A multi-level approach to speech rate in British English: towards an analysis-by-synthesis method

Cyril Auran & Caroline Bouzon

CNRS UMR 8163, Laboratoire Savoirs Textes Langage
Université Lille 3 - Charles-de-Gaulle, Lille, France
{cyril.auran, caroline.bouzon}@univ-lille3.fr

Abstract

This paper provides a detailed account of the durational differences induced at different structural levels (inter-silence segments, rhythmic units, syllables, syllabic constituents, phones) by changes in speech rate from normal to slow speech in read British English. Using the data described in this study, we present preliminary results concerning a regression tree model predicting phone durations in slow speech with an average precision of 16 ms in slow speech.

1. Introduction

The research described here fits in a more global project concerning the modelling of orthogonal prosodic dimensions in discourse ([2], [16]). These orthogonal prosodic dimensions have been shown to play an important part in the marking of the topical structure of discourse, for instance with increased pitch level and range signaling the beginning of a new topic ([9], [12], [22]), and lower pitch level and range ("final lowering") at the end of topics ([17], [22]).

More specifically, our current research focuses on speech rate in British English. With relation to discourse structure, slower speech rate is involved in the marking of topic beginnings ([15], [19]) and topic ends together with final lengthening ([19], [22]).

This paper proposes the construction of a two-way model capable of predicting the speech rate category (normal *vs*. slow) of annotated and time-aligned utterances provided as input; conversely, this model is to be used in the synthesis of artificially slowed down utterances from "normal speech rate" input, thus paving the way for the integration of speech rate modelling in automated discourse segmentation and discourse annotated speech synthesis. This paper tackles this second aspect with preliminary results from a Classification And Regression Tree (CART) model.

2. Corpus and method

2.1. Corpus

Three British native speakers (two female and one male, henceforth speakers F, I and J) were asked to read a series of sentences in an anechoic room. In the absence of any particular instructions to the speakers, these utterances belong in the "normal speech rate" category (Phase 1). After each sentence, a message was displayed, indicating whether the sentence had been successfully recognized by a decoy automatic speech recognition system or if it had to be read again slower. These "slower" sentences constitute our "slow speech rate" Phase 2 items. In the end, 26 sentences were retained in both the normal (Phase 1) and slower (Phase 2) conditions for speakers F and I, and 22 for speaker J, providing a total of 146 utterances.

2.2. Annotation

Using Praat [5], the data was manually transcribed into phonemes and aligned at different levels: phones, syllables and words. Stress was also manually marked (distinguishing unstressed, primary and secondary stressed syllables). From this annotation, we automatically derived multi-level TextGrids following the Aix-MARSEC methodology [3]. The data is therefore aligned at the following levels: phonemes, syllabic constituents and syllables (maximal onset principle), stress feet [1], Narrow Rhythm Units/ Anacruses [13] and Inter-Silence Segments (ISS).

Our study involves two different rhythmic models (Abercrombie's and Jassem's), the main difference between the two lying in the definition of the rhythmic unit: in Abercrombie's model [1], the stress foot starts with a stressed syllable and includes all the following unstressed syllables up to the next stressed syllable (without including it); however, in Jassem's model [13], the Narrow Rhythm Unit/ NRU starts with a stressed syllable and also includes the following unstressed syllables but only up to the end of the word, all the other unstressed syllables belong to the Anacrusis/ ANA (see [6] and [7] for a comparison of the two models).

The following information was then automatically derived at all levels, thus leading to a total of 173 parameters:
- raw and normalized durations (using the z-score method) of these different units,
- number of sub-constituents (complexity),
- position within higher constituents,
- F0 and global intensity means and standard deviations.

3. Results

3.1. Global results

The most important differences between Phase 1 and Phase 2 lie in an increased number of pauses in Phase 2 and in a stark difference in normalized mean phoneme durations. Indeed, for each speaker, there are significantly more pauses in Phase 2 than in Phase 1, ranging from 0 to 3 in Phase 1 to 0 to 6 in Phase 2. This difference is very significant for speakers F and I (Fisher's exact tests p=5.319e-07 and p=5.566e-06 respectively), and less important but still significant for speaker J (Fisher's exact test p=0.001493).

Normalised mean phoneme durations are significantly (ANOVA F(1)=185.431 p<2.2e-16) less important in Phase 1 (mean=0.113) than in Phase 2 (mean=0.652).

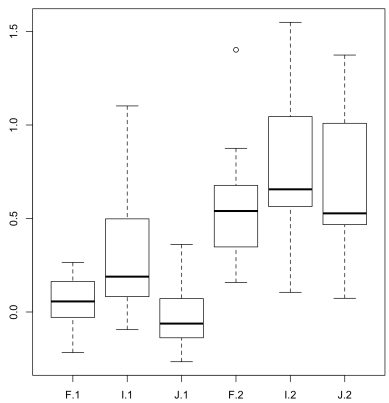

Figure 1: M*ean normalized phoneme durations by Phase and by Speaker (global results).*

Differences were also found between speakers for a given phase (ANOVA F(2)=16.598 p=3.329e-07), with speaker I displaying significantly higher values in both phases.

No significant interaction between Phase and Speaker was observed (ANOVA F(2)=1.479 p=0.2313).

3.2. Narrow Rhythm Units

The duration of units within Jassem's model is influenced by Phase, Speaker, Unit type (ANA *vs*. NRU) and Position in the ISS. In Phase 1, ANAs are shorter than NRUs, but, for speakers I and J, this difference no longer is significant in Phase 2.

At the level of the ISS, there is a clear lengthening of the final NRU in both Phases and for all Speakers except J in Phase 2. This lengthening is localized on the last NRU and does not affect the penultimate (Welsh two sample t-tests; table 1). There is globally no effect on the duration of initial NRUs, or at least no clear tendency.

	Phase 1	Phase 2
Speaker F	**0.499** > -0.032	**0.798** > 0.384
Speaker I	**0.743** > 0.234	**1.192** > 0.517
Speaker J	**0.177** > -0.047	**0.767** = 0.577

Table 1. *Mean normalised durations of final* vs. *non-final NRUs (1st NRU excluded).*

3.3. Stress feet

We observe a strong influence of Phase, Speaker, Unit type (silent ictus/ SI *vs*. stress foot) and Position within the ISS. Differences between the duration of SIs and feet are not identical to those in Jassem's model. Indeed, SI are shorter in both phases for speaker F; for speaker I, SI are shorter than feet in Phase 1 but not in Phase 2 and for speaker J, there is no difference, independently of Phase.

Regarding sensitivity to Position in the ISS, the results are similar to those obtained with Jassem's model: clear lengthening of the final foot for all speakers with no particular effect on the penultimate (Welsh two sample t-tests; table 2). There is no clear effect on the initial foot.

	Phase 1	Phase 2
Speaker F	**0.514** > -0.0.71	**0.824** > 0.353
Speaker I	**0.743** > 0.165	**1.197** > 0.459
Speaker J	**0.193** > -0.106	**0.834** > 0.451

Table 2. *Mean normalised durations of final* vs. *non-final feet (1st foot excluded).*

3.4. Syllables

We observe a strong influence of Phase, Speaker, Stress, and Relative position within the ISS, with interactions between Speaker and Phase, Phase and Relative position within the ISS and between Stress and Relative position within the ISS.

In Phase 1, primary stressed syllables are on the whole significantly longer than both unstressed and secondary stressed syllables. In Phase 2, this difference is less significant for speakers F and I, and no longer significant for speaker J (LRM Student's t-tests; table 3).

We measure a clear lengthening of the ISS final (unstressed and stressed) syllable for all speakers in both phases. This lengthening is localised on the last syllable with, in some cases, a shortening of the penultimate syllable (Welsh two sample t-tests comparing the penultimate and final syllables with the other (un)stressed syllables; table 4).

	Phase 1	Phase 2
Speaker F	**1 > 0, 2** [0.186 > -0.057 > -0.046]	**1 > 0, 2** [0.612 > 0.396 > 0.2423]
Speaker I	**1 > 0, 2** [0.389 > 0.163 > -0.007]	**1 > 0, 2** [0.859 > 0.64 > 0.297]
Speaker J	**1 > 0, 2** [0.044 > -0.091 > -0.331]	**1 = 0, 2** [0.704 = 0.566 = 0.109]

Table 3. *Mean normalised durations of stressed (1,2) and unstressed (0) syllables.*

	Penultimate syllable	Last syllable
F - Ph 1	U: **-0.434** < -0.234	U: **1.05** > -0.25
	S: **= 0.22** = 0.178	S: **1.319** > 0.187
F - Ph 2	U: **-0.158** < 0.075	U: **1.353** > 0.017
	S: **0.531** = 0.606	S: **1.448** > 0.576
I - Ph 1	U: **-0.345** < -0.023	U: **1.472** > -0.04
	S: **0.389** = 0.389	S: **1.088** > 0.391
I - Ph 2	U: **0.124** = 0.289	U: **1.405** > 0.255
	S: **0.531** < 0.747	S: **1.937** > 0.669
J - Ph 1	U: **-0.511** < -0.173	U: **0.558** > -0.192
	S: **0.103** = 0.068	S: Not enough data
J - Ph 2	U: **0.158** = 0.383	U: **1.151** > 0.352
	S: **0.446** < 0.7	S: **1.772** > 0.636

Table 4. *Mean normalised durations of penultimate and last unstressed (U) and stressed (S) syllables.*

Things are less clear at the beginning of ISS and speakers seem to use different strategies: in Phase 1, there is a shortening of the first stressed syllable for F and J (Welsh two sample t-tests p=0.03788 and p=0.0003202), not for I; in Phase 2, there is a slight lengthening of the first syllable for I (unstressed and stressed syllables; Welsh two sample t-tests p=0.00478 and p=0.01207) and J (unstressed syllables only; Welsh two sample t-test p=0.006379).

3.5. Syllabic constituents

Here too, we observe a strong influence of Phase, Speaker, Stress, and Relative position within the ISS, with interactions between Speaker and Phase, Phase and Relative position within the ISS and between Stress and Relative position within the ISS.

Type (Onset *vs*. Nucleus *vs*. Coda) also plays a significant role, with interactions with Phase, Stress and Relative position within the ISS, but no interaction with Speaker.

The effects of final lengthening are significant for both nucleus and coda in final position in both Phase 1 and Phase 2 (Welsh two sample t-tests; table 5).

In penultimate position, the effects of final lengthening are significant for the nucleus both in Phase 1 (stressed) and in Phase 2 (unstressed and stressed). However, in Phase 1, the onset displays significant lengthening for unstressed syllables, but shortening for stressed syllables; in Phase 2, it displays no significant difference with other onsets elsewhere in the ISS (Welsh two sample t-tests; table 6).

In ante-penultimate position, the onset displays shortening for stressed syllables in Phase 1, and no significant difference in all other contexts (Welsh two sample t-tests; table 7).

	Nucleus	Coda
Ph 1	Unstressed: **1.827** > -0.19	Unstressed: **1.661** > -0.092
	Stressed: **2.269** > 0.136	Stressed: Not enough data
Ph 2	Unstressed: **2.5** > 0.378	Unstressed: **1.747** > 0.527
	Stressed: **2.823** > 0.58	Stressed: **1.887** > 0.574

Table 5. *Mean normalised durations of nuclei and codas in final position.*

	Onset	Nucleus
Ph 1	Unstressed: **0.367** > 0.011	Unstressed: **0.497** > -0.253
	Stressed: **-0.507** < 0.385	Stressed: **0.631** = 0.132
Ph 2	Unstressed: **0.465** = 0.328	Unstressed: **1.0698** > 0.238
	Stressed: **0.51** = 0.838	Stressed: **2.144** > 0.515

Table 6. *Mean normalised durations of onsets and nuclei in penultimate position.*

	Onset
Ph 1	Unstressed: **0.113** = 0.001
	Stressed: **-0.045** < 0.388
Ph 2	Unstressed: **0.371** = 0.316
	Stressed: **1.024** = 0.82

Table 7. *Mean normalised durations of onsets in ante-penultimate position.*

3.6. Phones

There is a strong influence of Phase, Speaker, Stress and Relative position in the ISS. In Phase 1, the effect of stress is quite clear: for all speakers, primary stressed phones are longer than unstressed and secondary stressed phones (LM Strudent's t=2.577 p=0.01). In Phase 2, we observe a loss of distinctiveness (LM Strudent's t=1.939 p=0.05259), in particular for speakers I and J.

As could be expected from the results given above, there is a clear lengthening of final phones for all speakers and for both phases; in most cases, this lengthening is localised on the last two phones, with no effect at all on the antepenultimate phone.

	Penultimate phone	Last phone
F - Ph 1	U: **0.557** > -0.195	U: **1.895** > -0.137
	S: **0.26** = 0.137	S: **2.244** > 0.138
F - Ph 2	U: **0.854** > 0.167	U: **2.232** > 0.267
	S: **0.566** = 0.595	S: **2.25** > 0.595
I - Ph 1	U: **0.86** > 0.048	U: **2.341** > 0.103
	S: **-0.433** = 0.387	S: **2.336** > 0.38
I - Ph 2	U: **1.127** > 0.426	U: **2.107** > 0.527
	S: **1.471** > 0.675	S: **2.331** > 0.717
J - Ph 1	U: **0.235** > -0.179	U: **0.978** > -0.153
	S: **-0.2975** = 0.053	S: Not enough data
J - Ph 2	U: **0.803** = 0.435	U: **1.77** > 0.467
	S: **1.274** = 0.633	S: **2.526** > 0.64

Table 8. *Mean normalised durations of penultimate and last unstressed (U) and stressed (S) phones.*

We observe no significant effect of the initial position on the duration of phones.

4. Predictive model

4.1. Method

The statistical analyses evoked in section 3 consistently signal complex interactions between such parameters as Speaker, Phase, Stress or Position within the ISS. Classical regression techniques, such as linear models, are not easily interpretable with such complex patterns. We therefore decided to resort to Classification And Regression Trees (CARTs; [8]), which have the advantage to (1) select the most significant parameters, (2) provide "honest" estimates of their performance, (3) allow both categorical and continuous features to be considered and (4) allow straightforward human exploration and interpretation of the results (see [18] for a detailed explanation of this method in segmental duration modelling).

The CARTs were generated within the R environment using the rpart package ([20]). Tree over-fitting was controlled through cost-complexity pruning and cross-validation over the entire data set (thresholds provided below), a method proposed in [8] which minimises the variance of the prediction error as a function of tree length.

Speakers displaying different strategies, we decided to implement this method by speaker and by phase. We further split between Abercrombie's and Jassem's models, thus allowing a comparison of the two models in terms of phone duration predictions. In this preliminary stage, we restricted the initial parameter set to Stress, Rhythmic unit type and Number of constituents and Position (from beginning and from end) in the ISS, the rhythmic unit and the syllable.

4.2. Results

In this section, we provide preliminary results obtained for Speaker F in both Phase 1 and 2, with a comparison of Abercrombie's and Jassem's contributions.

Figures 2 to 5 provide the optimised Phase 1 and Phase 2 CARTs obtained with Abercrombie's and Jassem's models through cost-complexity pruning with respective thresholds of 0.021, 0.024, 0.023, 0.023.

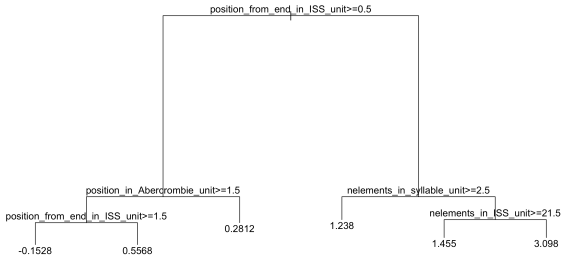

Figure 2: *CART for Speaker F (Ph 1, Abercrombie's model).*

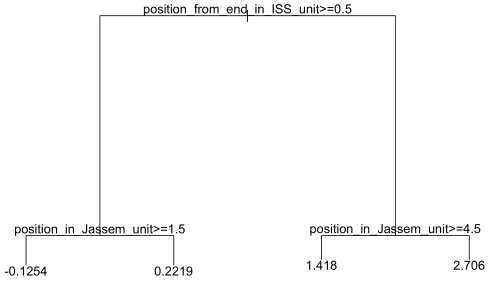

Figure 3: *CART for Speaker F (Ph 1, Jassem's model).*

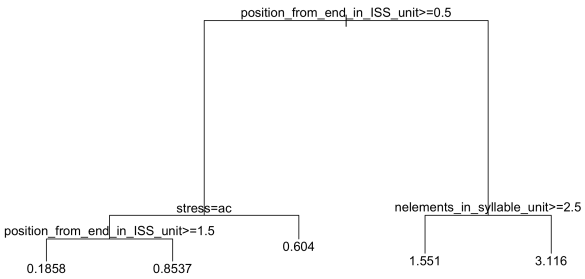

Figure 4: *CART for Speaker F (Ph 2, Abercrombie's model).*

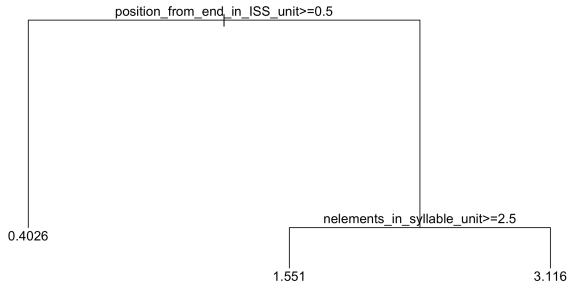

Figure 5: *CART for Speaker F (Ph 2, Jassem's model).*

As can be seen in table 9, Both models provide a mean absolute error of about 13 ms in phase 1 and 16 ms in phase 2. However, CARTs generated using Jassem's model display lower complexity (fewer splits) in both phases.

	Model	MAE	split complexity
Phase 1	Abercrombie	12.5	5
	Jassem	13	3
Phase 2	Abercrombie	15.5	4
	Jassem	16	2

Table 9. *CART mean absolute error and split complexity for speaker F.*

5. Conclusions and perspectives

The statistical results given in section 3 confirm classical findings about "slow speech", with an increased number of pauses and the lengthening of utterances (see [21] for a recent and exhaustive description). Another result is the clear final lengthening at all observed linguistic levels and in both types of speech rate. We show that this lengthening is localised on the final rime (among others [4], [19]).

Another interesting result is the relative loss of distinctiveness for primary stress in Phase 2 (*contra* [11] and [14]). This is to be linked with the more important lengthening of unstressed phones in Phase 2, which also accounts for the loss of distinctiveness observed between ANAs and NRUs and between SIs and Feet.

The models presented in section 4 offer a mean absolute error of about 13 ms in phase 1 and 16 ms in phase 2; these values are analogous to those given in [18] with a similar method. A finer-grained analysis of the distribution of errors depending on phone type, stress and position needs to be carried out and confronted with the perception of synthetic utterances produced by algorithms based on these models.

The cost-effectiveness of CARTs based on Jassem's model can easily be explained by the obvious interaction of NRUs and ANAs with word boundaries: in this case, the use of these rhythmic units somehow amounts to integrating word-related features in the model. It would thus be interesting to analyse word-related phenomena more thoroughly and to feed this information in CARTs based on Abercrombie's model.

Our study globally shows that speakers tend to use different strategies when changing their speech rates; further developments would beneficially include more speakers and other speaking styles.

6. References

[1] Abercrombie, D. "Syllable quantity and enclitics in English", in Abercrombie, D., Fry, P., MacCarthy, N. and Trim, J. [eds]: In Honour of Daniel Jones, 216-222, London: Longman, 1964.

[2] Auran, C. "Prosodie et anaphore dans le discours en anglais et en français : cohésion et attribution référentielle", PhD dissertation, Université de Provence, 2004.

[3] Auran, C., Bouzon, C. and Hirst, D. J. "The Aix-MARSEC project: an evolutive database of spoken English", in Proceedings of the Second International Conference on Speech Prosody:561-564, 2004.

[4] Beckman, M.E. and Edwards, J. "Lengthenings and shortenings and the nature of prosodic constituency", Laboratory Phonology I:152-178, 1990.

[5] Boersma, P. "Praat. A system for doing phonetics by computer", Glot International, 5 (9/10):341-345, 2001.

[6] Bouzon, C. "Rythme et structuration prosodique en anglais britannique contemporain", PhD dissertation, Université de Provence, 2004.

[7] Bouzon, C. & Hirst, D.J. "Isochrony and prosodic structure in British English", in Proceedings of the Second International Conference on Speech Prosody:223-226, 2004.

[8] Breiman, L., Friedman, J., Olshen, R. and Stone, C. "Classification and regression trees", Monterey, CA, Wadsworth & Brooks, 1984.

[9] Brown, G. & Yule, G. "Discourse Analysis", Cambridge, Cambridge University Press, 1983.

[10] Crystal, T.H. and House, A.S. "Segmental duration in connected speech signals: Preliminary results", JASA, 72(3):705- 716, 1982.

[11] Fant, G., Kruckenberg, A. and Nord, L., "Durational correlates of stress in Swedish, French and English", Journal of Phonetics, 19:351-365, 1991.

[12] Grosz, B. & Hirschberg, J. "Some Intonational Characteristics of Discourse Structure", in Proceedings of the International Conference on Spoken Language Processing, 1:429-432, 1992.

[13] Jassem, W. "Intonation in Conversational English", Warsaw, Polish Academy of Science, 1952.

[14] Klatt, D.H. "Review of text-to-speech conversion for English", JASA, 82:737-793, 1987.

[15] Koopmans-van Beinum, F.J. & van Donzel, M.E. "Discourse Structure and its influenece on local speech rate", Proceedings of the Institute of Phonetic Sciences Amsterdam, 20, 1996.

[16] Ladd, D. "Intonational phonology", Cambridge University Press, 1996.

[17] Liberman M. & Pierrehumbert J. "Intonational Invariance under Changes in Pitch Range and Length", in Aronoff, M.. & Oehrl, R.T. [eds], Language Sound Structure:157-234, Cambridge, MA, MIT Press, 1984.

[18] Riley, M. "Tree-based modelling of segmental durations", in Bailly, G., Benoit, C. and Sawallis, T.R. [eds]: Talking Machines: Theories, Models and Designs, Elsevier, 1992.

[19] Smith C. " Topic transitions and durational prosody in reading aloud: production and modelling", Speech Communication, 42:247-270, 2004.

[20] Therneau, T.M. and Atkinson, B. rpart Package, computer software (R library), 2002-04-24 (R port by Ripley B., 2009-08-05).

[21] Trouvain, J. "Tempo Variation in Speech Production. Implications for Speech Synthesis", PhD dissertation, Saarbrücken, 2003.

[22] Wichmann, A. "Intonation in Text and Discourse: Beginnings, Middles and Ends", Pearson Education, London, Longman, 2000.