same syntactic structure with either the same (interrogative-interrogative pairs and neutral-neutral pairs) or different (neutral-interrogative pairs) intonation. (The neutral-interrogative pairs were not used for contrasts in the fMRI analysis, they were used to keep subjects' attention during the paradigm instead). In other words, in one sentence pair, neutral and interrogative sentences were syntactically exactly the same, only sentence intonation differentiated the semantic meaning. Since Hungarian uses variations in pitch to signal different sentence intonations, the syntactic form of questions and statements are identical, and there is no transformation from one to another as in English. (In this study, neutral sentences with monotonic intonation could be statements from a syntactic viewpoint, only intonation differentiates them from questions.) It is important to note that in Hungarian the fundamental frequency of interrogative sentences appears to start high and generally move upward through the duration of the sentence. Besides, the fundamental frequency of the used neutral sentences of monotonous intonation in our design starts much lower and stays flat through the whole sentence. Also, sentences were strictly matched in terms of syntactic structure, semantic complexity and length, which is important from a methodological point of view, in order to make sure, that the neutral and interrogative sentences differ only in prosody. All of the pairs were presented in a randomized order, which was the same across subjects. We used an event-related design. Between sentence pairs an inter-trial interval of 4-6 s (jittered) was used. Participants were asked to press an answer button when the two sentences sounded different, e.g.in the case of neutral-interrogative sentence pairs. As mentioned earlier, the neutral-interrogative pairs were not used for contrast in the fMRI analysis, they were only used to keep subjects' attention focused during the paradigm.

2.3. Functional data analysis

Functional data sets were analyzed using FSL 4.1.3. In the fMRI data analysis of the linguistic prosody task the question "Where is the response to the interrogative sentences greater than the response to the neutral sentences?" was asked by defining contrast of regressors: interrogative pairs compared to neutral pairs (interrogative>neutral). Since we hypothesized that the neutral and the interrogative sentences differ only in sentence intonation, analyzing the above mentioned contrast shows neural circuitry underlying pitch perception associated with sentence intonation. LIs were calculated using the LI toolbox available as part of the SPM8. Because most areas of the frontal cortex are activated during verbal fluency task, language lateralization analysis was focused on the frontal lobe. Language dominance was classified as left hemispheric (LI>0.2), bilateral (-0.2≤LI≤0.2) or right-sided (LI<-0.2).

2.4. Examples

Neutral-neutral pair:

János a könyvtárban tanul./John is studying in the library.
János a könyvtárban tanul./John is studying in the library.
Interrogative-interrogative pair:

Péter a konyhába megy?/Is Péter going to the kitchen?
Péter a konyhába megy?/Is Péter going to the kitchen?
Neutral-interrogative pair:

Mária a szobában sír./Mary is crying in the room.
Mária a szobában sír?/Is Mary crying in the room?*
*In the original Hungarian examples no structural change signals sentence type, only prosody and intonation distinguishes declarative and interrogative sentences.

Discussion

As far as we know, this is the first fMRI study investigating the relationship of hemispheric representation and functional lateralization of sentence intonation with both atypical and typical language lateralization in a preliminary sample of healthy, left-handed subjects. During the sequential analysis of fMRI data, we found that the two groups had markedly distinct neural activation patterns. The 'atypical group' (AG) – in the interrogative versus neutral sentence contrast – recruited language related brain areas only in the right hemisphere, such as the posterior division of the right middle temporal cortex and the right anterior paracingulate gyrus. On the other hand the 'typical group' (TG) activated language related areas mainly in the left hemisphere, such as the left frontal operculum reaching the left middle temporal gyrus, the left superior temporal gyrus, the left inferior parietal lobule and the pars triangularis of the left inferior frontal gyrus. Moreover, the left superior frontal gyrus, the left gyrus lingualis, the left thalamus was also recruited by the TG. Even more, the TG activated brain regions in the right hemisphere as well, such as the posterior division of the middle temporal gyrus and the insula. Between-group comparison showed significantly stronger activations in subjects with typical language lateralization than in subjects with atypical language

lateralization only in left sided language areas: the pars triangularis of the inferior frontal gyrus, the superior frontal gyrus and the inferior parietal lobule. This finding is consistent with our previous data [13], which found that a reduced microstructural integrity of the left hemisphere was associated with atypical language lateralization.

Results in the sentence intonation processing task revealed activations mostly in the dominant hemisphere in both groups. Interestingly, the AG showed a much poorer activation pattern than the TG. Considering the AG poorer activation pattern, we can speculate that this finding may potentially be the result of a neurodevelopmental disorder behind atypical language lateralization [14,15]. In contrast, the typical group recruited a more widespread brain activation network including language areas in the dominant hemisphere and also in the right (subdominant) middle temporal gyrus - which is classically associated with the encoding of prosodic information - and in the right insula.

Like most of the higher-order cognitive functions, it is very likely that linguistic prosody is processed within a neural network mainly in the frontal, temporal and parietal cortex [6,9,2,11]. Our fMRI data demonstrate that during the sentence intonation task the TG recruited the fronto-temporo-parietal activation network predominantly in the language-dominant hemisphere. This predominantly dominant hemispheric activation pattern is supposedly due to the processing of the modality (i.e. the semantic meaning) of the interrogative sentences. This finding is in line with the functional lateralization hypothesis which proposes that linguistic prosody is processed within the language-dominant hemisphere. Besides, for the processing of non-linguistic acoustic signals subdominant temporal activations were also observed in the TG.

Conclusions

Our finding is in harmony with the functional lateralization hypothesis, since we have found that the representation of sentence intonation is predominantly lateralized within the language-dominant hemisphere in both investigated groups.

Acknowledgements

This research was supported by the National Brain Research Program (NAP) KTIA_NAP_13_1_2013_001 Grant.

Schnell, Zs. was supported by TÁMOP 4.2.4. A/2-11-1-2012-0001 „National Excellence Program – Elaborating and operating an inland student and researcher personal support system convergence program" The project was subsidized by the European Union and co-financed by the European Social Fund.

Schnell, Zs. was supported by SROP-4.2.2.C-11/1/KONV-2012-0005 (Well-Being in the Information Society project) based on the operation of the Theoretical, Computational and Cognitive Linguistics Research Team (ReALIS) at the University of Pécs, Department of Linguistics.

References

[1] H. Ackermann., I. Hertrich. and W. Ziegler. “Prosodische Störungenbei neurologischen Erkrankungen: Eine Literaturübersicht.” *Fortschritten der Neurologie und Psychiatrie*, vol. 61, pp. 241-253, 1993.

[2] D. Wildgruber, H. Ackermann, B. Kreifelts and T. Ethofer. “Cerebral processing of linguistic and emotional prosody: fMRI studies.” *Progress in Brain Research*, vol. 156, pp. 249-268, 2006.

[3] R. Ivry and L. Robertson. “The two sides of perception.” Cambridge, MA: MIT Press, 1998.

[4] S. Charbonneau, B.P. Scherzer, D. Aspirot and H Cohen. “Perception and production of facial and prosodic emotions by chronic CVA patients.” *Neuropsychologia*, vol. 41, pp. 605-613, 2003.

[5] D. Poeppel. “Pure world deefness and the bilateral processing of the speech code.” *Cognitive Science*, vol. 25, pp. 679-693, 2001.

[6] J. Gandour, M. Dzemidzic, D. Wong, M. Lowe, Y. Tong, L. Hsieh, N. Satthamnuwong and J. Lurito. “Temporal intergration of speech prosody is shaped by language experience: An fMRI study.” *Brain and Language*, vol. 84, pp. 318-336, 2003.

[7] A. Cancelliere and A. Kertesz. “Lesion localization in acquired deficits of emotional expression and comprehension.” *Brain and Cognition*, vol. 13, pp. 133-147, 1990.

[8] M.D. Pell and L.C. Leonard. “Processing emotional tone from speech in parkinson disease: role for the basal ganglions.” *Cognitive, Affective, & Behavioral Neuroscinece*, vol. 3, pp. 275-288, 2003.

[9] Y. Tong, J. Gandour, T. Talavage, D. Wong, M. Dzemidzic, Y. Xu, X. Li and M. Lowe. “Neural circuitry underlying sentence-level linguistic prosody.” *Neuroimage*, vol. 28, pp. 417-428, 2005.

[10] D. Wildgruber, I. Hertrich, A. Riecker, M. Erb, S. Anders, W. Grodd and H. Ackermann. “Distinct frontal regions subserve evaluation of linguistic and affective aspects of intonation.” *Cerbral Cortex*, vol. 14, pp. 1384-1389, 2004.

[11] L. Aziz-Zadeh, T. Sheng and A. Gheytanchi. “Common premotor regions for the perception and production of prosody and correlations with empathic and prosodic ability.” *Plos One*, vol. 5, 2010.

[12] S. Knecht, M. Deppe and B. Dräger. “Language lateralization in healthy right-handedness.” *Brain*, vol. 123, pp. 74-81, 2000.

[13] G. Peraki, R. Horvath, G. Orsi and M. Aradi. “White-matter microstructure and language lateralization in left-handers: a whole-brain MRI analysis.” *Brain and Cognition*, vol. 82, pp. 319-328, 2013.

[14] Schaafsma S.M., Riedstra B.J., Pfannkuche K.A., Bouma A., Groothuis T.G.G. “Epigenesis of behavioural lateralization in humans and other animals.” *Philosophical Transactions of the Royal Society B*, vol. 364, pp. 915-927, 2009.

[15] J.P. Szaflarski ,J.R. Binder, E.T. Possing, K.A. McKiernan, B.D. Ward, T.A. Hammeke. “Language lateralization in left-handed and ambidextrous people.” *Neurology*, vol. 2, pp. 238-244, 2002.

The acquisition of multimodal cues to disbelief

Meghan E. Armstrong[1], *Núria Esteve-Gibert*[2], *Pilar Prieto*[3,2]

[1] Department of Languages, Literatures and Cultures, University of Massachusetts, Amherst, Massachusetts, USA
[2]Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra, Barcelona, Spain
[3]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain
armstrong@umass.edu, nuria.esteve@upf.edu, pilar.prieto@upf.edu

Abstract

In this study, we examine how 3-, 4-, and 5-year-old Catalan-acquiring children are able to make use of audio (intonational) and visual (facial gesture) modalities in the comprehension of speaker disbelief, as well as the role of the child's developing Theory of Mind. Our results suggest that in this case, facial gesture provides children with scaffolding for linguistic meaning, and that explicit belief-reasoning also helps children to infer speaker disbelief. We discuss the implications of these findings for the study of intonational development.

Index Terms: intonational development, multimodal comprehension, acquisition, prosodic meaning

1. Introduction

When inferring meaning, adults may rely on different modalities (auditory and/or visual) in different ways. [1] showed, for example, that when incredulity or *disbelief* meaning is marked through intonation and facial gesture in polar questions, Catalan listeners rely more on the visual modality than Dutch listeners, since Dutch encodes incredulity intonationally in a way that is quite different from other types of questions. Children, then, must learn to make use of both the audio and visual modalities to guide them to meaning. Often times the information from the visual modality *reinforces* the message from the audio modality. [2] found that when preschoolers and kindergartners were faced with the task of comprehending a syntactically complex message, reinforcing gestures facilitated comprehension for preschoolers, but not for kindergartners. They suggested that reinforcing gestures serve as "scaffolding", by guiding the child toward the intended meaning of the utterance. They also pointed out that when a message is conveyed through two modalities (e.g. audio + visual), younger children might disregard one channel altogether for working memory reasons. Older children, on the other hand, are more capable of integrating both modalities. While [2] investigated manual gestures, to our knowledge no studies have investigated the reinforcing nature of facial gestures with respect to speech comprehension in children.

One limitation in the literature is the precision with which the audio modality is described, especially with respect to the prosodic cues that guide listeners to meaning, and that children learn to attend to. Most recently, [3] refer to children's understanding of "emotional cues in the voice", but describe their materials as neutral versus "affectively inflected stimuli" with no acoustic characterization of the prosodic differences between the different emotions. Such descriptions do not allow us to understand the types of prosodic cues that children learn to attend to. Recent work has applied the Autosegmental Metrical framework [4,5] in order to investigate children's comprehension of the intonational aspect of the audio modality. [6] showed that 4-, 5- and 6-year-old children performed at above-chance levels in a linguistic comprehension task where children had to identify a "disbelieving" speaker based on differences in the ¡H* L% and L* HL% nuclear configurations in Puerto Rican Spanish [7]. 6-year-olds, however, significantly outperformed the 4- and 5-year-olds. Therefore, in the absence of visual information, children used linguistic information to perceive belief-state meaning. Linguistic meaning associated with belief states is of particular interest since it involves Theory of Mind (ToM) reasoning [8]. That is, in order to fully comprehend linguistic forms that encode information about speaker and hearer belief states, children must have some ability to "mind-read", i.e. infer the belief states of others. One measure of ToM in children is the false belief task [9, 10]. In recent work assessing ToM and emotion understanding, [11] pointed out that awareness of false belief is needed to understand the human state of surprise since one needs to recognize that something must have contradicted the beliefs of a speaker. They compared children's (ages 3-5) scores on a facial expression task where the child had to select a target label about how a person in a picture felt, based on his/her facial expression. They used a battery of ToM tasks, and found a relationship between belief-based emotion labeling such as surprise, and ToM. The children found the false belief tasks in their study easier than labeling facial gestures of surprise or fear. They claim that children exhibit the ability to pass explicit false belief tasks before they are able to label belief-based facial gestures.

In this study we sought to understand the role that facial gestures might play in children's belief state comprehension, specifically how they use audio, visual or audiovisual cues in the comprehension of a speaker's state of *disbelief.* If facial gesture provides scaffolding to linguistic meaning, there should be an ordered path of acquisition - children should be more successful at comprehending disbelief meaning from facial gesture cues than they are for audio cues. Audiovisual cues could, on the one hand, be more useful than audio only cues because of the presence of the facial gesture, but might also be more difficult since the child must integrate the two modalities. Thus it is possible that children use different strategies for the audiovisual condition. When no visual information is available at all (i.e. audio-only modality), we hypothesize that comprehension of belief states should be more difficult for younger children. If children do not have access to the "scaffolding" they use for meaning, they might simply not have access to the meaning. Further, we hypothesize that children with explicit false belief reasoning should be more successful at the task, regardless of the condition. We tested these hypotheses using the tasks outlined below.

2. Methods

2.1.Participants

Seventy-seven Central Catalan-speaking children participated in the experiment, which consisted of two tasks: a ToM false belief task and a comprehension task. The age range was between 34 and 75 months (M = 53.6 mo.). For the comprehension task, there were three conditions, with a between subjects design: Audio Only (AO), Visual Only (VO) and Audiovisual (AV). A total of twenty-six children received the AO condition: one 2-year-old (35 mo.), six 3-year-olds (range 40-44 mo., M=41 mo.), ten 4-year-olds (range 49-59 months, M=53.5 mo.), six 5-year-olds (range 62-68 mo., M=64 mo.) and three 6-year-olds (range 73-75 mo., M=74 mo.). Twenty-three children received the VO condition: two 2-year-olds (both 34 mo.), six 3-year-olds (range 36-47 mo., M=41.5 mo.), six 4-year-olds (range 52-58 mo., M=54.7 mo.), nine 5-year-olds (range 60-71 mo., M=68 mo.) and two 6-year-olds (range 74-75 mo., M=74.5 mo.). Finally, twenty-eight children received the AV condition: one 2-year-old (34 mo.), eight 3-year-olds (range 37-47 mo., M=41.8 mo.), seven 4-year-olds (range 48-59 mo., M=47.8 mo.), eleven 5-year-olds (range 60-71 mo., M=65.4 mo.) and one six-year-old (72 mo.). The participants were all students at Catalan public elementary schools where Catalan was the primary language of instruction.

2.2.Materials

2.2.1. False belief task

The false belief task was an adaptation of the Sally Ann task [8], a classic ToM task. The materials for this task consisted of a short video (0:54) featuring two puppets. At the beginning of the video a princess puppet appears with a ball, announcing that she will leave her ball in one of two containers in front of her. She leaves the ball in one of the containers and subsequently states that she will leave for school. The princess then leaves for school, disappearing from the video. While the princess is gone, a lion puppet appears. The lion takes the ball out of the container and transfers it into the other container in the scene, covering it so that the ball cannot be seen. The lion laughs in a sneaky way and leaves. The princess puppet then appears, announcing that she has returned from school.

2.2.2. Comprehension task

For all three conditions mentioned above, the materials consisted of a Powerpoint presentation containing AO, VO or AV materials. The premise of the task was such that the child had to decide which member from a set of twins did not believe his or her friend about an animal that the friend claimed to have seen while on vacation. Thus, for each slide (which constituted one trial), participants saw four images: the two twins (upper left and lower left in Fig. 1), the twins' friend (lower right) and an image of the animal the friend claimed to have seen (upper right). The images seen in the upper left and lower left regions of Figure 1 were either a female child or a male child, depending on the block of presentation (2 blocks per participant).

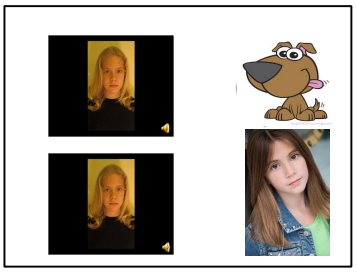


Figure 1. *Slide from AO condition (neutral faces) depicting the two twins (left panels), their friend (lower right panel) and the animal the friend claimed to have seen (upper right panel)*

2.2.2.1. Stimuli

The stimuli for the AV condition were created first. Two child actors, a female (11 years old) and a male (13 years old) were video recorded producing two types of echo questions: neutral echo questions and disbelief echo questions. While there can be variation in terms of the specific pitch contour that might appear for these contexts, the actors produced two different contours labeled L+¡H* L% (neutral echo question) and L* H% (echo question with disbelief marking) in the Cat_ToBI system [12]. (1) and (2) show examples of neutral versus disbelief echo questions:

(1) A and B are talking about what time to leave in the morning in a noisy restaurant. B strains to hear what A has said.

A: I think we should be on the road by eleven.

B: **Eleven**?

(2) B knows A has been a vegetarian for the last ten years.

A: I had the best filet mignon last night. It was cooked to perfection.

B: **Filet mignon!?!** When did you start eating meat?

In (1) B simply repeats an element from A's prior turn and does not convey any type of attitudinal information towards the proposition. In (2) B produces an echo question that repeats linguistic information from the prior discourse, and at the same time expresses her belief state about the propositional content – that she can't believe it.

The intonation contours used for the neutral vs. disbelief echo questions are presented in Figures 2 and 3. Each stimulus was phonetically analyzed in Praat to confirm that the appropriate contour was produced by the actors.

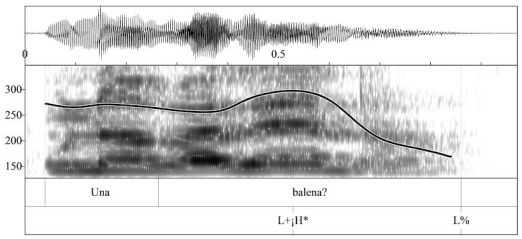


Figure 2: *Pitch track, spectrogram and waveform for the neutral echo question* ***Una balena?*** *'A whale?'*

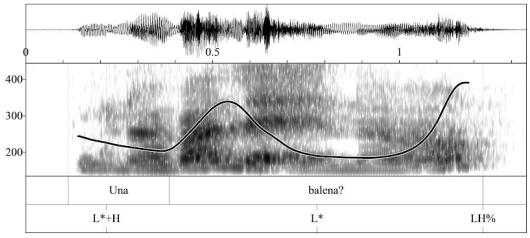


Figure 3: *Pitch track, spectrogram and waveform for the disbelief echo question* ***Una balena?!*** *'A whale?!*

For the AO condition, the audio was extracted from the original AV videos. For the VO condition, the audio was removed so that only the visual information was available. Figures 4 and 5 show typical facial gestures that were presented for the AV and VO stimuli: brow raising and eye-widening for the neutral echo questions, and brow furrowing accompanied with a backwards movement of the head for the disbelief echo questions. Children that received the AO condition saw two identical neutral faces (as shown in Figure 1) presented as screenshots (as in Figure 1).

Figure 4: *Screenshots of facial expressions for neutral echo (left panel) and disbelieving echo (right panel)*

2.3 Procedure

The experimenter was a female native speaker of Central Catalan (the second author of this paper). First, the children were given the ToM task. The experimenter was seated with the child in a quiet room in the child's school. The child was told to watch a video presented on a laptop computer and listen carefully to what the characters said, because afterwards s/he would have to answer some questions. After watching the video, the child was asked two questions: 1.) *On buscarà la pilota, la nena?* 'Where will the girl look for the ball?'; and secondly 2.) *On és la pilota, en realitat?* 'Where is the ball really?' 1.) was considered correct if the child responded that the girl would look for the ball in the container where she left it. 2.) was considered correct if the child said that the ball was in the container where it was moved to.

After the child finished the false belief task, s/he was administered the comprehension task for either the AO, VO or AV condition. This was done with a between subjects design, such that each child saw only one of the three conditions. Like the false belief video, the Powerpoint was presented on a laptop computer. At the beginning of the task each child received familiarization trials. S/he was told that there was a set of twins, and their friend Marta (lower right hand panel in Figure 1). Marta was telling the twins about what she saw when she was on vacation. The children were also told that there would always be a twin that did not believe what Marta said, and that they would have to decide which twin that was based on how the twins reacted to Marta. For instance, the experimenter told the child that Marta was telling the twins that she saw a whale – *La Marta els explica que va veure una balena* 'Marta tells them that she saw a whale'. The experimenter then showed the child a reaction from each twin, one on top, and the other below. For each test trial, one twin produced a neutral echo question and the other produced a disbelief question. The children that received the AV condition saw one of the twins producing the facial gesture for a neutral echo question with the L+¡H* H% intonation contour. The other twin produced the facial gesture for the disbelieving echo question with the L* H% intonation contour. For the AO condition children saw still images and heard only the audio stimuli. For the VO condition the children saw the videos but without audio. After each twin spoke (or gestured as in the case of VO), the experimenter asked *Quin/a bessó/na no es creu la Marta, el/la de dalt o el/la de baix? Assenyala'l/-la.* 'Which twin does not believe Marta, the one on top or the one below? Point to him/her.' The child then pointed to the twin s/he thought did not believe Marta. The answer was marked correct by the experimenter if the child picked the twin that produced brow furrowing/backwards movement of head along with the L* H% for the AV condition, and the one that picked the relevant component of these for the AO (L* H%) and VO (brow furrowing/backwards head movement) conditions. Each participant received two blocks of stimuli. Children were either exposed to the female actor in Block 1, and the male actor in Block 2, or vice versa. They received four familiarization trials in Block 1, and two more in Block 2 in order to familiarize them to the new actor. For the familiarization trials, the neutral versus disbelief distinction in meaning of the test trials was maintained, but it was expressed lexically rather than intonationally (*Ah, que bé, que veiessis un gos*. 'Oh, that's nice that you saw a dog.' produced with a positive nodding head movement and *No m'ho crec, que veiessis un gos.* 'I don't believe that you saw a dog.' with a negative head shaking movement). Trials of this type were also used as fillers throughout the experiment. Therefore across the two blocks there were four filler trials in order to reorient the child were they to forget the intended meanings. For each actor, the child received six test trials and two fillers, for a total of twelve test trials and four fillers per child.

3. Analysis and Results

Two separate analyses were performed in order to account for performance on the two tasks described above. We first sought to test our predictions about children's performance for the three conditions. A total of 924 trials were analyzed. We first ran a set of mixed-effects logistic regression model in R [12] with CONDITION (three levels: AO, AV, VO), AGE (3, 4, 5[1]) and THEORY OF MIND (two levels: pass vs. not pass) as fixed effects and PARTICIPANT as a random effect. The dependent variable was CORRECT RESPONSE (correct vs. incorrect). Models were compared using ANOVAs. The best-fit model included CONDITION and AGE, but not THEORY OF MIND (pass or fail[2]). No interactions were included in the best-fit model. AGE was selected as a significant predictor of CORRECT

[1] The older 2-year-olds mentioned in 2.1 were included in the category "Age 3" and younger 6-year-olds were categorized as "Age 5".

[2] We did not include the second ToM question in the pass vs. fail decision, since all ages were shown to be at ceiling for this question.

RESPONSE. 4-year-olds performed significantly better than 3-year-olds (Estimate=1.35 vs. .38, $p<0.05$, SE=.38, z=-2.56) and 5-year-olds performed significantly better than 4-year-olds (Estimate = 2.70 vs. 1.35, $p<0.01$, SE=0.42, z=3.24). In order to investigate further any differences between conditions, a simple regression analysis was then performed. This analysis revealed differences for conditions. The slopes for the AO condition ($p<0.01$, r=0.70) and the AV condition ($p<0.01$, r=0.65) were both significant, while the slope for VO was not significant (p=0.09, r=0.39).

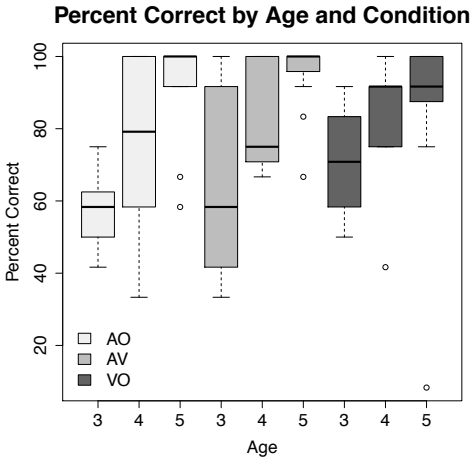


Figure 5: *Percent correct by Age for AO, AV & VO*

The fact that the slope for VO was not significant is evident if we inspect the boxplots in Figure 5, where less change between age groups is apparent for the VO condition when compared to the AO and AV conditions. Also evident in Figure 5 is the fact that the youngest children perform the worst on the AO task. 4-year-olds are better, but with a great deal of variability, suggesting that acquisition is still in progress. By Age 5, children are no longer struggling and show very little variability. A great deal of variability is observed for younger children that received the AV condition, especially for 3-year-olds. This could reflect the fact that even though children had gestural information available to reinforce the disbelief meaning in the AV condition, it may have been difficult to integrate the two cues. It is also possible that they ignored one modality because of working memory, which could have helped or hindered them depending on which one it was. Interestingly, 5-year-olds show more variability for the VO condition than for any other condition. This is perhaps because as children get older, they learn that facial cues do not always reinforce linguistic meaning.

With respect to ToM performance, Table 1 shows a general tendency observed. Across ages, children that pass the Sally Ann task tend to be more successful at the comprehension task, regardless of the condition.

Table 1. *% correct trials on linguistic comprehension task based on performance on Sally Ann task.*

Condition	Pass SA	Fail SA
AO	80	68
AV	89	73
VO	89	74

T-tests revealed that both the pass and fail groups performed at above-chance levels for all conditions. We conclude, then, based on the tendency observed in Table 1, that while explicit false belief reasoning (i.e. success on the Sally Ann task) may be helpful to children for the comprehension task, it is not a prerequisite for successful performance.

4. Discussion and Conclusions

In this experiment, each age group differed significantly from the next, showing that in general, children are making great strides in their comprehension of disbelief between the ages of 3 and 5. The simple regression analysis confirms our hypothesis that children might be better at inferring disbelief through the visual modality (significant slopes for AO and AV conditions but not VO). Inspection of the boxplot in Figure 5 reveals very different patterns for the three age groups across conditions. The amount of variability found for 3-year-olds depends on the condition. The 3-year-olds performed worst on the AO condition, but with less variability than 3-year-olds that received the AV condition, where the most variability for this age group is found. This suggests that combining the audio and visual modalities may help some children, but hinder others. The variability observed for 4-year-olds in the AO task, however, could be explained by a 'transition' stage of acquisition between the 3- and 5-year-olds. Four-year-olds are still in the process of abandoning the stage during which they rely on visual scaffolding, and are moving towards full acquisition of the linguistic meaning. For this reason, less variability is found for the 4-year-olds that received the AV condition when compared to 3-year-olds for that condition. Finally, 5-year-olds no longer require scaffolding from the visual modality, and do not seem to be thrown off by the presence of two modalities in the AV condition (there is very little variability). With respect to how explicit false belief reasoning might help children with our task, our results reveal a tendency for children that pass the Sally Ann task to perform better on the comprehension task. But across ages, those that failed the Sally Ann task still perform at above-chance levels, indicating that success on the Sally Ann task is not a prerequisite for success on the comprehension task. These results should be interpreted with caution, since our task tested *explicit* false belief judgments. Younger children may exhibit *implicit* knowledge of false belief [8]. It is obvious that some kind of mind-reading ability (i.e. ToM) must be present for children to perceive their interlocutor's state of disbelief, and this study shows that the modality through which this belief state is communicated matters. In linguistic theory, gesture is often thought to be peripheral to the study of speech. We argue that gesture is a critical component of the study of intonational development that should not be ignored.

5. Acknowledgements

We thank Page Piccinini for statistical analysis, Llorenç Andreu for help running participants, CE Jacint Verdaguer, Escola Sants Abdó, CEIP Sant Martí, the families that participated in this research and the actors, Anna and Lluís Gifra Prieto. The research was funded by a Spanish Ministry of Science and Innovation grant (FFI2012-31995 "Gestures, prosody and linguistic structure"), by a Generalitat de Catalunya grant (2009SGR-701) to the Grup d'Estudis de Prosòdia, and by the the grant RECERCAIXA 2012 for the project "Els precursors del llenguatge. Una guia TIC per a pares i educadors" awarded by Obra Social 'La Caixa'. Finally, we thank Maria del Mar Vanrell and Jill de Villiers for their helpful feedback.

6. References

[1] Crespo-Sendra, V., Kaland, C., Swerts, M. and Prieto, P., "Perceiving incredulity: The role of intonation and facial gestures", Journal of Pragmatics, 47:1-13, 2013.

[2] McNeil, N.M., Alibali, M.W., and Evans, J.L., "The role of gesture in children's comprehension of spoken language: now they need it, now they don't", Journal of Nonverbal Behavior 24(2), 131-150, 2001.

[3] Sauter, D.A., Panattoni, C., Happé, F., "Children's recognition of emotions from vocal cues", British Journal of Developmental Psychology 31: 97-113, 2013.

[4] Pierrehumbert, J., "The phonology and phonetics of English intonation". MIT PhD dissertation, 1980.

[5] Ladd, D. R., "Intonational Phonology", Cambridge University Press, 1996/2008

[6] Armstrong, M.E., "Child comprehension of intonationally-encoded disbelief", Proceedings of the 38th Boston University Conference on Language Development, accepted.

[7] Armstrong, M.E., "Puerto Rican Spanish intonation", in P. Prieto and Roseano, P., [Eds.], Transcription of intonation of the Spanish Language, Lincom EUROPA, 155-190.

[8] de Villiers, J., "The interface of language and Theory of Mind", Lingua: 117(11), 1858-1878, 2007.

[9] Baron-Cohen, S., Leslie, A. and Frith, U., "Does the autistic child have a 'Theory of Mind'?, Cognition 21: 37-46, 185.

[10] Wimmer, H., and Perner, "Beliefs about beliefs: representation and the containing function of wrong beliefs in young children's understanding of deception", Cognition: 13, 103-128, 1983.

[11] Nelson, N.L., Widen, S.C. and Russell, J.A., "The development of preschooler's Theory of Mind and emotion understanding", poster presented at the Bi-Annual Meeting of the Cognitive Development Society, 2007.

[12] Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R. and Vanrell, M., "Intonational phonology of Catalan and its dialectal varieties", in S. Frota and Prieto, P. [eds.], Intonational variation in Romance, Oxford University Press, in press, to appear in 2014.

[13] R Core Team, "R: a language and environment for statistical computing", R Foundation for Statistical Computer, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org/, 2013.

The pragmatic interpretation of intonation in Greek *wh*-questions

Amalia Arvaniti[1], *Mary Baltazani*[2], *Stella Gryllia*[3]

[1] English Language and Linguistics, University of Kent, UK
[2] Linguistics, Philology & Phonetics, University of Oxford, UK & University of Ioannina, Greece
[3] Leiden University Centre for Linguistics, Leiden University, Netherlands

a.arvaniti@kent.ac.uk, mary.baltazani@ling-phil.ox.ac.uk, s.gryllia@hum.leidenuniv.nl

Abstract

We experimentally investigated the pragmatics of two melodies commonly used with Greek *wh*-questions, L*H L-!H%, described as the default, and LH* L-L% considered less frequent and polite. We tested two hypotheses: (a) the !H%-ending melody is associated with information-seeking questions, while the L%-ending melody is pragmatically more flexible and thus appropriate also for non-information-seeking *wh*-questions expressing bias; (b) the !H%-ending melody, being more polite, is more appropriate for female talkers, all else being equal. In Experiment 1, comprehenders rated !H-ending and L%-ending versions of the same questions for politeness and appropriateness for the context in which they were heard (which favored either information-seeking or "biased" *wh*-questions). In Experiment 2, comprehenders heard the same questions and chose between two follow-up responses, one providing information, the other addressing the bias of the wh-question. Comprehenders rated !H%-ending questions more appropriate than L%-ending questions and judged the !H%-ending questions of female talkers more polite. They also chose information-providing answers more frequently after !H%- than L%-ending questions, but the preference was higher for female talkers and depended on comprehender gender. The results argue in favor of a compositional view of intonational meaning which depends not only on the tune but also on *context*, broadly construed.
Index Terms: *wh*-questions, intonation, pragmatics, gender

1. Introduction

We present data from two perception experiments on the intonational pragmatics of Greek *wh*-questions to argue that intonation requires, in addition to a description of its phonetic realization, a phonological representation which must take into consideration differences in meaning in tandem with differences in form. These results, in combination with the production study of [1] show that a phonological analysis is required to explain variation in the realization of intonation as well as differences in meaning and pragmatic interpretation.

In Greek *wh*-questions, the *wh*-word is utterance initial; thus the questions are marked both morphologically and syntactically as such, e.g., [pos se ˈlene] *what's your name?* (*lit.* how you.acc call.3pl). In addition to the morphosyntactic information, *wh*-questions are marked by the use of a particular melody (this melody can be used with other constructions as well [2]; a discussion of these cases is beyond the scope of this paper). In addition to this default melody, *wh*-questions can sometimes be uttered with another melody and a different pragmatic interpretation ([1], [2], [3], [4]). Past reports on these differences, combined with our own assessment as native speakers, constitute the background of our experiments, which probe the pragmatic interpretation of the *wh*-questions when used with these two melodies.

1.1. Melodies of Greek wh-questions

Illustrations of the melody used by default with *wh*-questions are presented in Figure 1 below (based on [1]). A comparison between the two panels of Figure 1 shows that there is variation in the realization of the melody. In Figure 1a, which shows a short question, F0 starts high, reaching a peak on the stressed vowel of the *wh*-word [pu] "where," after which it quickly dips before a final small rise. Figure 1b shows a longer question, in which F0 starts low, has a late peak that occurs on the syllable *following* the *wh*-word [apo ˈpu] "from where"), and shows a rather extensive low F0 stretch before the final rise. Despite the obvious phonetic differences, these two contours are recognized by native speakers as instances of the same melody. Further, the variation illustrated in Figure 1 is systematic: it depends on the length of the question, the length of the *wh*-word itself, and the position of the stressed syllables with respect to each other and the utterance edges [1]. The representation L*H L-!H% ([1], [3]), abstracts away from phonetic detail and allows us to predict systematic differences in realization, including those illustrated in Figure 1.

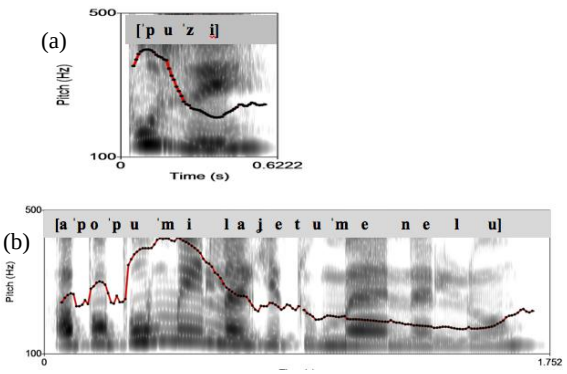


Figure 1: *Spectrograms and F0 contours of wh-questions: In panel (a), [ˈpu ˈzi] "where does he live?"; in panel (b), [apo ˈpu ˈmilaje tu ˈmenelu] "from where was she speaking to Menelos?". For details see text.*

Although the description above covers the typical realization of contours used with *wh*-questions in Greek and their phonological analysis, certain issues remain unresolved. First, as noted, Greek uses an additional melody with *wh*-questions. This has been analyzed as L*H L-L%, a melody similar to L*H L-!H% but ending at the bottom of the speaker's range, rather than with a final rise ([1], [3]). The fact that the two contours can be elicited under the same conditions, [1], puts into question the posited phonological difference between !H% and L%; !H% could represent simply a return to a default mid-level pitch rather than a meaningful difference (cf. [6]). If so, then the difference between !H% and

L% is one of phonetic realization and as such it need not be included in the phonological representation.

As a first step in addressing this issue, we conducted an exploratory production study in which two female and two male native speakers of Greek produced sixteen questions in two types of contexts which, based on our assessment as native speakers, should lead to the use of either the !H%- or the L%-ending melody; the contexts were similar to those presented in (1a) and (1b) below.

Both our male and our female speakers produced distinct melodies in response to the different contexts suggesting that the two melodies convey different pragmatic meaning and therefore, that they are phonologically distinct. Acoustic analysis of these data indicates that the melodies differ systematically not only in the way they end but also in the pitch accent associated with the *wh*-word. The !H%-ending melody has an accent best represented as L*H ([1], [7]), as it starts with a marked rise (Figure 2, filled symbols, solid lines), while the L%-ending melody has an accent best represented as LH* ([3], [4], [8]), which typically starts with a peak (as its L tone is truncated when the *wh*-word is short (Figure 2, unfilled symbols, broken lines). These results support previous descriptions ([1], [3]) about a systematic difference in the way the melodies end but also establish differences in regards to the pitch accent on the *wh*-word (cf. [4]).

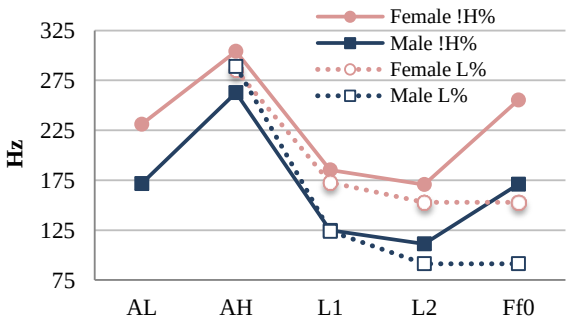


Figure 2: *Average F0 values (in Hz) of the stimuli used in the two experiments, separated by gender and melody; AL = F0 onset; AH = accentual peak; L1, L2 = beginning and end of low-F0 stretch respectively; Ff0 = final F0 value in melody. Note that L2 in the L%-ending contours is added for clarity.*

Here we investigate the differences in meaning associated with these differences in realization by means of two perception experiments conceived on the basis of the production data briefly discussed above and the following observations by [1] and [3]. According to [3], L*H L-L% sounds less polite or less "involved" than L*H L-!H%, as if the speaker does not care for an answer; [1] note that the L%-ending tune was rare in their data (accounting for only 8% of tokens) and most instances were elicited from male talkers.

Based on the above, our hypotheses regarding the two melodies were as follows. First, the !H%-ending melody is the default melody for *wh*-questions and therefore the most appropriate when questions serve their primary function of seeking information. Second, the L%-ending melody is appropriate for both information-seeking and non-information seeking questions. Non-information-seeking questions can serve various functions; e.g. [4] discusses rhetorical questions and notes they are produced with the L%-ending melody. Here we investigated questions indicating questioner bias for a specific answer (cf. [9]); such questions serve as an indirect way of making a statement. We hypothesized that !H%-ending questions would be deemed inappropriate in this context, while L%-ending questions would be highly preferred. Finally, we hypothesized that the L%-ending melody, being less polite, would not be as appropriate for female as for male talkers, especially in requests for information [3].

2. Exp. 1: Appropriateness and politeness

In Experiment 1 comprehenders rated LH* L-L% and L*H L-!H% versions of six *wh*-questions for their appropriateness and politeness in a given context.

2.1. Participants

Eighty-nine comperhenders took part in the experiment. They provided information about their linguistic background and history on the basis of which 13 were not considered for further analysis as they turned out to be either bilingual or have a history of speech or hearing disorders. Two more comprehenders were excluded as they failed to respond to more than 20% of the trials. Results reported here are based on 74 comprehenders, 56 female and 18 male. They were all monolingual native speakers of Greek studying at the University of Ioannina, and ranged in age from 18 to 22 years.

2.2. Stimuli

The stimuli were six pairs of Greek *wh*-questions, one !H%- and one L%-ending version per pair. The questions were selected from our corpus of 128 questions discussed in section 1.1. and were evenly divided among the four speakers of that corpus. The six pairs of questions were chosen on the basis of their naturalness. The total number of stimuli was 48 questions (6 *wh*-questions × 2 melodies × 4 speakers).

We constructed two contexts for each question, so that each context in a pair would most likely lead to a different response: an information-seeking question, as in (1a) or a biased question as in (1b). Specifically, a question following a context such as (1b) would be interpreted in Greek as an attempt by the speaker not to seek information but, rather, to elicit addressee acquiescence to an indirect point (which reflects the questioner's bias for a particular answer). In our example, this indirect point is recognition on the part of the addressee that going to Syntagma would be difficult, if not impossible, under the circumstances. The contexts were read by a different native speaker of Greek. Contexts and questions were crossed for a total of 96 trials (48 melodies × 2 contexts) so that each question was heard after (a) a context that made asking for information a plausible action or (b) a context that did not necessitate an information-seeking action.

(1a) Context: *Lena, who is visiting Athens for the first time, stops a passerby for directions:*
Question: ['pos θa 'pao sto 'sidaɣma]
'How will I get to Syntagma?'

(1b) Context: *A protest march in Syntagma is scheduled for the time Kostas has an interview there; as they listen to the news, Kostas says to his wife:*
Question: ['pos θa 'pao sto 'sidaɣma]
'How will I get to Syntagma?'

We expected that !H-ending questions would be rated more appropriate after contexts like (1a), while L%-ending questions would be rated more appropriate after contexts like

(1b). In addition we expected that in information-seeking contexts !H%-ending questions would be rated more polite than L%-ending questions, and that melody would interact with talker gender so that !H-ending melodies would be rated more polite if the talker was female.

2.3. Procedures

The comprehenders heard each context followed by a question over loudspeakers in a classroom at the University of Ioannina and filled in hard copy response sheets. They were tasked with rating how appropriate and polite each question was in the context that preceded it, using a 1-7 rating scale. The timeline of each trial is presented in Figure 3.

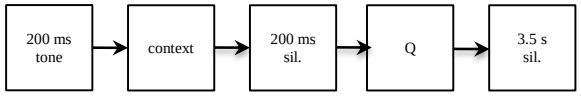


Figure 3: *Timeline of a trial in Experiment 1.*

2.4. Results

Ordinal logit regression showed that comprehenders judged questions more appropriate when they were preceded by a context that made information-seeking a plausible action [Wald = 219.6, $p < 0.0001$; Figure 4a]. They also rated !H%-ending questions more appropriate than L%-ending ones [Wald = 155, $p < 0.0001$; Figure 4b].

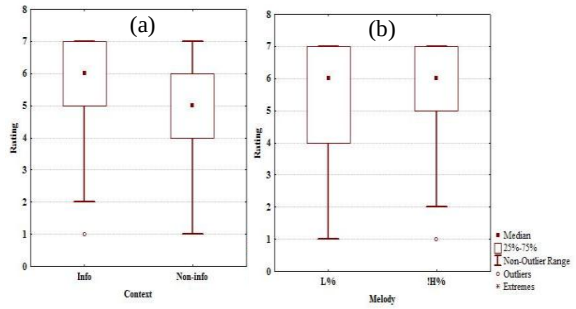


Figure 4: *Appropriateness ratings as a function of context (left) and melody (right).*

Regarding politeness, results showed an interaction between melody and talker gender [Wald = 15.5, $p < 0.0001$]. Melody did not affect the rating of questions uttered by male talkers, but it did affect questions by female talkers: their questions were judged more polite when !H%-ending (Figure 5).

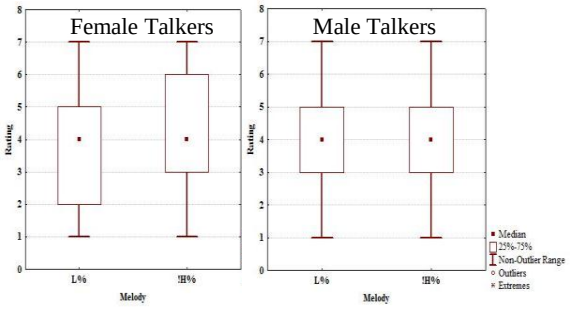


Figure 5: *Politeness ratings as a function of melody, separately for female (left) and male (right) talkers.*

Post-hoc analysis also showed that both appropriateness and politeness were affected by *comprehender* gender, with females giving overall lower politeness ratings than males [Wald = 35.5, $p < 0.0001$] but higher appropriateness ratings [Wald = 52.2, $p < 0.0001$].

3. Experiment 2: Pragmatic interpretation

3.1. Participants and stimuli

A different set of 79 comprehenders took part in Experiment 2. The data of six of them were discarded for the same reasons as before. The results reported here are based on 73 comprehenders (55 female and 18 males) with the same demographics as in Experiment 1. The same !H%- and L%-ending versions of the six questions used in Experiment 1 were also used here.

3.2. Procedures

The questions were presented aurally out of context under the same conditions as in Experiment 1. The comprehenders' task was to choose one of two possible responses to each question, presented to them in hard copy response sheets: (i) an information-providing response or (ii) a response that agreed with the bias implied by the question. There was a total of 48 trials (6 *wh*-questions × 2 melodies × 4 speakers).

The setup is illustrated in (2): comprehenders heard a question like "how will I get to Syntagma?" (stimulus) and had to choose between two possible responses (counterbalanced across trials): Response A which provides information and Response B, which concurs with an implicit bias attributed to the questioner. The timeline of a trial is presented in Figure 6.

(2) Stimulus: ['pos θa 'pao sto 'sidaɣma]
'How will I get to Syntagma?'
Response A: *You will take line 3 and get off at (stop) Syntagma.*
Response B: *You're right, you can't go. There'll be mayhem.*

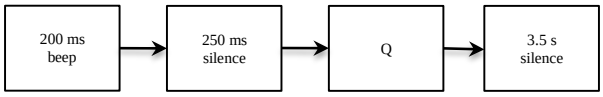


Figure 6: *Timeline of a trial in Experiment 2.*

It was expected that the comprehenders would be more likely to interpret !H%-ending questions as information-seeking and thus choose an information-providing response, such as Response A in (2). On the other hand, L%-ending questions would be more likely interpreted as indicating the questioner's bias, rather than seeking information. Thus comprehenders would be more likely to select the answer that did not provide information but concurred with this bias, such as Response B in (2).

3.3. Results

Logit regression showed that comprehenders preferred information-providing responses to responses concurring with (the questioner's implied) bias when questions were !H%-ending [Wald = 49.1, $p < 0.0001$]. The preference was stronger for female than male talkers [Wald = 7.6, $p < 0.01$]. Panels (a) and (b) of Figure 7 illustrate these two points respectively.

Post-hoc analysis showed an additional effect of comprehender gender [Wald = 6.9, $p < 0.01$], indicating that among comprehenders females chose information providing responses less often than males (Figure 7c). Post-hoc investigation of the interaction between comprehender and talker gender [Wald = 4.7, $p < 0.05$], illustrated in Figure 8 further indicates that the difference between male and female comprehenders was due to the fact that female comprehenders chose information providing responses less often when the talkers were male; i.e., they more frequently interpreted male than female stimuli as more likely to indicate bias rather than be genuine requests for information.

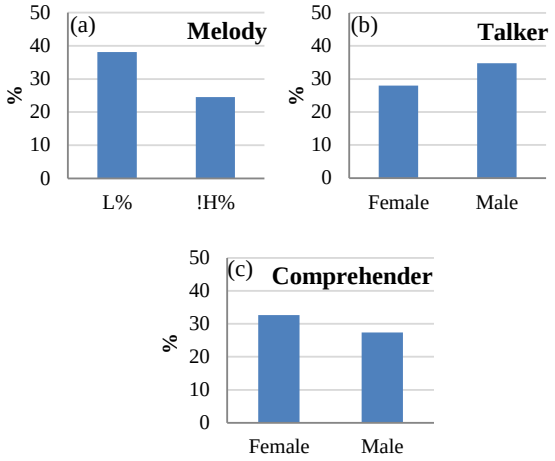


Figure 7: *Percentages of bias concurring responses as a function of melody (panel a), talker gender (panel b) and comprehender gender (panel c).*

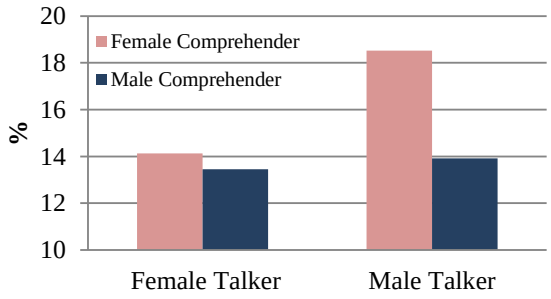


Figure 8: *Percentages of bias concurring responses as a function of comprehender and talker gender.*

4. Discussion

Our results confirmed that the two boundary tones cannot be seen as "allophonic" as each one leads to a different evaluation of the questions, in terms of their politeness and appropriateness as responses, as well as to a different interpretation of their pragmatic intent. In addition, our data showed that the assumption held in the literature, [1], [3], that the accent of the two melodies is the same must be incorrect, at least for the questions used in our experiments. Overall then, our results support the view that we are dealing with two different melodies, L*H L-!H% and LH* L-L% (cf. [4]).

Here we offer a preliminary compositional pragmatic analysis of the two melodies, which is based on attributing different pragmatic interpretations to the two melodic components that vary between the two tunes: the pitch accent (L*H or LH*) and the boundary tone (!H% or L%). Each of these components contributes to the pragmatic interpretation of the whole question.

Specifically, we maintain that the two melodies differ in givenness and completeness status: the L*H L-!H% melody is composed of a pitch accent which marks new information in Greek ([7]), and a boundary tone which marks the utterance as incomplete thereby inviting an answer ([1]; cf. [10]). The LH* L-L% melody is composed of a pitch accent typically used to mark contrastive focus in Greek: it conveys that the accented item (and not some alternative) should be believed (cf. [8]) and marks the remainder of the utterance as given. The L% boundary tone marks the utterance as complete ([1]; cf. [10]). The result of combining these components, L*H with !H% and LH* with L%, is that while the former questions are interpreted as requiring an answer, the latter need not function as such. This in turn explains why the L*H L-!H% melody is restricted to questions proper, while the LH* L-L% melody can be used more widely.

The results are also of interest from the point of view of processing and the value of experimental research in intonational pragmatics. First, our experiments showed that the comprehenders preferred taking *wh*-questions at face value, i.e. interpreting them as requests for information. This is not surprising given that, as noted, *wh*-questions in Greek are morphosyntactically marked as such. It is significant however that this preference was modulated by melodic changes which shifted responses towards alternative interpretations without concomitant morphosyntactic changes. Crucially, both talker and comprehender gender played a part in the interpretation and evaluation of questions, with female comprehenders in Experiment 2 being more likely to interpret stimuli from male talkers as statements than as questions (as compared to how they treated the stimuli from female talkers). This suggests that intonational pragmatics does not depend only on the interaction of melody with semantics and linguistic context, as is often maintained, but can be affected by additional factors, such as talker and addressee gender. Thus focusing exclusively on speaker intent when examining intonational pragmatics may be unnecessarily constricting, since, clearly, all participants in a conversation play an active part in constructing intonational meaning.

5. Conclusion

The results confirmed L*H L-!H% as the default melody for Greek *wh*-questions, supporting our hypothesis about a difference in the pragmatic interpretation of the L*H L-!H% and LH* L-L% melodies that string-identical Greek *wh*-questions are uttered with. Since these interpretations were available to comprehenders out of context and despite the presence of a fronted *wh*-word clearly marking the stimuli as *wh*-questions, our findings suggest that intonation can win over conflicting morphosyntactic information. Evidence was also found that both talker and comprehender gender must be factored into the pragmatic analysis, in addition to melody, semantics and linguistic context. These results strongly suggest that acknowledging comprehender expectations about the social use of melodies is crucial for fully understanding intonational pragmatics. Finally our results show that closer attention is due to the systematic differences in meaning that relate to melodic variation.

6. References

[1] Arvaniti, A. and Ladd, D. R., "Greek *wh*-questions and the phonology of intonation", Phonology, 26: 43-74, 2009.

[2] Baltazani, M., Quantifier Scope and the Role of Intonation in Greek. Doctoral thesis, UCLA. 2002.

[3] Arvaniti, A. and Baltazani, M., "Intonational analysis and prosodic annotation of Greek spoken corpora", in S. Jun (Ed], Prosodic Typology: The Phonology of Intonation and Phrasing, 84-117, Oxford University Press, 2005.

[4] Dimos, K., "Intonational features of Greek questions [Epitonika xaraktiristika erotimatikon protaseon tis ellinikis]". Master's thesis, Univeristy of Ioannina, Greece, 2011.

[5] Ladd, D.R., Intonational Phonology. Cambridge University Press, 2008.

[6] Grabe, E., Comparative Intonational Phonology: English and German. (MPI Series in Psycholinguistics 7), Ponsen and Looijen, 1998.

[7] Arvaniti, A., Ladd, D.R. and Mennen, I., "Stability of tonal alignment: the case of Greek prenuclear accents", Journal of Phonetics, 26: 3-25, 1998.

[8] Arvaniti, A., Ladd, D.R. and Mennen, I., "Effects of focus and 'tonal crowding' in intonation: Evidence from Greek polar questions", Speech Communication, 48: 667-696, 2006.

[9] Gunlogson, C., True to Form: Rising and Falling Declaratives as Questions in English. New York: Routledge, 2003.

[10] Pierrehumbert, J. and Hirschberg, J., "The meaning of intonation in the interpretation of discourse", in P. Cohen, J. Morgan, and M. Pollack [Eds], Intentions in Communication, 271-311, MIT Press, 1990.

Temporal stability of long-term measures of fundamental frequency

Pablo Arantes[1], *Anders Eriksson*[2]

[1]Languages and Linguistics Department, São Carlos Federal University, Brazil
[2]Department of Linguistics, Stockholm University, Sweden

pabloarantes@gmail.com, anders.eriksson@ling.su.se

Abstract

We investigated long-term mean, median and base value of F0 to estimate how long it takes their variability to stabilize. Change point analysis was used to locate stabilization points. In one experiment, stabilization points were calculated in recordings of the same text spoken in 26 languages. Average stabilization points are 5 seconds for base value and 10 seconds for mean and median. Variance after the stabilization point was reduced around 40 times for mean and median and more than 100 times for the base value. In another experiment, four speakers read two different texts each. Stabilization points for the same speaker across the texts do not exactly coincide as would be ideally expected. Average change point dislocation is 2.5 seconds for the base value, 3.4 for the median and 9.5 for the mean. After stabilization, individual differences in the three measures obtained from the two texts are 2% on average. Present results show that stabilization points in long-term measures of F0 occur earlier than suggested in the previous literature.

Index Terms: fundamental frequency, long term measurements, forensic phonetics

1. Introduction

The study of statistical measures of location or preferred value of the voice fundamental frequency (F_0) of an individual has at least two applications. The first one is the development of F_0 contour normalization procedures [1][2][3], that are used to factor out the most common value of a particular speaker so that variation due to linguistic components of the contour becomes more evident. In the other important usage, knowledge of preferred F_0 value of individual speakers is often relevant to speaker comparison in forensic case work (see [4] and references therein), speech technology applications [5] and the development of security systems that require user authentication by voice, for instance (see [6] for a recent survey of the field).

When it comes to the estimation of preferred F_0 value, important issues are (i) what long-term statistical measures are better suited to do it and (ii) how long should a speech sample be in order to the estimated value be representative of an individual's speech? Regarding the choice of estimator, the literature mostly cites the mean and standard deviation (for an overview of the subject see [4]). Given that long-term F_0 distributions are usually skewed towards higher values, it seems advisable to compare the mean to alternative measures that make no assumptions regarding normal distribution of sample values. To fill this gap, we propose to compare the mean to two other measures, the median and the base value. The base value can be conceived as a neutral and speaker-specific value of F_0 below which maintaining phonation becomes difficult and to which speakers return to after excursions of linguistic or expressive value (see [7] for a detailed explanation and section 2.4 for the definition implemented in this study). Both the base value and the median are quantile-based measures that are robust to the presence of skewness or extreme values in the F_0 sample, although the median does not share all of base value's properties.

Regarding the appropriate minimum length of speech sample required for long-term mean to stabilize, the literature shows no definite consensus. Eriksson [4] makes reference to five different estimates, ranging from 14 seconds to two minutes. There is also no agreed upon objective way of estimating when a long-term measure has reached a stable point, most researchers resorting to visual inspection of trajectories of cumulative mean. In this study we explore a statistical technique called change point analysis as a way to objectively compare the performance of the three measures studied here.

2. Materials and Methods

2.1. Language effect

To study the possible effect of language on the variability of measure of location of F_0, a set of recordings of speakers of 25 languages reading the "The North Wind and the Sun" passage were analyzed. The recordings are publicly available on the website of the International Phonetic Association (IPA). One recording of a Brazilian Portuguese speaker reading the same text was included in the sample. Sixteen speakers are male. The sample includes languages of eight linguistic families: Afro-Asiatic, Sino-Tibetan, Indo-European, Uralic, Niger-Congo, Altaic, Tai-Kadai and Turkic.

2.2. Text effect

To test the possible effect of the text being read on the time it takes the long term measures of F_0 to stabilize, recordings of four speakers reading two different texts were analyzed.

The texts are the "North Wind and the Sun" passage translated to Brazilian Portuguese and a passage of "A Menina do Narizinho Arrebitado" by Brazilian writer Monteiro Lobato. The first is coded text 1 and the second text 2. Text 2 is phonetically balanced in the sense of having all Brazilian Portuguese (BP) phonemes, while in translating and adapting the "North Wind and the Sun" passage to BP the goal was to be faithful to the semantic content and not to make sure all the phonemes in the language were being used.

The recordings are by two male and two female speakers of two Brazilian states, São Paulo and Minas Gerais. Speakers from São Paulo and Minas Gerais are referred to by the sp1 and sp2 labels, respectively, followed by -f or -m to indicate if it is a female or male speaker.

2.3. Acoustical analysis

F_0 contours for every recording analyzed were extracted with the help of a Praat script that implements a heuristic suggested by Hirst [8] that tries to minimize extraction errors such as octave or fifth jumps by optimizing floor and ceiling values passed to Praat's auto-correlation F0 extraction algorithm[1]. Remaining errors were hand-corrected. Further processing of F_0 contours to obtain cumulative measures of location was done by a second Praat script written specifically for this purpose.

2.4. Measures of location

The following statistical measures of location were investigated:

- Arithmetic mean
- Median (50th-quantile of the sample F_0 values)
- Base value (7th-quantile of the sample F_0 values)[2]

All measures were taken cumulatively from the first voiced frame up to the last in non-overlapping steps of 200 ms. All F_0 values within each 200 ms interval are included in the computation of the measures. The number of F_0 samples contained in each 200 ms step depends on the floor parameter provided to Praat's F_0 extraction algorithm. In the IPA languages sample, the average minimum value for male speakers was 70 Hz and 120 Hz for female speakers, which gives us 20 values for male speakers and 32 for female speakers each 200 ms.

In the IPA languages sample, the median duration of recordings was 38 seconds with values ranging from 25 seconds (Galician) to 66 seconds (Thai). In the text effect experiment, recordings of text 1 have an average duration of 32 seconds and recordings of text 2 have an average duration of 41.3 seconds.

Mean and median values are usually close, but base value, by definition, is smaller than both. Since here we are more interested in how their variability changes over time and not specially in their absolute values, a normalization procedure was applied so that the three time series can be seen in the $[0, 1]$ interval. This was done by means of formula 1, where f_i is the i^{th} raw F_0 value in a given contour and $f_{\min}$ and $f_{\max}$ are respectively the minimum and maximum values in the contour:

$$(f_i - f_{\min})/(f_{\max} - f_{\min}) \quad (1)$$

The normalized values were used for visualization purposes only. The statistical analysis were carried out on the raw values (in Hz) of the cumulative measures.

2.5. Statistical analysis

Our main interest is to determine how long it takes for the time series defined by the cumulative measures of location of F_0 studied here to have its variability reduced to what could be considered a stable value. In most of the literature on the subject, what we are calling the stabilization point has been determined by visual inspection of the time series of cumulative measure of location. Although the visual inspection can be useful, it would be important to develop a less subjective and more automatic way of determining stabilization points.

A statistical technique called change point detection analysis was used to attain a greater level of objectivity in determining stabilization points. This technique estimates the point in time at which underlying statistical properties (mean, variance or both) of a time series change. A function of the R package changepoint [9] was used in the analysis to find a point that divides the time series in in two parts having different variances and tests the hypothesis that the two values are significantly different. We searched for single variance change points in cumulative mean, median and base value time series. Since distribution of cumulative mean, median and base value are highly skewed, an algorithm that does not assume that the values in the time series follow a normal distribution was used.

3. Results and discussion

3.1. Language effect

Figure 1 shows temporal evolution of cumulative normalized measures of location. Wide-range fluctuations in the three estimators are a general trend across languages, specially at the first seconds of the recordings. As estimators' values are computed over longer stretches of time, the range of fluctuations gets increasingly smaller, but with notable differences between the languages: in some cases, variability quickly drops (e.g, Arabic, German, Galician, Hungarian, Slovene and Swedish) whereas, in other cases, the reduction seems to be more gradual (e.g., Brazilian Portuguese, Catalan, French, Irish and Korean). For all languages, with the possible exception of Turkish, the cumulative value of the three estimators tend to reach a stable value at some point, usually within the first fourth of the recording duration.

Table 1 lists the temporal location of change points for the 26 languages investigated as determined by the change point analysis described in section 2.5 as well as ratio of variance before and after the change point. Figure 2 shows a box plot of change points broken down by typical measure type.

One of the main findings is that change points in this sample seem to happen in the low range of values suggested in most of the previous literature on the subject or even earlier than that (see section 1). The other main finding is that the base line tends to stabilize a little earlier (5 seconds) then mean and median (about 10 seconds). The base value change points are also less variable (median absolute deviation of 2.2) than mean and median (MAD of 6.2 and 7.6 respectively). Inspection of variance reduction factors in Table 1 suggests that in fact the points identified by the change point analysis can be considered stabilization points. Base value also has a superior performance in terms of variance reduction: base line has an average reduction factor of 120 and the mean and the median an average factor of 48.

3.2. Text effect

Figure 3 shows cumulative values of long-term measures of typical F_0 for four speakers reading two different texts. Table 2 lists change points in the three measures as well as variance reduction factors.

For a given long-term estimator, the change point and the estimator value for a given speaker would ideally be the same, regardless of the text being read. Strictly speaking, that was not the case for the four speakers in our sample: average absolute difference between change point for text 1 and text 2 is 4.9 seconds, with a minimum of 0.4 and a maximum of 21.1 seconds. 75% of the differences are under 6 seconds. Considering

[1] Available at http://code.google.com/p/praat-tools/

[2] In [7], the authors say that the base value can "as a rule of thumb, be expected to be about 1.5 σ below his average F_0". To Define the base value in terms of a quantile is for all practical purposes equivalent to the definition based on standard deviation, assuming F_0 values are normally distributed.

Table 1: *Change point locations (in seconds) in the mean, median and base value time series for the 26 languages investigated. Ratio of variance before and after change point are shown in parentheses.*

language	mean	median	base value
Amharic	11 (44)	16.2 (137)	4.8 (69)
Arabic	4.2 (242)	4.6 (226)	4.6 (619)
Brazilian Portuguese	10.4 (33)	10.8 (33)	5.2 (41)
Bulgarian	16.2 (48)	15.6 (79)	3.8 (440)
Cantonese	6.2 (26)	7 (8)	6 (41)
Catalan	11 (40)	10.6 (33)	11.2 (31)
Croatian	0.8 (62)	12.4 (162)	0.6 (180)
Czech	6.8 (86)	4.8 (77)	8 (204)
Dutch	8.2 (50)	10.2 (36)	4.4 (778)
English	11.6 (48)	1.6 (5)	2.4 (63)
French	15 (7)	16 (8)	3.4 (22)
Galician	5.2 (28)	4.8 (67)	5.2 (100)
German	5.2 (141)	5.8 (135)	3.8 (26)
Hindi	10 (112)	16.4 (322)	10 (84)
Hungarian	2.4 (194)	3.8 (171)	4 (217)
Igbo	7.8 (19)	8.8 (2)	21 (222)
Irish	14.4 (7)	12.6 (17)	15 (10)
Japanese	6 (183)	11.6 (58)	0.4 (1772)
Korean	21.4 (18)	21.6 (13)	1 (3)
Persian	20.6 (27)	20.6 (21)	4.6 (3)
European Portuguese	11 (116)	6.4 (81)	5.8 (219)
Sindhi	15 (66)	15.4 (44)	6.4 (172)
Slovene	9 (141)	14.8 (21)	0.8 (232)
Swedish	6.2 (108)	2.8 (504)	15.4 (271)
Thai	22.2 (23)	24.4 (52)	27.6 (137)
Turkish	20 (11)	3.2 (3)	4.6 (16)

that the standard deviation of change point location on the IPA language sample is 6 seconds, the differences in change point location between text 1 and text 2 are less then what would be expected when comparing samples of different languages. A comparison of the differences between the raw values (in Hz) of the cumulative mean, median and base value of text 1 and text 2 at the time of change point shows that on average the difference is 2%, with a range going from 0 to 9%, the four speakers polled. 90% of the differences are under 4%, i.e., less than one semitone. These data indicate there is an effect due to text whose magnitude is slightly smaller than that due to language.

None of the texts yielded overall earlier change points or greater variance reduction factors. The only exception to that is the long-term median of text 2, whose change points are earlier than those of text 1 for all four speakers. It's not clear if the behavior of the median can be attributed to the fact that text 2 is phonetically balanced and why only the median should be affected by this particular feature of text 2.

4. Conclusions

We set out to compare three long-term measures of typical value for F_0, namely the mean, median and base value, in two respects: how long it takes for each measure to achieve a more or less stable level of variability and how much they are affected by language and text.

Our results indicate that long-term measures tend to stabilize at most 30 seconds after the begining of a recording of read speech, with median times around 10 seconds. Of the three measures, the base value seems to yield the earliest and less variable change points, confirming earlier findings suggesting its robustness against factors such as variation in speaker emotional state, vocal effort and channel quality.

Table 2: *Change point (seconds) in the mean, median and base value for the four speakers and the two texts. Ratio of variance before and after change point are shown in parentheses.*

speaker	text	mean	median	base value
sp1-f	1	9.8 (18)	9.6 (21)	9.6 (6)
sp1-f	2	15.2 (28)	5.2 (13)	5.6 (83)
sp1-m	1	10.2 (92)	10.2 (58)	6.2 (33)
sp1-m	2	4.8 (22)	4.8 (42)	5.2 (7)
sp2-f	1	7.2 (11)	7.2 (4)	9.4 (9)
sp2-f	2	28.2 (8)	5.6 (6)	9.8 (39)
sp2-m	1	10.4 (33)	10.8 (33)	5.2 (41)
sp2-m	2	24 (20)	8 (14)	13.8 (39)

The results also show that both language and the specific text being read seem to cause variability in stabilization points. The base value is less affected by the language effect than the mean and median measures. The text effect is slightly smaller than the language effect and the three measures seem to be equally affected by it.

In order to increase the accuracy of the results obtained in the present study, in follow-up studies we are going to increase the number of speakers in the languages investigated and the length of the recordings. Two promising avenues of investigation worth exploring are the effects of different speaking styles (reading vs. spontaneous speech, for instance) and non-contemporaneous recordings on the temporal stability of long-term measures.

5. References

[1] Jassem, W. "Normalisation of F0 curves", in Fant, G. and Tahtam, M. [Eds], Auditory Analysis and Perception of Speech, 523-530, Academic Press, 1975.

[2] Rose, P. "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency?", Speech Communication 10, 229–247, 1991.

[3] Maidment, J. A. and Garca Lecumberri, M. L. "Pitch Analysis Methods for Cross-Speaker Comparison", Proceedings of ICSLP 1996, v. 4, 2247–2249, 1996.

[4] Eriksson, A., "Aural/Acoustic vs. Automatic Methods in Forensic Phonetic Case Work", in Neustein, A. and Patil, H. A. [Eds], Forensic Speaker Recogntion: Law Enforcement and Counter-terrorism, 41-69, Springer-Verlag, 2011.

[5] Ferrer, L, Shriberg, E. and Stolcke, A. A prosody-based approach to end-of-utterance detection that does not require speech recognition. Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2003.

[6] Müller, C. [Ed], Speaker Classification I: Fundamentals, Features and Methods, Springer-Verlag, 2007.

[7] Traumüller, H. and Eriksson, A. "The frequency range of the voice fundamental in the speech of male and female adults". Manuscript. Retrieved via http://www2.ling.su.se/staff/hartmut/f0_m&f.pdf.

[8] Hirst, D. "The Analysis by Synthesis of Speech Melody: from Data to Models", Journal of Speech Sciences 1(1):55-83, 2011.

[9] Rebecca Killick and Idris Eckley. "changepoint: An R package for changepoint alysis". R package version 1.1. http://CRAN.R-project.org/package=changepoint, 2013.

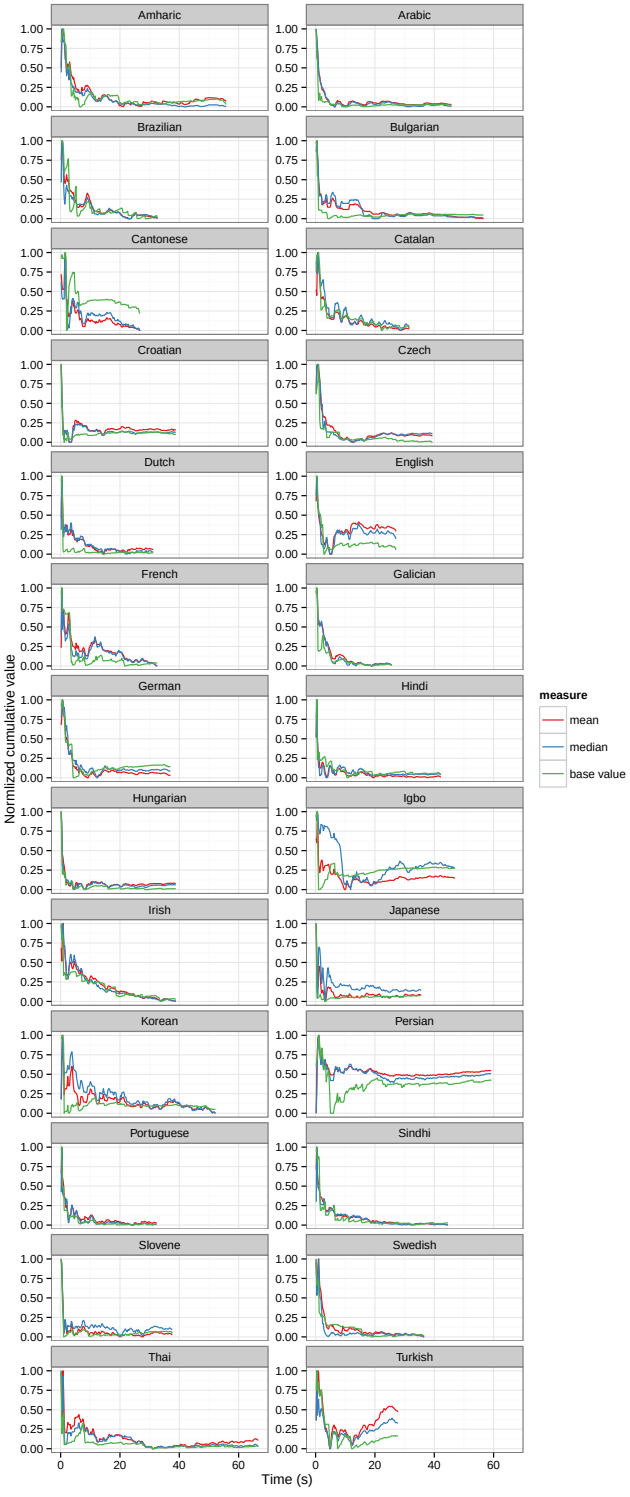


Figure 1: *Language effect on typical F_0 value. Vertical axis shows normalized cumulative mean, median and base value for 26 languages.*

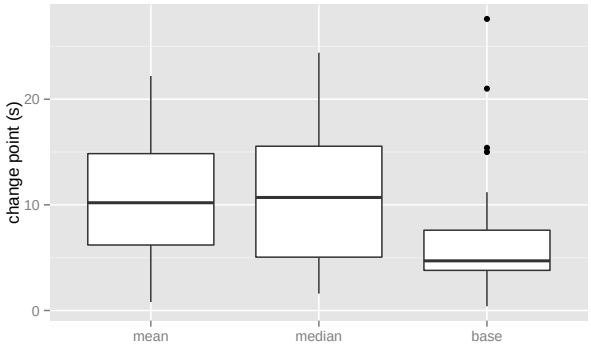


Figure 2: *Box plot of change points (seconds) of mean, median and base value time series for the 26 languages in the IPA sample.*

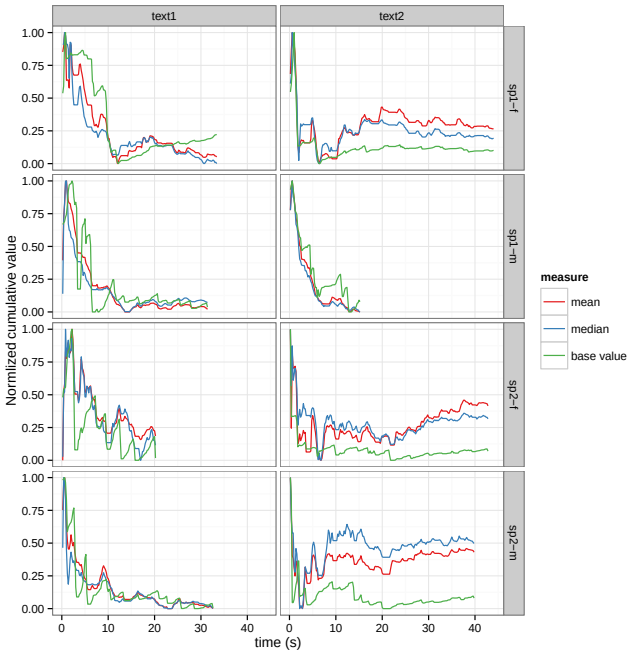


Figure 3: *Normalized cumulative mean, median and base value for the four speakers and the two texts.*

17 Friday 3

Probabilistic prosody: Effects of relative speech rate on perception of (a) word(s) several syllables earlier

Meredith Brown[1], *Laura C. Dilley*[2], *Michael K. Tanenhaus*[1]

[1]Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY, US
[2]Department of Communicative Sciences & Disorders, Michigan State University, East Lansing, MI, US

mbrown@bcs.rochester.edu, ldilley@msu.edu, mtan@bcs.rochester.edu

Abstract

Speech perception depends on the ability to rapidly accommodate considerable variability in speech rate. We present results from two eye-tracking experiments indicating that listeners use context speech rate to generate, maintain, and update probabilistic hypotheses about the timing and number of constituents in upcoming speech. Participants heard utterances containing polysyllabic nouns preceded by indefinite articles and followed by [s]-initial words (e.g. *...saw a raccoon slowly...*). We altered the speech rate of the indefinite article and of the [s] with respect to surrounding context, manipulating the likelihood that the item would be perceived as singular (*a raccoon*) vs. plural (*raccoons*). Shorter indefinite articles elicited higher proportions of fixations to plural target pictures than longer articles both before and after the processing of [s], demonstrating that listeners made rapid use of prosodic cues to the presence or absence of the article. Importantly, fixations were also influenced by the duration of [s] relative to context speech rate. These findings suggest that listeners maintain and update provisional speech-rate hypotheses across multiple morphophonemic units. We interpret these results with respect to probabilistic approaches to spoken language understanding.

Index Terms: perception of prosody, speech rate, expectations, eye movements, language comprehension

1. Introduction

The realization of prosodic information in speech, such as pitch accents, speech tempo, and other intonational features, is highly variable (e.g. [1], [2], [3]). This variability poses numerous challenges for spoken language processing. For example, the realization of temporal speech cues like voice onset time depends on an individual's overall speech rate. Comprehension therefore crucially depends on the ability of listeners to interpret prosodic cues and rate-dependent speech cues with respect to surrounding context (e.g. [4], [5], [6]). A comprehensive understanding of the role of prosody in spoken language comprehension requires an explanation of how listeners accommodate contextual information during real-time processing.

In this paper we explore a possible explanation for how listeners interpret prosody in context based on emerging *data-explanation* approaches to perception and cognition. Data-explanation approaches posit a central role for *generative processes* within perceptual systems that give rise to probabilistic expectations about incoming sensory input based on high-level representations and contextual information (e.g. [7], [8]). During speech perception, we hypothesize that listeners continuously make and update inferences about the source of the speech signal (i.e. the communicative intention of the speaker) by comparing internally generated probabilistic expectations about the acoustic realization of the speech signal to the actual speech signal as it unfolds. This provides a potential explanatory framework for the integration of multiple distinct and temporally distributed constraints during spoken language processing [9]. Particularly compelling from this perspective are so-called *distal prosody* effects [10], [11]. For example, manipulating pitch and timing patterns across utterance material several syllables before a temporarily ambiguous word (e.g. *panda*, which can initially be interpreted as *pan*) influences the time course of lexical competition, even when the prosodic characteristics of the target word itself are unaltered [12], [13]. These findings suggest that listeners develop expectations based on prosodic patterns in speech that extend across a relatively wide window.

The data-explanation framework provides a potential explanation for how expectations based on preceding prosody are mapped onto the acoustic-phonetic properties of the unfolding speech signal. It also makes the prediction that listeners continuously update their provisional hypotheses about the source of the speech signal on the basis of additional downstream information. The present study investigates this prediction by capitalizing on effects of context speech rate on the number of words that are perceived within a stretch of speech [11]. Compressing the speech rate of portions of an utterance surrounding a highly coarticulated word (e.g. the underlined segments surrounding the determiner "a" in *The Smiths wouldn't buy a Butterball...*) reduces the likelihood that listeners report hearing this word. Likewise, when the word in question is not present in the signal (e.g. *buy Butterball...*), slowing down the segments surrounding its potential location makes listeners more likely to perceive the word within the slow portion of speech.

In the present study, we investigate whether and how listeners maintain and update provisional prosodic percepts based on downstream cues, by manipulating the relative speech rate of multiple temporally-distributed cues to the plurality of a noun phrase – the presence or absence of the indefinite determiner *a* before the noun and of the plural marker *–s* following the noun. Our experiments use the *visual world paradigm* to examine listeners' eye movements to pictures depicting singular and plural versions of the target noun phrase, providing an index of the time course of interpretation [14], [15]. The goals of this work were to determine (a) whether and how listeners combine multiple temporally distributed prosodic cues to the plurality of a referring expression; and (b) whether the time course of prosodic cue integration is consistent with the hypothesis that listeners rapidly update their previous prosodic expectations and provisional percepts in light of downstream information.

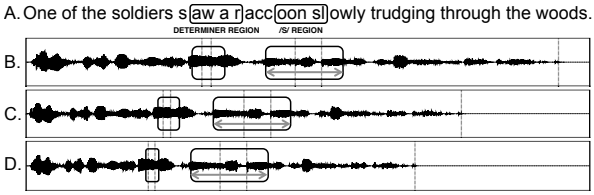


Figure 1: *Illustration of the distal /s/ speech rate manipulation in Experiment 1: (a) Example stimulus sentence; (b) 86% determiner, 95% distal /s/; (c) 86% determiner, 75% distal /s/; (d) 86% determiner, 65% distal /s/. Boxes indicate the position and duration of determiner and /s/ regions, respectively.*

2. Experiment 1

2.1. Methods

2.1.1. Participants

We recruited 36 students from the University of Rochester to participate in Experiment 1. All participants were native speakers of American English and had normal hearing and normal or corrected-to-normal visual acuity.

2.1.2. Materials

The stimuli were 54 grammatical declarative sentences containing a highly imageable singular noun phrase (e.g. *a raccoon*) followed by a word starting with an /s/ (e.g. Fig. 1a). Recordings of these sentences were elicited from 12 speakers. These speakers also recorded 81 filler items whose target nouns were equally likely to be singular or plural and whose number was unambiguous (e.g. *those bathrobes*). We selected one token of each item for use in the experiment, consisting of 4-5 critical item tokens and 6-7 filler item tokens from each speaker. Critical item tokens were selected such that they had a relatively high degree of coarticulation on the indefinite determiner "a" and continuous articulation of the target word and the following /s/, such that the presence or absence of the plural marker *–s* was not clearly signaled.

Stimuli were manipulated in two ways (Fig. 1b-d). First, the speech rate of the region consisting of the determiner and the segments immediately surrounding it (*determiner region*) was compressed to 92%, 86%, or 78% of its original rate. Then, we compressed the speech rate of utterance context preceding and following the segments surrounding the /s/ following the target word (*/s/ region*), such that it was 95%, 75%, or 65% of its original rate. The absolute physical duration of the /s/ region remained the same across conditions, but its speech rate relative to surrounding context increased with successive levels of context speech rate compression. Levels of each manipulation were selected based on norming data. The global speech rate of filler items was manipulated such that equal numbers of items were compressed to 95%, 75%, or 65% of their original rate.

2.1.3. Procedure

On each trial, participants were presented with a computer screen containing a four-picture visual display. The display contained singular and plural versions of the target word and of a distractor picture. Each participant heard a single version of each item over Sennheiser HD 570 headphones after 500 ms of display preview. Their task was to click on the picture that they heard referred to in each sentence. Eye movements were recorded using a head-mounted SR Research EyeLink II system sampling at 250 Hz, with drift correction procedures performed following every fifth trial.

Two lists were created by pseudo-randomizing trial order and rotating picture positions 180 degrees. An additional set of two lists was created by reversing the order of these lists. An equal number of critical items in each list were assigned to each of the nine pairings of determiner condition and distal speech rate condition. The assignment of items to conditions was counterbalanced across participants. All lists started with six filler items to ensure that participants were familiar with the task prior to encountering critical items.

2.1.4. Analyses

Response choices and fixations were analyzed separately. Data from trials on which the participant incorrectly selected one of the two distractor pictures (less than 0.5% of trials) were excluded. Selections of singular vs. plural target pictures were analyzed using multilevel logistic regression. Proportions of fixations to singular and plural target pictures on each trial were averaged across two windows of interest: (a) an *early window* 400 ms in duration (the mean duration of the target word), starting 200 ms before and ending 200 ms after the onset of the /s/; and (b) a *late window* between 200–1000 ms following the onset of the /s/. Both windows were selected to take into account a 200 ms delay for programming and executing fixations. Mean proportions of fixations to the plural target picture were divided by the mean propotion of fixations to both target pictures to calculate a *plural target advantage ratio* across each window for each trial, which was transformed using the empirical logit function [16], [17]. Plural target advantage ratios were analyzed using linear regression. The significance of predictors in the linear regression models was estimated by assuming convergence of the t distribution with the z distribution [18]. Models were computed using the *lme4* package in R (version 2.15.0) [19], [20]. All regression models had determiner speech rate, distal speech rate, and their interactions as fixed effects, and full random effects structure except as noted due to lack of convergence [21]. Factors were contrast coded with the least rate-manipulated level set as the reference level (i.e. 92% determiner, 95% distal /s/). Model comparison procedures were used to remove fixed effects that did not contribute significantly to model fit according to the likelihood ratio test [18].

2.2. Predictions

We predicted that compressing determiner speech rate would result in increased selections of plural target pictures and higher plural target advantage ratios in both the early and late windows, replicating early effects of relative speech rate on spoken language processing observed in previous work [22]. Importantly, we predicted that compressing speech rate of material distal to the /s/ (effectively slowing down the speech rate of /s/ relative to surrounding context) would also result in increased plural target picture choices and proportions of fixations to plural target pictures following the onset of /s/.

2.3. Results and discussion

2.3.1. Picture choices

Participants' response choices were consistent with our predictions (Fig. 2). The multilevel logistic regression model of response choices confirmed that participants were more likely to select plural target pictures when the determiner had a faster

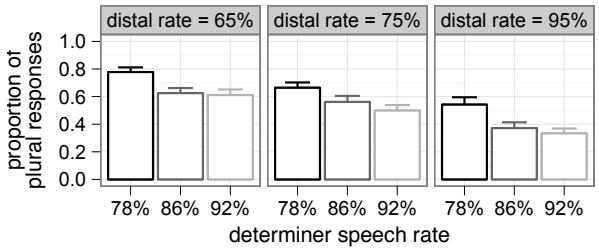


Figure 2: *Proportions of plural responses in Experiment 1.*

speech rate (66% plural responses for the most compressed determiner compared to 48% for the least compressed determiner; β=1.42, z=5.76, $p<.0001$). The distal /s/ manipulation also had the predicted effect. Participants were most likely to choose plural target pictures when the /s/ was surrounded by the most compressed speech, and therefore had a slower speech rate with respect to the utterance context (67% plural responses). They were least likely to choose plural target pictures when the /s/ was surrounded by the least compressed speech (42% plural responses; β=1.89, z=7.96, $p<.0001$).

2.3.2. Proportions of fixations

Figure 3 shows fixations to plural and singular target pictures over time with respect to the onset of the /s/. Analysis of plural target advantage ratios across this window revealed early effects of determiner speech rate on fixation proportions (Fig. 3, top). More compressed determiners were associated with more looks to plural target pictures and fewer looks to singular target pictures, compared to less compressed determiners (β=0.11, t=1.92, p_{est}=.055). This finding replicates early effects of speech rate on determiner perception previously observed in related work and suggests that effects of speech rate manipulation on the perception of short words have a locus in perceptual expectations [22]. Effects of determiner speech rate persisted into the late analysis window (β=.13, t=3.97, $p_{est}<.0001$).

The distal /s/ speech rate manipulation also exhibited the predicted effects within the late analysis window, during and following the processing of the /s/ (Fig. 3, bottom)[1]. Plural target advantage ratios were highest when the /s/ was surrounded by the most compressed speech, and therefore had a slower speech rate than the utterance context, than when the /s/ was surrounded by relatively slow speech (β=.17, t=3.61, $p_{est}<.0005$).

These results suggest that determiner perception is influenced by prosodic information influencing whether the /s/ multiple syllables downstream is perceived as containing the plural morpheme *–s*. However, it is also possible that these effects are instead attributable to the absolute duration of the determiner. The determiner is compressed first as part of the determiner region manipulation, and then as part of the distal /s/ speech rate manipulation (which involves compressing all material distal to /s/, including the determiner). The absolute duration of the determiner therefore covaries with the relative speech rate of the /s/. For example, in the 86% determiner speech rate condition, the absolute duration of the determiner is in fact 82% of its initial duration in the slowest distal speech rate condition and 56% in the fastest (cf. Fig. 1b-d). It is therefore possible that apparent effects of the speech rate of /s/ relative to surrounding context could instead have resulted from effects of distal speech rate compression on the absolute duration of the determiner. These possibilities cannot be distinguished on the basis of the time course data, because of the overall effects of speech rate compression on fixations prior to the onset of the target noun phrase.

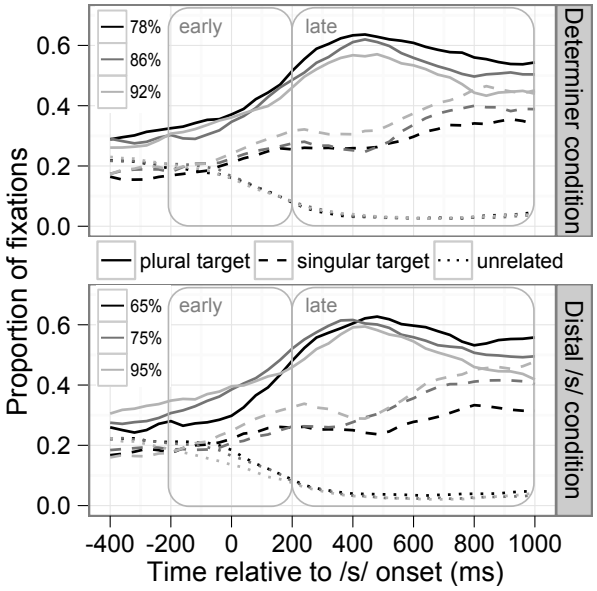


Figure 3: *Proportions of fixations in Experiment 1, by determiner (top) and distal /s/ (bottom) conditions. Superimposed shapes depict early and late analysis windows.*

To address this potential confound, we conducted a second visual world experiment in which we manipulated the speech rate of the /s/ by slowing down the /s/ region itself, rather than by compressing surrounding context. This proximal /s/ speech rate manipulation had no effect on the absolute duration of the determiner. Thus, if the effects that we observed in Experiment 1 were merely attributable to effects of the distal speech rate manipulation on the absolute duration of the determiner, we would not expect to observe effects of the proximal /s/ speech rate manipulation on picture choices or fixations to singular or plural pictures. If, however, the effects found in Experiment 1 were due to the speech rate of the /s/ relative to surrounding context, we would expect to see effects of the proximal /s/ manipulation that are similar to those that we observed in Experiment 1.

3. Experiment 2

3.1. Methods

Participants were 36 University of Rochester students meeting the same criteria as for Experiment 1. In addition, the experi-

[1] We also found marginally significant effects of the distal /s/ speech rate manipulation on fixations within the early window (i.e. before the processing of the /s/; β=.10, t=1.74, p_{est}=.082). Post-hoc analyses revealed that the logit-transformed sum of fixations to both target pictures differed as a function of distal speech rate condition (β=-0.17, t=-3.89, $p_{est}<.0005$). Increased compression of the utterance context was associated with lower proportions of fixations to target pictures within the early analysis window, because the analysis window was time-locked to the onset of the /s/. It is therefore likely that these effects of distal /s/ speech rate manipulation on pre-/s/ fixations are merely attributable to information about the target word becoming available at different times and different rates. Importantly, because the baseline effects observed in the early analysis window were in the opposite direction of the predicted effects of distal speech rate following the onset of the /s/, they did not complicate interpretation of effects in the later analysis window.

ment setup, stimulus lists, data collection, and analysis procedures were the same as in Experiment 1.

We used the recordings from Experiment 1 to create the stimuli for Experiment 2. The determiner manipulation was the same as in Experiment 1. However, instead of manipulating the relative speech rate of /s/ by compressing the speech rate of material surrounding the /s/ region, we instead manipulated the /s/ region directly by slowing its speech rate to 110%, 150%, or 170% of its original rate. Following this proximal /s/ speech rate manipulation, all critical and filler items were globally compressed to 85% of their original duration, to maintain similarity with Experiment 1 in terms of global stimulus characteristics and overall experiment duration.

3.2. Results and discussion

3.2.1. Picture choices

The multilevel logistic regression model of response choices. Participants again selected plural target pictures more frequently when the determiner region had a faster speech rate (65% in the fastest condition compared to 46% in the slowest condition; β=1.43, z=6.86, $p<.0001$). Crucially, the speech rate of /s/ also influenced response choices in the predicted direction, such that participants were more likely to select plural pictures when the /s/ region had a slower speech rate (62% plural responses in the slowest condition compared to 43% in the fastest condition; β=1.48, z=8.76, $p<.0001$). This suggests that the effects of the speech rate of /s/ relative to surrounding context influenced listeners' judgments in both experiments, rather than simply the absolute duration of the determiner.

3.2.2. Proportions of fixations

The multilevel linear regression of plural target advantage ratios in the early analysis window indicated no significant effects of determiner or /s/ speech rate manipulation. The determiner manipulation used in these experiments was subtle relative to manipulations used in previous related work, in which early effects of determiner speech rate were reliably observed (Brown et al., 2012). Previous work also used a multi-word target expression, providing a larger window of analysis with more statistical power. Although the manipulation we used in the present experiments elicited robust effects in response choices and in overall fixation behavior, it is possible that it was nevertheless too subtle to elicit large enough effects to be reliably observed immediately after the processing of the determiner, even though numerical trends in the predicted direction emerged within this window. In addition, other factors such as a lack of variation in the global speech rate of filler items may have contributed to subtle differences in time course and/or magnitude of effects across experiments.

Crucially, however, analysis of plural target advantage ratios during the late analysis window revealed significant effects of not only determiner but also /s/ speech rate. As predicted, more compressed determiners were associated with higher plural target advantage ratios (i.e. more fixations to plural pictures and fewer to singular pictures; β=.28, t=5.80, $p_{est}<.0001$). In addition, plural target advantage ratios were higher when the speech rate across the /s/ region was the slowest (β=.21, t=4.33, $p_{est}<.0001$). This finding, together with the significant effect of /s/ speech rate on picture choices, indicates that the effects of Experiment 1 cannot be explained merely on the basis of the absolute duration of the determiner across different levels of the distal /s/ speech rate manipulation. Eliminating this confound provides stronger evidence that the interpretation of function words can be modulated by information encountered considerably later in the utterance.

4. Discussion

Our results provide evidence that listeners combine multiple temporally distributed prosodic cues to the plurality of a referring expression during real-time spoken language comprehension. Prosodically conditioned percepts of short words like determiners can be influenced by congruent or conflicting information in the speech signal that occurs substantially downstream. These findings suggest that listeners maintain and update provisional hypotheses about previously encountered material across multiple morphophonemic units.

Our findings are closely aligned with recent work demonstrating that language processing is influenced by information spanning a wider temporal integration window than standardly assumed [23], [24], [25]. For example, when listeners hear a target word *leash* in a sentence context, lexical competition with *leaves* is stronger when the target word follows the verb *shakes* (whose rhyme *rakes* is semantically related to *leaves*) than when it follows the verb *rattles* [23]. This suggests that residual uncertainty about the perceived verb influences lexical competition effects several syllables downstream.

The present work further demonstrates that these "right-context" effects also extend to uncertainty about the timing and number of constituents in preceding speech, based on preceding prosodic information. These findings are difficult to explain with respect to traditional feed-forward models of spoken language comprehension that assume that listeners map acoustic patterns onto more abstract representations (e.g. words and morphemes) prior to interpreting the perceived representations with respect to sentence- and discourse-level context. Rather, our results suggest that listeners maintain and update probabilistic inferences about speakers' intended meaning (such as the intention to produce a singular or plural construction) based on available prosodic information across a relatively wide window.

5. Conclusions

Prosody influences spoken language processing in a gradient, probabilistic fashion. Further, prosodically-conditioned percepts are maintained and updated across multiple morphophonemic units. These findings are most naturally explained within a probabilistic data-explanation account of spoken language processing, involving probabilistic inference about the communicative intentions that give rise to the acoustic realization of an utterance. These inferences inform fine-grained probabilistic expectations about how aspects of lexical alternatives will be realized in context. They can be also updated in light of subsequent information encountered in the unfolding utterance.

6. Acknowledgements

This work was supported by an NSF predoctoral fellowship to MB, NSF grant BCS-0847653 to LCD, and NIH grants HD073890 and HD027206 to MKT. We gratefully thank Dana Subik and Chelsea Marsh for assistance with recruiting and testing participants, and the audiences at AMLaP 2013 and AP-CAM 2013 for feedback on earlier versions of this work.

7. References

[1] Ladd, D. R. (2008). *Intonational phonology*, (2nd Ed), Cambridge Studies in Linguistics.

[2] Badino, L. & Clark, R. A. J. (2007). Issues of optionality in pitch accent placement. In *Proc. 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007.

[3] Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128*, 839-850.

[4] Miller, J. (1987). Rate-dependent processing in speech perception. In A. Ellis (ed.), *Progress in the psychology of language* (pp. 119-157). London: Erlbaum Associates.

[5] Repp, B. H., Liberman, A. M., Eccardt, T., & Pesetsky, D. (1978). Perceptual integration of acoustic cues for stop, fricative, and affricate manner. *Journal of Experimental Psychology: Human Perception and Performance, 4(4)*, 621-637.

[6] Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance, 37(3)*, 978-996.

[7] Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11(10)*, 428-434.

[8] Clark, A. (2013). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Brain and Behavioral Sciences, 36,* 181-204.

[9] Farmer, T. A., Brown, M., & Tanenhaus, M. K. (2013). Prediction, explanation, and the role of generative models in language processing [Commentary]. *Behavioral and Brain Sciences, 36,*, 31-32.

[10] Dilley, L., & McAuley, J. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59,* 291-311.

[11] Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21(11),* 1664-1670.

[12] Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin and Review, 18(6),* 1189-1196.

[13] Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (under review). Metrical expectations from preceding prosody influence spoken-word recognition. *Journal of Experimental Psychology: Human Perception and Performance.*

[14] Cooper, R. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6*, 84-107.

[15] Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632-1634.

[16] Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

[17] Cox, D. R. (1970). *The analysis of binary data.* London: Chapman and Hall.

[18] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed- effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390-412.

[19] Bates, D., Maechler, M., & Bolker, B. (2011). lme4: Linear mixed-effects models using S4 classes (R package, version 0.999375-42) [Computer software]. Online: http://CRAN.R-project.org/package=lme4.

[20] R Development Core Team. (2012). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: http://www.R-project.org.

[21] Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255-278.

[22] Brown, M., Dilley, L. C., & Tanenhaus, M. K. (2012). Real-time expectations based on context speech rate can cause words to appear or disappear. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 1374-79).

[23] Johnstone, S., Trueswell, J., & Dahan, D. (2013). Partially activated words participate in combinatory semantic interpretation during sentence processing. Talk presented at the 26th CUNY Conference on Human Sentence Processing (Columbia, SC).

[24] Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences, 106*, 21086-2109.

[25] McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language, 60(1),* 65-91.

The role of intonation in early word recognition and learning

Jill C. Thorson and James L. Morgan

[1] Department of Cognitive, Linguistic and Psychological Sciences,
Brown University, Providence, USA
Jill_Thorson@brown.edu, James_Morgan@brown.edu

Abstract

The motivation for our study is to investigate how English-acquiring toddlers are guided by the mapping between intonation and information structure during on-line reference resolution and in novel word learning tasks. We ask whether specific pitch movements (*deaccented*, *monotonal*, *bitonal*) more systematically predict patterns of attention and subsequent novel word learning abilities depending on the referring or learning condition (*new*, *given, contrastive*). Experiment 1 examines the attentional patterns of 18-month-old toddlers when referents are either *new* or *given* in the discourse, and carry one of the three pitch movement types. Contrary to previous work, results show increased attention to the target in the deaccented condition if the referent is *new* to the discourse. Also, both monotonal and bitonal pitch movements direct attention to the target even when the target is *given*. Thus, pitch type interacts with information structure in directing toddler attention. Experiment 2 tests two-year-olds in a novel word learning task, varying pitch type and contrastiveness during learning. Preliminary results show that learning is aided when the novel word is introduced in contrast to a previous referent. Together, these two experiments demonstrate the role of pitch type and information structure in guiding attention and aiding early word learning.

Index Terms: intonation, information structure, first language acquisition

1. Introduction

From birth, infants are sensitive to native language rhythm and pitch patterns ([1], [2]). They can then approximate adult-like intonation contours from the onset of production ([3], [4], [5]) and align these contours with felicitous semantic and pragmatic intentions ([6]). However, little research has been conducted on the early comprehension of contours as they reflect information status.

Previous research shows that toddler attention to referents can be mediated by both intonation and information structure in discourse ([7]). In turn, attention to a referent is essential for making the correct word-to-object or word-to-action mappings necessary for early word learning ([8]). The motivation for our study is two-fold: 1) to investigate how American English-acquiring 18-month-olds are guided by mappings from intonation to information structure during on-line reference resolution in discourse (Experiment 1), and 2) to investigate how the pitch accent on a novel word interacts with its referential status to aid early word learning in American English-acquiring 24-month-olds (Experiment 2).

Social pragmatics and intentionality are often cited as essential in order to achieve successful word learning ([9], [10]). Our study tests outcomes when live interactions are removed from the experimental design and any degree of intentionality is only accessible through the utterances themselves and their corresponding prosody.

This study focuses on two pitch accents in American English, the simple monotonal H* and the complex bitonal L+H*. Previous work in adult speech perception shows that the simple H* pitch accent can be associated with either new or contrastive information in discourse, and the complex L+H* is more typically associated with a contrastive interpretation ([11]). Our experiments exploit these mappings in American English in order to test how these pitch accents interact with referential newness, givenness, and contrastiveness during early attentional processes and word learning.

2. Experiment 1

Previous research by Grassmann and Tomasello (2010) claimed that German-acquiring two-year-olds attend to a referent of a familiar word if and only if the word is both *stressed* and *new* to the discourse ([7]). Their experiment consisted of three conditions where the target referent could be introduced in a short discourse context with (a) stress only, (b) newness only, or (c) stress and newness. They used a live speaker to present the stimuli to each subject in order to ensure a level of intentionality on the side of the speaker and measured looking time and pointing to a referent as their dependent variables.

Our experiment expands upon the work by Grassmann and Tomasello (2010) but makes a number of methodological changes. First, we add in a deaccented condition to the design to act as a control against the accented (or 'stressed') conditions. Second, we ask whether specific pitch movements more systematically predict patterns of attention to a referent, rather than using one "stressed" category. Third, since a live speaker may have produced varying intonation contours, we control for speaker variations by using pre-recorded stimuli. Finally, we isolate the role of pitch in guiding attention, holding duration and intensity constant. This is a first step towards understanding how each of the acoustic correlates of intonation contribute and interact in guiding attention.

Specifically, we consider how unique pitch movements facilitate attention when a referent is either *new* or *given* in the discourse. The pitch types tested are a simple monotonal rise (~H*), a complex bitonal rise (~L+H*), and a deaccented control pattern. First, we predict that regardless of pitch type, newness will guide attention to a referent. Second, a semantically and pragmatically appropriate pitch accent will guide attention to both *new* and *given* referents.

2.1. Method

2.1.1. Participants

Data were analyzed for 48 American English-acquiring 18-month-old toddlers (27 female). The age of participants ranged from 529 to 589 days, with a mean age of 551 days. Thirteen additional participants were discarded due to fussiness (9), experimental error (1), or equipment malfunction (3). All participants were from Providence, RI, USA, and surrounding areas.

2.1.2. Stimuli

An adult female speaker produced all target and distractor utterances. In order to create the three different pitch types, the speaker first produced carrier sentences with H* accents on the target words using careful speech (slow and clear), but not child-directed speech. The pitch contour of only the target word was then digitally manipulated in Praat ([12]) to create the simple monotonal and complex bitonal versions. Deaccented target words were spliced into the carrier phrase and matched for duration and intensity to the two accented versions. All test stimuli were resynthesized. Naïve listeners judged the resynthesized target speech sounds as natural.

There were 6 (C)CVC monosyllabic target words and 18 (C)CVC monosyllabic distractors used in the procedure (See Table 1 for a full list of stimuli by trial). All target words were phonologically distinct within a trial and primarily sonorant in nature. All of the target and distractor words are commonly known by 18-month-olds. Previous knowledge of the words was confirmed by a vocabulary questionnaire completed by the caregiver. If the toddler was not familiar with a particular word, then that particular trial was eliminated during analysis (this was not a common occurrence, and it only affected at most one trial per participant).

Table 1. *Stimuli for Experiment 1.*

Trial Number	Target Word	Distractors/Fillers
Practice	spoon	cake
		bear
		fish
1	ball	sock
		lamb
		cat
2	moon	dog
		shoe
		book
3	cow	pig
		tree
		duck
4	doll	bus
		cup
		sun
5	plane	star
		truck
		dress

2.1.3. Design and procedure

We used a 2x3-mixed design to test toddlers' responses to changes in information status and intonation, isolating the specific role that pitch plays in directing attention to *new* or *given* referents. The independent variable of Information Structure *(new* vs. *given)* was manipulated within subjects, whereas the independent variable of Pitch Type (*deaccented, simple, complex)* was manipulated between subjects.

Each trial consisted of a Context Phase and a Test Phase. During the Test Phase, the test utterance played and the *proportion of looking time (PLT) to the target* was collected using the SMI iView X™ RED eye tracker. The Context Phase consisted of two parts (or slides), which established the Prior Discourse Context before the Test Phase was introduced (See Figure 1). Importantly, the target referent in the Test Phase was either *new* or *given* to the discourse as well as in contrast to an item in the previous slide, making both the simple (~H*) and the complex (~L+H*) movements acceptable in this location.

There were 5 familiarization trials and 2 test blocks per between-subjects condition. A test block included 1 practice trial and 5 test trials for a total of 6 trials in each block. With two blocks per condition (one *new* and one *given* condition), there were a total of 2 practice trials and 10 test trials (12 total trials). Trial order within a block was randomized and block order was counter-balanced across participants. The location of the target items on the screen (left or right) was also counter-balanced within and across conditions.

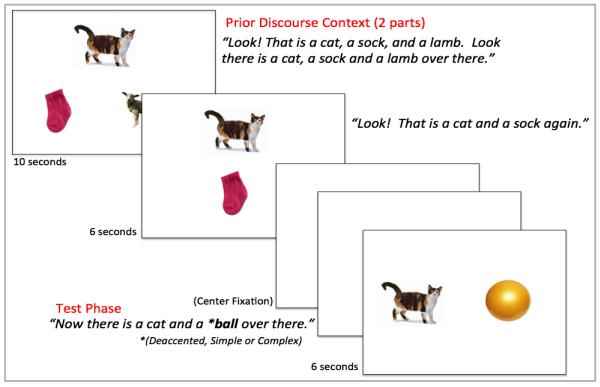


Figure 1: *Example new trial from Experiment 1.*

2.2. Results

A 2x3 repeated measures ANOVA shows a significant main effect of Information Structure ($F(1,45) = 32.36$, $p < .001$, $\eta_p^2 = .418$) and a significant main effect of Pitch Type ($F(2,45) = 6.34$, $p = .004$, $\eta_p^2 = .220$). The two-way interaction of Information Structure by Pitch Type approached but did not quite reach significance ($F(2,45) = 3.01$, $p = .059$).

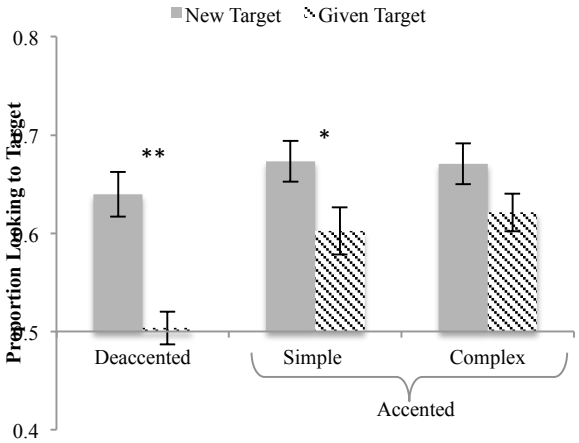


Figure 2: *Bar graph of proportion looking to a new or given target referent for each pitch type condition. Error bars show +/- SE. *: p < .05, **: p < .01.*

Planned comparisons between groups reveal more looking to a target than a distractor when the referent is *new* to the discourse context, regardless of pitch accent type (See Figure 2). Additionally, both types of accentuation (*simple* and

complex) guide more looks to the target than to the distractor when the target referent is either *new* or *given* to the discourse context. In the *simple* condition, there was significantly longer looking to the target when it is *new* to the discourse than when it is *given*. This difference between *new* and *given* target referents does not reach significance in the *complex* condition. Crucially, there are more looks to the target than the distractor referent in all conditions except when the target is *deaccented* and *given* in the discourse.

2.3. Discussion

Contrary to previous literature, *newness* is sufficient to draw 18-month-olds' attention to a referent in a discourse, even without pitch accentuation. A preference for the novel (or *new*) stimulus item over a familiar (or *given*) one suggests a more mature level processing by 18-months-olds ([13], [14]). Toddlers prefer to look at the more prominent or salient item in the discourse, where salience in this case is achieved through pitch movement and newness effects.

Additionally, even in the case of a target referent that is *given* in the discourse, both *simple* and *complex* pitch movements guide attention to this referent. Thus, the presence of either *newness* or a *pitch accent* shifts attention to a target referent in a discourse, regardless of pitch movement type. This suggests that the more salient or prominent the stimulus item, the more a toddler will look. When a target is both new and carries a pitch movement, the result is even greater looking to the target. For *given* information, attention is being driven by the pitch movement on the referent word.

Importantly, we observe robust effects for the variable of pitch type even when only the acoustic dimension of pitch is manipulated. With intensity and duration held constant, any observed pitch accent effects were a consequence of pitch (f0) manipulation. Thus, even with a more complex discourse phase and limited acoustic cues to stress, we find a very different pattern of results from previous research.

In this experiment, the two pitch types analyzed were a simple monotonal and a complex bitonal pitch movement. Future work will extend analyses to other types of pitch accents, discourse contexts, as well isolate the roles of other acoustic cues to prosody (i.e. duration and intensity). Critically, significant methodological differences change the complexion of results in comparison to previous work. Understanding the mechanisms for guiding attention is essential for subsequent early word learning.

3. Experiment 2

As demonstrated in Experiment 1, toddler attention to a referent is mediated by both the intonation and the information structure of the discourse. Experiment 2 extends these findings to investigate how pitch type interacts with contrastiveness during a novel word learning task.

Both newness and pitch were shown to aid in guiding attention during a discourse context in Experiment 1, where the target item was also in contrast to a previous referent. Contrastiveness here is defined as introducing a referent in direct opposition to a one that is previously mentioned. This type of contrast without pitch accentuation on the target word was not sufficient to guide attention to the intended referent. Interestingly, with the addition of a context appropriate pitch accent movement, attention increased to a target referent over a distractor. The goal of experiment 2 is to explore how discourse contrastiveness interacts with intonation to aid in early word learning.

First, we predict that contrastive learning situations are more likely to shift attention to a referent and aid in word learning, particularly when paired with the more prominent contrastive accent ([15]). Second, from Experiment 1 we know that accentuation facilitates attention to a referent. Thus, we predict that contrastiveness and the presence of a pitch accent will aid in the learning of a novel word.

3.1. Method

3.1.1. Participants

Preliminary data were analyzed for 12 American English-acquiring 24-month-old toddlers (7 female). The age of participants ranged from 722-767 days, with a mean age of 743 days. Four additional participants were discarded due to fussiness (3) or inattentiveness (1). All participants were from Providence, RI, USA, and surrounding areas.

3.1.2. Stimuli

The same female speaker who recorded the stimuli for Experiment 1 produced all utterances using careful speech (slow and clear), but not child-directed speech. The speaker was trained in intonational phonology and was able to produce H* and L+H* accents consistently across the different stimulus items. Each target word was produced naturally in an utterance with an H* accent, a L+H* accent, or as deaccented. The target word was then spliced out of the original utterance and into a carrier sentence. This ensured that the only difference between the different pitch type conditions was the pitch accent of the target word, and not the other parts of the sentence. All splices were made at zero-crossings.

Stimuli included two CVC monosyllabic target novel words and 4 CVC monosyllabic distractors (See Table 2 for a full list of stimuli by learning condition). All target words were phonologically distinct. All of the distractor words were animals commonly known by 24-month-olds. Previous knowledge of the words was confirmed by a vocabulary questionnaire completed by the caregiver. Novel words were associated with novel animals designed for this experiment. There are four novel animals, two for each learning condition.

Table 2. *Audio stimuli for Experiment 1.*

Learning Condition	Novel Target Word	Distractors/Fillers
Noncontrastive	*wug* (IPA: /wʌg/)	sheep bear pig
Contrastive	*neem* (IPA: /nim/)	duck bear pig

3.1.3. Design and procedure

We tested participants using a 2x3 mixed design with *contrastiveness* (within-subjects) and *pitch type* (between-subjects) as independent variables. *Contrastiveness* varied in how the target word was presented during learning, either *contrastive* or *noncontrastive* in relation to a preceding discourse element/referent (see Figure 3 for an example of a

noncontrastive learning condition). The *pitch type* variable consisted of three levels. A novel target word could bear a *simple* monotonal (H*) pitch accent, a *complex* bitonal (L+H*) pitch accent, or a *deaccented* pattern.

In a *contrastive* learning condition, the target word was always presented in opposition to one referent in the preceding discourse element (e.g. *Look! There is a bear and pig over there. Oh! There is a bear and a* ***WUG***$_{contrastive}$ *over there.*). In a *noncontrastive* learning condition, the target novel word was presented as presentationally new in relation to the previous discourse referents (e.g. *Look! There is a bear and pig over there. Oh! There is a sheep and a* ***wug***$_{noncontrastive}$ *over there.*) (See Figure 3). In this condition, the novel target was never presented in contrast to a previous referent.

An experimental condition consisted of a two-part Learning Phase and a Test Phase. The Learning Phase consisted of two 12-trial learning blocks (one *contrastive* and one *noncontrastive*). A distinct novel word was presented in each of these learning blocks. Each block included 4 target trials (where the novel word-animal pairing was presented), 4 distractor trials (where a second unnamed novel animal was presented), and 4 filler trials of familiar animals. The Test Phase consisted of 10 trials: 6 test trials and 4 control trials (See Figures 3 and 4). During a control trial, two familiar animals were presented and the toddler was asked to identify one of the referents. During a test trial, the target novel animal was presented with the distractor novel animal. The *proportion of looking time (PLT) to the target* was collected during the Test Phase to assess whether or not the child learned the novel word-animal pairings. We used the same eye-tracking paradigm as Experiment 1.

Block order and novel animal assignments were counter-balanced across conditions. Test trials were randomized and the location of the target item (left or right) was counter-balanced within and across conditions.

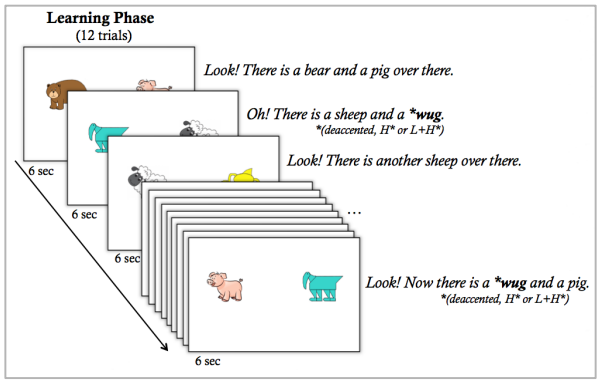


Figure 3: *Example noncontrastive learning phase from Experiment 2. The novel target word 'wug' could bear one of three pitch accent types depending on the condition.*

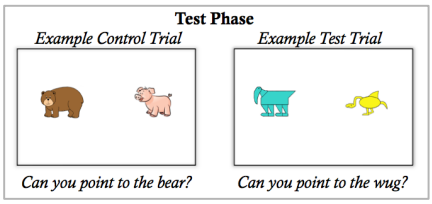


Figure 4*: Example control and test trials from the Test Phase of Experiment 2.*

3.2. Results

Preliminary results exhibit two primary patterns (See Figure 2). The conditions that show substantially more looking to the target novel animal at test are the *contrastive* accented conditions, both *simple* and *complex*. In addition, there is also more looking to the target novel animals in the *noncontrastive-complex* condition.

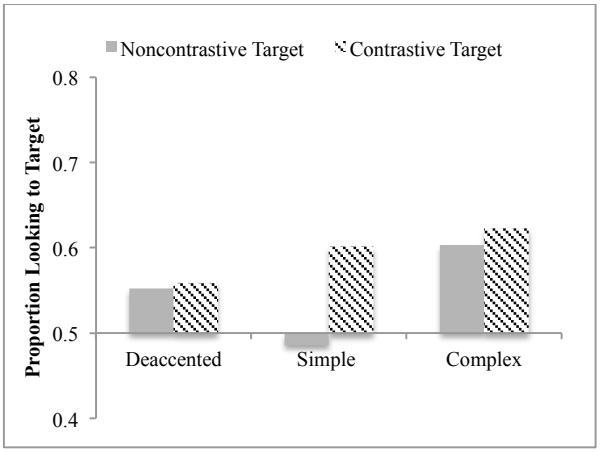


Figure 5: *Bar graph of proportion looking to a contrastive or noncontrastive target novel word referent for each pitch type condition.*

3.3. Discussion

As predicted, the initial pattern of results suggest that contrastiveness, with either a simple or a complex pitch accent, guides more looks to the target novel animal at test. Additionally, the use of a complex pitch accent during the noncontrastive learning situation also directs more looks to the target novel animal than the distractor. These results suggest learning of the novel word-animal association in these conditions, which we predicted to be the most salient during learning. Overall, these data show a preliminary pattern for how intonation and contrastiveness interact to aid in early word learning. Data collection is ongoing.

4. General Discussion

Intonation and information structure both play a role in directing toddler attention and facilitating early word learning. Experiment 1 demonstrates that either newness or the presence of a pitch movement (i.e. only f0 variation) guide 18-month-olds' attention during reference resolution. Future work will test other acoustic correlates of intonation (intensity and duration), and analyze how they interact to guide attention. Preliminary results from Experiment 2 show that contrastiveness and/or a complex pitch accent aid 24-month-olds in learning a novel word. Even with live interactions removed from both methodologies, this set of experiments show the interacting effects of prosody and information structure on attention as well as in successful word learning.

5. Acknowledgements

Special thanks to the subjects, their families, Laura Kertz, Pilar Prieto, Lori Rolfe, Megan Keough, Glenda Molina-Onario, and the Metcalf Infant Research Lab. This work was supported by NIH grant R01HD068501 to JLM.

6. References

[1] Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756-766.

[2] Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development, 21*(4), 779-784.

[3] Prieto, P., & Vanrell, M.M. (2007). Early intonational development in Catalan. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the XVI International Congress of Phonetic Sciences* (ICPhS) (pp. 309-314). Saarbrücken, Germany.

[4] Chen, A., & Fikkert, P. (2007). Intonation of early two-word utterances in Dutch. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the XVI International Congress of Phonetic Sciences* (ICPhS) (pp. 315-320). Saarbrücken, Germany.

[5] Frota, S., & M. Vigário. (2008, August). The intonation of one-word and first two-word utterances in European Portuguese. Paper presented at the *XI International Conference for the Study of Child Language Conference* (IASCL), Edinburgh, Scotland.

[6] Prieto, P., Estrella, A., Thorson, J. & Vanrell, M.M. (2012). Is prosodic development correlated with grammatical development? Evidence from emerging intonation in Catalan and Spanish. *Journal of Child Language*, *39*(2).

[7] Grassmann, S., & Tomasello, M. (2010). Prosodic stress on a word directs 24-month-olds' attention to a contextually new referent. *Journal of Pragmatics*, *42*(11), 3098-3105.

[8] Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development, 10*, 201-224.

[9] Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child development, 67*, 635-645.

[10] Grassmann, S. & Tomasello, M. (2007). Two-year-olds use primary sentence accent to learn new words. *Journal of Child Language, 14*, 23-45.

[11] Watson, D., Gunlogson, C., & Tanenhaus, M. (2008). Interpreting pitch accents in on-line comprehension: H* vs L+H*. *Cognitive Science, 32*, 1232-1244.

[12] Boersma, P. & Weenink, D. (2012). *Praat: A system for doing phonetics by computer*, available at http://praat.org.

[13] Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science, 146,* 668-670.

[14] Rose, S. A., Gottfried, A. W., Melloy-Carminar, P., & Bridger, W. H. (1982). Familiarity and novelty preferences in infant recognition memory: Implications for information processing. *Developmental Psychology, 18*, 704-713.

[15] Katz, J. & Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language, 87*(4), 771-816.

Prominence perception in and out of context

Rory Turnbull, Adam J. Royer, Kiwako Ito, Shari R. Speer

Department of Linguistics, Ohio State University, Columbus, OH, USA

turnbull@ling.osu.edu

Abstract

The perception of prosodic prominence is known to be influenced by several distinct factors. In this study, we investigated the role of context, both global and local, in the prominence judgements of naïve listeners. Monolingual English listeners marked where they heard prominence on pairs of two-word phrases (e.g. *blue ball*, *green drum*). Stimuli varied in whether or not the first phrase implied a contrastive focus on the second phrase. We found clear evidence of a hierarchy of prominence across pitch accent types: L+H* > H* > !H* > unaccented. Additionally, we found that contrast status only affected prominence markings when the participants were made explicitly aware of the discourse context and were instructed to imagine themselves physically present to observe the conversation. This effect of global context suggests that information structure cannot be reliably interpreted in the absence of an established discourse context. Taken together, these results suggest that naïve listeners are sensitive to prominence differences at levels corresponding to categorical annotations. Perception of a word's relative prominence was consistently influenced by phonetic and phonological factors, while pragmatic factors (such as contrast-evoking context) required more elaborate plausibility manipulations in order to affect prominence perception.

Index Terms: prominence, perception, discourse, focus, contrast

1. Introduction

Intonational phonology assumes a strictly layered hierarchical prominence organization [1]. Words with pitch accent are more prominent than words without pitch accents, and the nuclear pitch accent—in American English, the final one in the utterance—is more prominent than other pitch accents [2]. For the most part, these assumptions have gone largely unchallenged.

Previous studies [3, 4, 5] have shown that the perception of prosodic prominence is based on the listeners' expectations in addition to properties of the signal. For instance, [3] found that information structural considerations, such as anticipating a contrastive focus, can influence perception of prominence such that a semantically or pragmatically salient word can be perceived as prominent even in the absence of acoustic or phonological salience. The scientific goal of this research project is to establish what factors underlie the perception of prominence—factors such as acoustics, phonology, lexico-syntactic phrasing, pragmatic context—and how these factors interact. Additionally, we seek to describe the constraints on the use of these factors predictively in speech comprehension [6].

Several investigations of prosodic prominence have obtained judgements from naïve, untrained listeners in tasks where they are asked to mark prominences on a transcript of aurally presented speech [4, 7, 8, 9, 10, 11]. For longer speech samples, such a task can involve considerable memory load. In the current study, we used a similar metalinguistic judgement task where native speakers of American English were asked to indicate which words in a short phrase sounded prominent. In particular, we examined the role of discourse-level factors and paradigmatic phonological structure in influencing listeners' judgements about which words are prominent.

2. Methods

2.1. Materials

Materials were selected from a ToBI-annotated corpus of spontaneous speech collected from naïve speakers instructing Christmas tree decoration [12]. Twenty-two utterances consisting of adjective-noun combinations, each denoting a particular tree ornament (e.g. red house), were extracted from one female speaker. Utterances either had the pitch accent tune [H* !H*] or [L+H* 0] on the adjective and noun respectively, where 0 represents an unaccented word. The first of these tunes can be considered a 'neutral' prosody, while the second is commonly associated with contrastive focus on the adjective. The stimulus phrases involved eight different adjectives (*beige*, *blue*, *brown*, *clear*, *gray*, *green*, *navy*, and *orange*) and six different nouns (*ball*, *bell*, *candy*, *drum*, *house*, and *onion*).

2.2. Procedure

In each trial, a pair of two-word phrases were displayed on screen (see Figure 1), and after 250ms, the recordings were presented over headphones with 500ms of silence between the two. The participant's task was to highlight, using a button box, which words out of the four on screen sounded prominent. They could highlight as many or as few words as they liked, and there was no time constraint on their choices.[1]

Each pair of utterances was played twice, allowing the participants to double-check their marking before proceeding to the next trial. Each phrase pair either contained a lexically established contrast between first and second utterances (e.g. *blue ball*, *green ball*), and thus an implied contrastive adjective focus on the second adjective, or no contrast (e.g. *red house*, *green ball*). The pitch accent tune types in the first and second utterances were fully crossed, leading to 48 trials. Thus, this offered a total of 192 possible prominence markings, with an equal number of presentations of each pitch accent type; see Table 1.

Additionally, the pragmatic context of the phrases was manipulated between subjects. In the 'monologue' condition, the two phrases were presented one after another, with no intervening material during the 500ms interval. This way, the audio

[1] Participants also completed three other similar tasks; data from these other tasks are not analyzed in the current paper.

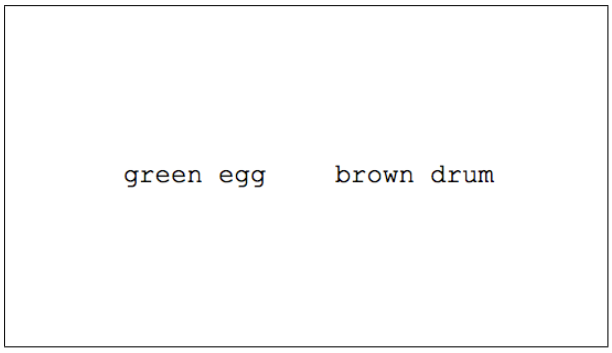


Figure 1: *Example of visual presentation of orthographic representation of stimuli. Once a word was selected as prominent, it turned red.*

sounded like a monologue, a person saying a string of phrases. In the 'plain dialogue' condition, the two phrases were presented with a 'connective' utterance intervening between them. These connectives were extracted from the speech of the male Christmas tree decorator from the same corpus, and consisted of short utterances such as "okay, next", "alrighty, next" and similar. In this condition, there was still a total of 500ms of silence between the first and second phrases.

Finally, the 'narrative dialogue' condition was exactly the same as the plain dialogue condition with the exception of the instructions to the participants. In this condition, participants were made fully aware of the provenance of the recordings and the purpose of the utterances (*viz.*, decorating a Christmas tree). They were presented with a short extract of the conversation between the instructor and the decorator, the visual materials used to elicit speech (see Figure 2), a diagram of the experimental apparatus and procedure, and the physical location the recordings were made. At the beginning of each block, they were instructed to imagine that they were physically present with both participants in the dialogue they heard, and to mark the words that sounded "important to the conversation".

The experiment thereby constituted a 2×2×2×3 fully crossed design, with the relevant factors being pitch accent sequence of the first phrase ([H* !H*] vs [L+H* 0]), pitch accent sequence of the second phrase ([H* !H*] vs [L+H* 0]), contrast status of the pair of phrases (contrastive or non-contrastive), and pragmatic context of the utterances (monologue vs plain dialogue vs narrative dialogue). Pragmatic context was manipulated between subjects; all other factors were within subjects.

2.3. Participants

A total of 125 monolingual speakers of American English participated in the experiment for either partial course credit or a

First phrase		Second phrase	
Adj	Noun	Adj	Noun
[H*	!H*]	[H*	!H*]
[H*	!H*]	[L+H*	0]
[L+H*	0]	[L+H*	0]
[L+H*	0]	[H*	!H*]

Table 1: *Summary of pitch accent sequences presented to participants.*

Figure 2: *Example visual materials from the original elicitation experiment [12] that was shown to participants in the narrative dialogue condition to demonstrate the illocutionary force of the extracts they will listen to. Depicted in this picture is a brown egg.*

$10 payment. 43 participated in the monologue condition, 42 in the plain dialogue condition, and 40 in the narrative dialogue condition.

2.4. Analysis

The results were analyzed using a mixed effect logistic regression model to predict whether or not the *second adjective* (the potentially putatively focused element) was marked as prominent. The model had fixed effects of the pitch accent sequence of the first phrase ([H* !H*] vs [L+H* 0]), pitch accent sequence of the second phrase ([H* !H*] vs [L+H* 0]), contrast status of the pair of phrases (contrastive or non-contrastive), and the condition (monologue vs plain dialogue vs narrative dialogue). All of the fixed effects were coded with sum contrasts, with the exception of condition which used treatment contrasts with the monologue condition as baseline. Additionally, since a number of participants reported during debriefing that they attended to the article *a* at the beginning of some phrases, the presence or absence of the article at the beginning of the first phrase and the second phrase were also added as fixed effects. Additionally, all possible three-way interactions between contrast status, 1st phrase pitch accent sequence, 2nd phrase pitch accent sequence, and condition were included, except for any interactions involving both of the pitch accent sequences.[2] Random intercepts of the second adjective lexical identity, second noun lexical identity, and subject identity were used, and a random slope for contrast status by subject.

In selecting the stimuli, we have relied upon ToBI transcriptions of the target phrases to classify them into groups for analysis. However, different coding systems exist, and it is theoretically possible to achieve different results from the use of different systems. In order to minimize this possibility, we sought independent phonetic evidence for classifying our stimuli, and so each stimulus phrase underwent acoustic analysis. For each phrase, measurements were made on both the adjective and the

[2]For clarity, the interaction term formula, in R-style syntax, was `contrast * condition * (1stpitchaccent + 2ndpitchaccent)`.

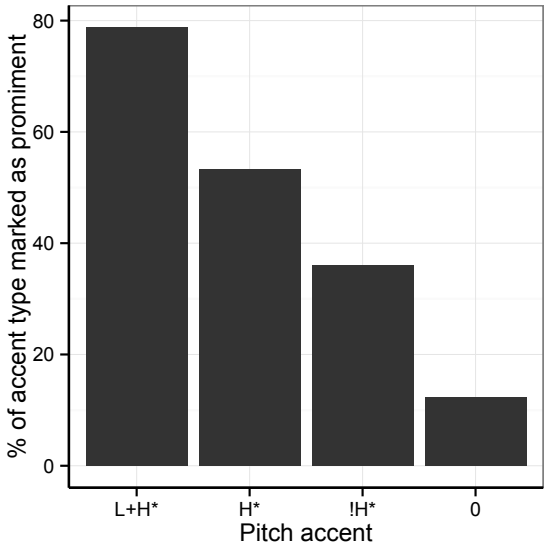


Figure 3: *Percentage of prominence endorsements on each type of pitch accent presented. 0 represents unaccented words.*

noun to extract the word duration, the vowel duration, the peak f0, the mean f0 within the vowel, and two measures of spectral tilt. The spectral tilt measures were the difference between the mean intensity of two different spectral bands, either 2kHz in bandwidth (i.e. 0-2kHz minus 2-4kHz) or 4kHz in bandwidth (i.e. 0-4kHz minus 4-8kHz). Additionally, two relative measures were taken which relate the adjective to the noun: the slope from the adjective peak f0 to the noun peak f0; and the slope from the adjective vowel mean f0 to the noun vowel mean f0. Each of these fourteen acoustic variables was entered as a predictor into a regression tree analysis [13] predicting the pitch accent sequence of the phrase (either [H* !H*] or [L+H* 0]). The resulting tree was pruned to minimize the cross-validation standard error, and of the predictor variables, only the f0 peak of the noun was found to act as a substantial cue to pitch accent. The tree correctly classified the phrases' pitch accent sequences 86.4% of the time (chance: 50%). This cue takes advantage of the relatively large f0 difference between !H* nouns and unaccented noun in our stimuli set. Therefore, with a small degree of error, it is possible to classify a phrase as [L+H* 0] if the noun peak f0 is low, and as [H* !H*] if the noun peak f0 is high.

A second mixed-effects logistic regression model was constructed, identical to the first except with the noun peak f0 measurements in place of the pitch accent transcriptions. This acoustically-based model allowed for a comparison of the phonological transcription with the acoustic details in their ability to account for the observed data. This comparison between the pitch-accent-based model and the acoustics-based model ensured that any observed effects were not simply artifacts of the transcription scheme or idiosyncrasies of the stimuli.

3. Results

Figure 3 depicts the overall prominence marking counts for each kind of pitch accent presented to the participants. This figure collapses over word position and condition; i.e., it depicts all of the prominence markings on all of the words in all of the conditions, split by pitch accent type. As can be seen, the endorsement rates depict a clear hierarchy of prominence, bolstering previous claims that the distinctions between these pitch accents in English (particularly between H* and L+H*) are mainly ones of prominence [14, 15, 16]. We now turn to the modeling results.

Effect	β	z	p
Intercept	0.922	2.228	0.026
PA2	1.676	10.462	< 0.001
Art1	-0.509	-6.220	< 0.001
Art2	0.358	2.232	0.026
Contrast $\times$ PA2	0.499	2.192	0.028
Contrast $\times$ PA2 $\times$ narrative	0.853	2.646	0.008

Table 2: *Summary of significant effects for the pitch-accent-based model. PA2 = pitch accent sequence of the second phrase; Art1 = presence of article in the 1st phrase; Art2 = presence of article in the 2nd phrase.*

The significant fixed effects of the pitch-accent-based model are summarized in Table 2. A significant effect of the pitch accent sequence of the second phrase was observed, such that adjectives with L+H* were more likely to be endorsed as prominent (78.9%) than those with H* (52.8%), as expected given previous research on accent type prominence (e.g. [14]). Two effects of article presence were observed: when the first phrase bore an article, the second adjective was significantly less likely to have a prominence marking (62.4%) than when the first phrase did not have an article (68.5%); similarly, when the second phrase bore an article, the second adjective was more likely to have a prominence marking (70.7%) than when there was no article (59.1%).

Additionally, an interaction between the pitch accent sequence of the 2nd phrase and contrast status was observed. This interaction is depicted in Table 3; when the sequence of phrases led to an implied contrastive focus on the adjective (e.g. *blue ball*, *green ball*), more prominence markings were observed on adjectives with a L+H* pitch accent. Interestingly, this effect of implied contrast was not observed on H* adjectives, suggesting that listeners did not interpret the [H* !H*] phrases with contrastive focus, despite the context.

Finally, and most crucially, a three-way interaction between contrast status, pitch accent, and pragmatic condition was observed. In contrastive contexts in the narrative dialogue condition, more endorsements were observed on L+H* adjectives when compared to the monologue condition. Essentially, the contrast effect for L+H* words (the two-way interaction mentioned in the preceding paragraph) is even stronger in the narrative dialogue condition. Recall that in the narrative dialogue condition, participants were encouraged to imagine themselves actually being present as the conversation took place. This effect is visualized in the right panel of Figure 4; note that in the other conditions, where the participants were not made aware

	H*	L+H*
Non-contrastive	54.1%	75.9%
Contrastive	51.5%	82.0%

Table 3: *Endorsement rates for second adjectives with different pitch accents in different contrast conditions.*

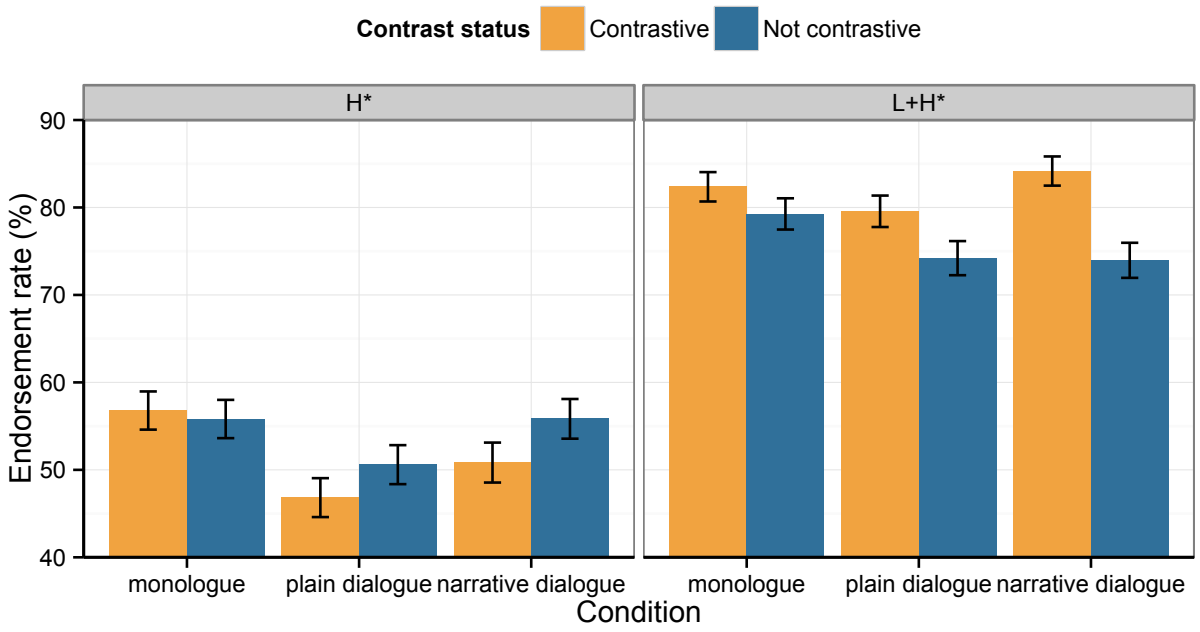


Figure 4: *Percentage of prominence markings on the second adjective of the phrase pair, broken up by pitch accent, pragmatic condition, and contrast status. Bars indicate standard error.*

of the discourse intent, the effect is much smaller.

The results of the acoustics-based model was comparable, in that the narrative dialogue condition sees contrast status effects that are not observed for the other conditions. See Table 4 for a summary of main effects. A likelihood ratio test comparing the two models confirmed that they did not differ in data likelihood ($p > 0.5$).

Effect	β	z	p
Intercept	9.721	8.835	< 0.001
Nf0	-0.040	-8.652	< 0.001
Art1	-0.440	-5.310	< 0.001
Contrast $\times$ Narrative	2.658	2.270	0.023
Contrast $\times$ Nf0 $\times$ Narrative	-0.012	-2.261	0.024

Table 4: *Summary of significant effects for the phonetic model. Nf0 = Noun peak f0; Art1 = presence of article in the 1st phrase.*

4. Discussion and conclusion

In addition to the expected effect where L+H* adjectives are marked as more prominent than H* adjectives, unexpected effects of article presence on the first and second phrases were observed. The trend in these effects appears to be that an article at the beginning of a phrase makes the adjective appear more prominent. During debriefing, some participants noted that they thought the article signaled "something really important coming up", particularly when it was pronounced as [eɪ]. This intuition is supported by a model of the entire dataset, collapsing across conditions and phrase position, predicting prominence marking based on article presence and word class. The model revealed that in phrases that follow an article, adjectives are more likely to be marked as prominent ($\beta = 0.994$, $z = 16.884$, $p < 0.001$), while nouns are less likely to be marked ($\beta = -1.072$, $z = -13.859$, $p < 0.001$).

The pitch accent prominence hierarchy depicted in Figure 3 is particularly striking. However, care must be taken in its interpretation, since in this study all of the L+H*s and H*s were associated with phrase-initial adjectives, and all of the !H*s and unaccented words were phrase-final nouns. Nevertheless, our findings are consistent with the expected patterns of prominence in American English (e.g. [17]). The consistency of these findings also illustrate the fact that prominence is indicated in the same locations by the ToBI annotators of these stimuli, by our naïve participants, and by the results of the acoustic analysis.

Finally, our main result is an effect of discourse context evoked by elaborately illustrated task instructions. Only when participants were made fully aware of the intent of the discourse and instructed to imagine themselves as being physically present in the conversation was an effect of contrast status observed. This is to say, only when participants are able to construct a common ground with the interlocutors does their prominence perception reflect information-structural concerns. Prosodic prominence on its own can be perceived in extremely impoverished contexts (the monologue condition), but the information-structural notion of contrast requires an established discourse context before perception and interpretation is possible.

5. Acknowledgements

For comments: Speerlab discussion group. For partial funding: Buckeye Language Network Undergraduate Research Award.

6. References

[1] E. O. Selkirk, *Phonology and syntax: the relationship between sound and structure*. Cambridge, MA: MIT Press, 1986.

[2] D. R. Ladd, *Intonational Phonology*, 2nd ed. Cambridge: Cambridge University Press, 2008.

[3] J. Bishop, "Information structural expectations in the perception of prosodic prominence," in *Prosody and Meaning*, G. Elordieta and P. Prieto, Eds. Berlin: Mouton de Gruyter, 2012.

[4] J. Cole, Y. Mo, and M. Hasegawa-Johnson, "Signal-based and expectation-based factors in the percetion of prosodic prominence," *Laboratory Phonology*, vol. 1, pp. 425–452, 2010.

[5] C. Smith, "French listeners' perceptions of prominence and phrasing are differentially affected by instruction set," *Proceedings of Meetings on Acoustics*, vol. 19, p. 060191, 2013.

[6] K. Ito and S. R. Speer, "Anticipatory effects of intonation: Eye movements during instructed visual search," *Journal of Memory and Language*, vol. 58, no. 2, pp. 541–573, 2008.

[7] J. Buhmann, J. Caspers, V. J. van Heuven, H. Hoekstra, J.-P. Martens, and M. Swerts, "Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus," in *Proceedings of LREC*, 2002.

[8] J. Cole, Y. Mo, and S. Baek, "The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech," *Language and Cognitive Processes*, vol. 25, no. 7, pp. 1141–1177, 2010.

[9] K. Kohler, "The perception of prominence patterns," *Phonetica*, vol. 65, no. 4, pp. 257–269, 2008.

[10] B. M. Streefkerk, L. C. W. Pols, and L. F. M. ten Bosch, "Prominence in read aloud sentences, as marked by listeners and classified automatically," *Proceedings of the Institute of Phonetic Sciences*, vol. 21, pp. 101–116, 1997.

[11] M. Swerts, "Prosodic features at discourse boundaries of different strength," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 514–521, 1997.

[12] K. Ito and S. R. Speer, "Using interactive tasks to elicit natural dialogue," in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleink, N. Richter, and J. Schließer, Eds. Berlin: Walter de Gruyter, 2006, pp. 227–257.

[13] T. M. Therneau and E. J. Atkinson, "An introduction to recursive partitioning using the RPART routines," Section of Biostatistics, Mayo Clinic, Rochester, MN, Tech. Rep. 61, 1997.

[14] S. Calhoun, "The theme/rheme distinction: Accent type or relative prominence?" *Journal of Phonetics*, vol. 40, pp. 329–349, 2012.

[15] D. R. Ladd, "Metrical representation of pitch register," in *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, J. Kingston and M. E. Beckman, Eds. Cambridge: Cambridge University Press, 1990.

[16] D. R. Ladd and R. Morton, "The perception of intonational emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[17] D. Büring, "On D-trees, beans and B-accents," *Linguistics and Philosophy*, vol. 26, pp. 511–545, 2003.

Table of Contents

The (sorted) Authors index