Is idiom identification possible from prosodic information? An experimental approach for the Greek language

Dimou Athanassia – Lida & Chalamandaris Aimilios

Université Paris 7
Denis Diderot, UFR Linguistique, Case 7003, 2 place Jussieu, 75005, Paris, France
Institute for Language and Speech Processing
Epidavrou & Artemidos 6, Maroussi, 15125, Athens, Greece.
ndimou@linguist.jussieu.fr, achalam@ilsp.gr

Abstract

The purpose of this paper is to investigate whether the identification of one's idiom is feasible through only prosodic information. As an experimental approach of this theory, we attempt to study whether native Greek adults can identify their mother idiom from synthesized stimuli which contain only prosodic - melodic and rhythmic - information. In the first section we present the main idea behind our work, in the second section we present the procedure we followed in order to complete this pilot study, while at the two final sections one can find the results and the conclusions of our experiments.

1. Introduction

Several theories and studies have been proposed about the possible identification of a language through only prosodic information [1], [5], [9]. Nevertheless, only some have attempted to prove that melody and prosodic information is indeed enough for identifying one's idiom [4], [11] from other idioms of the same language. In order to examine our hypothesis we have conducted a study, which includes a perceptive experiment; an identification task.

The utterances that were synthesized and served as stimuli in the experiment, came from recordings in two different regions in Greece, Athens, the capital city and Ayassos, a typical village of the island of Lesvos. In order to eliminate all lexical information, a Text-to-Speech engine, which has been developed at ILSP, was used for producing three different types of prosodically equivalent synthetic stimuli; the first two types of these stimuli contained only the phones [s] and [a], which were used to substitute all consonants and vowels of the original utterances respectively, with different prosodic contours; i.e. flat for the first type, intact in the second case. For the third type, different phones were used in order to replace the consonants of the naturally uttered sequences, while all vowels were replaced by the phone [a]. The use of these three different types of stimuli was made in order to verify our hypothesis, which was formed as follows: *"the rate of correct identification of a native listener for his mother idiom should be higher for the synthetic stimuli that provide him more prosodic information"*.

2. The Process

In order to avoid the recognition of specific voice patterns and voice quality, to which, all age and sociological differences of the speakers are reflected, we decided to use for all synthetic stimuli the voice of a professional speaker. The TtS engine we employed is tested and optimized for this professional speaker as the adaptation of the engine to the native speaker's voice would demand extra effort with potentially worse results. Even if this decision might have cost us in accuracy in matching the pitch contours of the original utterances with those of the synthetic stimuli, two additional reasons reinforced our decision:

- not everyone's voice is appropriate to be used for speech synthesis without producing distorted signal
- the phenomenon of allophones [13] was dominant in our case; this made it sometimes impossible to find 'clear' utterances of the phones needed, for the synthetic substitution of the phonemes of the original sequences.

Overall, the synthetic utterances sounded natural enough to make the listeners focus on the prosodic characteristics of the stimuli and not on the actual signals, a conclusion we reached to not only by the final results, but also from interviews with the subjects.

2.1. Segmentation procedure

The segmentation of the original recordings was carried out manually with the use of the open source program Praat [3]. A phonetician provided the transcription of the audio signals by hearing and performed their segmentation into individual phonemes. The transcription of the audio signals was carried out on the basis of the actual uttered phonemes and not on the grammatically correct Greek that should have been uttered. Hence, in cases where the speaker should normally pronounce a word of five phonemes, but he actually pronounced four of them, the transcription and the segmentation of that word was carried out only for the four materialized ones. Hence, the procedure of the manual segmentation helped us avoid possible errors that would have affected the final results if we had used an automatic speech recognition system; nevertheless extended testing of the procedure on larger corpora requires an automatic segmentation process, which still remains to be fine-tuned.

2.2. Pitch extraction

The algorithm that was used for the extraction of the pitch contour of every signal is the one suggested by Paul Boersma [2] as it is implemented, by him, in the Praat environment. The resultant contours were used as "transplants" for the synthesis of the experimental stimuli. The derived pitch contours were linearly interpolated at the silent parts of the audio signal, in order to be continuous and hence have meaning in the case of unvoiced consonants. For the case where all consonants were replaced by the phone [s] this has no significant meaning,

while for the case where the consonants were replaced by their respective group representative, as it is described in more detail in the following paragraph, was imperative in order to avoid pitch mismatches.

2.3. Synthetic stimuli creation

The production of the synthetic stimuli was performed with the help of the Text-to-Speech engine [7] that has been developed at ILSP; the engine is based on time-domain concatenative algorithms and makes use of pitch synchronous manipulation of pitch and phonemes durations The elemental units for the synthetic speech, i.e. the diphones with which we produced the synthetic speech, are derived from the original speech of a professional native Greek speaker, the voice of whom is used in the commercial ILSP Text-to-Speech system, "Ekfonitis+" [7]. In order to ensure that the synthetic stimuli would sound as natural as possible without much distortion, the target pitch contour was normalized to fit the pitch characteristics (mean value and bandwidth) of the professional speaker.

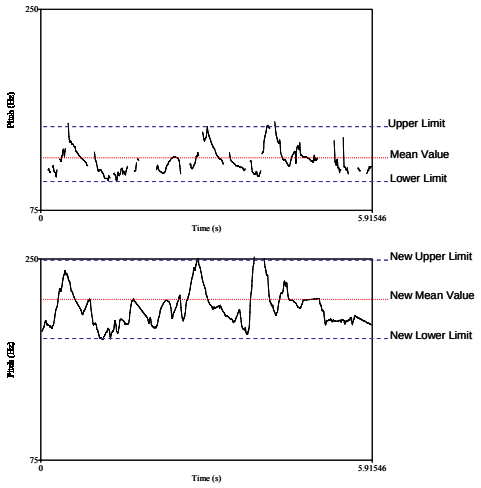

Figure 1: *(A) Original speech contour and normalized pitch contour extracted by synthetic stimulus. (B) The synthetic pitch contour extracted is normalized in order to meet the speaker's characteristics.*

After the normalization of the original pitch contour, a script was written, by making use of the ILSP TtS engine, which produced a synthetic speech signal, as close as possible to the original recordings, as far as the prosodic characteristics are considered.

The synthesis process gave birth to the three following stimuli sets:

1. A set of stimuli, which will be referred to as Sasasa flat group, where all vowels were substituted with the phone [a] and all consonants with the phone [s], all phones durations were kept intact as the original utterance, but with a flat pitch contour set at the mean value of the speaker's voice.
2. A set of stimuli, which will be referred to as Sasasa Intonated group, produced as above, but with the pitch contour normalized and transplanted as well into the synthetic stimuli.
3. A set of synthetic stimuli, which will be referred to as Saltnaja Intonated group, where all vowels were replaced by the phone [a] and all consonants were replaced by the phones [t] [d] [s] [z] [m] [l] [j] (voiceless plosives, voiced plosives, voiceless fricatives, voiced fricatives, nasals, liquids, and glides respectively). All phonemes durations and the pitch contour were transplanted as well into the synthetic stimuli.

These three sets were used in order to investigate our hypothesis of the escalation of the success rate in analogy to the amount of the prosodic information provided to the subject.

2.4. The original utterances

The purpose of this paper is to show that a native speaker can identify his mother idiom from its prosodic proprieties when this one is compared to another idiom of the same language; according to our hypothesis the native listener will identify more easily those utterances that contain most prosodic information – rhythmic, melodic and phonotactic.

The original sentences used for this experiment were all pronounced by Greek native speakers.

In order to maintain their traditions and their idiom, the habitants of the village of Ayassos set up each year several cultural events, such as carnivals and theatric plays which are written in their local idiom, the recordings of which they kindly provided us. After selecting several parts among these recordings, the selection of which was made principally upon the criterion of good audio quality, we translated them into the so called "standard Greek"; this procedure involved substitution of words used exclusively in the idiom of Ayassos [12], elimination of the allophonic variations proper in the idiom of Ayassos and rephrasing in the cases where this was necessary. The transcribed text was uttered by two professional actors, both originating from Athens, following the direction instructions that existed in the original theatrical manuscripts. In total, four different voices were used; one male and one female voice from each region.

After having established our corpus of recordings, and once the filters of synthetic speech were applied to all these utterances, a second selection was performed in order to limit the number of stimuli that would be used for the perceptive experiment. This selection was made upon the criterion of the stimuli length that should be no longer than 4 seconds, as well as good audio quality of the same stimulus in all three synthetic filters, for both idioms.

3. The experiment

After having produced the three sets of the synthetic stimuli, we carried out a listening experiment, aiming to study the validity or not of our hypothesis.

3.1. The Identification Task

In this experiment, the subjects listened to pairs of synthetic audio stimuli; each pair represented the same original utterance in both idioms. The airs of the synthetic stimuli were presented to the listeners in a shuffled order, different one for each of the three sets of synthetic stimuli. The listeners were asked to identify the Athenian ones. The total number of the stimuli was seventy-two, separated into three groups – according to the three different types of stimuli we described in paragraph 2.3. The listeners, who were all native Athenians, 30 women and 30 men, of the age between 20 and 45, after

having received the same instructions, performed the identification task in a specially designed interface in Praat program. Between two sequential stimuli a distinguishable beep sound was played in order to discriminate the two stimuli. The listening tests took place in a sound-proof room with the same headset for all subjects. The subjects could only hear once every pair of stimuli and there were no time restrictions for their answers, since they were also asked to define the degree of their answer certainty, on a three-level scale.

In order to allow the listeners' ear to adapt to the nature of the experiment and to the synthetic texture of the audio signals, 6 stimuli, two from each set, were used as introductory samples; the answers for these stimuli were not considered in the overall results. During the experiment and in order to measure the consistency of the listeners' answers, two pairs of stimuli in each synthetic filter were played twice throughout the test. By doing so we wanted to measure the consistency of the listeners' answers.

3.2. The statistic results

The statistic results [6], [10] of the perceptive experiment seem to validate our hypothesis. As shown in figure 2 below, the rate of correct identification increases from group Sasasa Flat to Saltanja Intonated.

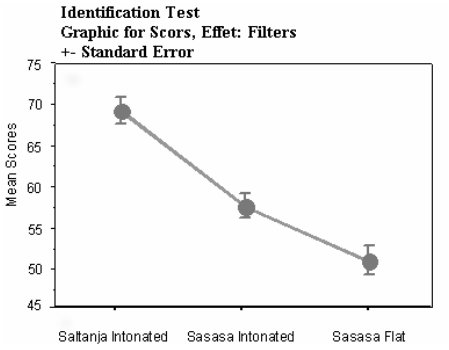

Figure 2: *The success rate in recognizing the Athenian accent versus the accent from Ayassos; correct answers in stimuli*

For the first group of synthetic stimuli the mean score obtained was 51%. The listeners, as expected, had difficulties in identifying the correct answers and actually answered randomly. This was due to the fact that almost no information, prosodic or other, was provided to them, as the synthetic stimuli did not contain anything but the syllabic duration. For the second group of the synthetic stimuli the prosodic information provided was richer, since apart from the syllabic durations – rhythmic information –the melodic information was also provided, i.e. the pitch contour. This additional information resulted to a higher mean percentage of correct identification of 58%. According to the variation analysis this result is significantly different from the results of the first group, namely the probability of this being attributed to accidental effects is less than 5%. Finally, in the third group of synthetic stimuli, the one named Saltanja Intonated, the mean score for the correct identification was of 69%. In this set of stimuli the listeners were given not only the rhythmic and the melodic information of the two previous conditions, but also a certain amount of phonotactical information; this was achieved through the use of a synthesis filter that was richer in phonemes, one different consonant for every phonemic category. We believe that this additional phonotactical information helped them identify the differences in the metric organization between the two idioms; potential invalid syllabic sequences, that did not match any retrievable to the native listener prosodic pattern, were subliminally attributed to the unknown to them idiom.

Table 2 below actually validates our hypothesis as it illustrates that the differences between the rates of correct identification obtained in these three sets of samples are significantly different from one to the other.

Identification Test
Test PLSD de Fisher for Scores, Effect: Filters

	Mean Diff.	Crit. Diff.	p value	
Saltanja Int., Sasasa Int.	11,528	4,490	<,0001	S
Saltanja Int., Sasasa Plat	18,108	4,490	<,0001	S
Sasasa Int., Sasasa Plat	6,580	4,490	,0043	S

Table 2: Test a posteriori PLSD de Fisher; correct identification rate in the three synthetic filters used in the identification task.

Figure 3 below illustrates in a 1 – 3 scale the degree of certainty of the listeners for each of the three synthetic sets. In this graph, 1 represents uncertainty while 3 represents absolute certainty; both "correct" and "all answers" are plotted. Regarding the level of certainty of all responses, the listeners' mean score was 1,895 out 3 for the first filter (Sasasa Flat), 1,932 out of 3 for the second one (Sasasa Intonated) and 1,977 out 3 for the third one (Saltanja Intonated).

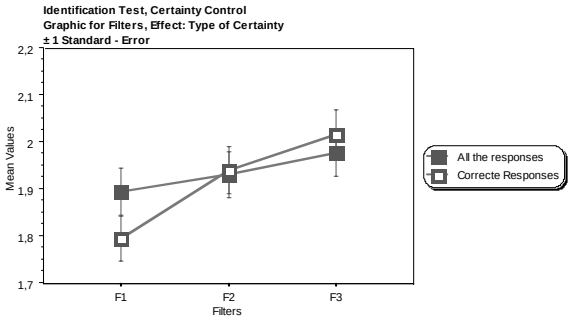

Figure 3: *The rate of the certainty degree of the Athenian listeners when recognizing their mother accent versus the accent from Ayassos.*

According to table 3 below, the only difference that is not due to chance is the one between the first filter and the third one. Globally, the impression we obtain from these results is that the listeners were in the whereabouts of "certain enough" during the whole test.

Nevertheless, this impression is somehow updated when we take a look at the results of the variation analysis effectuated for the answers where the listeners actually identified the right stimuli, in figure 3. For the first filter the degree of certainty of the listeners that correctly identified the Athenian stimuli is 1.796. For the second one, this rate increases up to 1.941 and for the third one it reaches 2.018. The variation analysis, the results of which are resumed in the table 4 below, depicted that the difference between all scores for the three filters is significant and hence not attributed to chance. These results

can be qualified as previously, i.e. as verification to our hypothesis. The increase of prosodic information provided from filter 1 to filter 3 made the listeners feel more certain about their answers, which in addition were more correct.

Identification Test
Test PLSD de Fisher for Filters, Effect: Type of Certainty
(All the Responses)

	Mean Diff.	Crit. Diff.	p value	
Saltanja Int., Sasasa Int.	-,038	,063	,2406	
Saltanja Int., Sasasa Plat	-,083	,063	,0109	S
Sasasa Int., Sasasa Plat	-,045	,063	,1615	

Identification Test
Test PLSD de Fisher for Filters, Effect: Type of Certainty
(Correct Responses)

	Mean Diff.	Crit. Diff.	p value	
Saltanja Int., Sasasa Int.	-,145	,074	,0002	S
Saltanja Int., Sasasa Plat	-,222	,074	<,0001	S
Sasasa Int., Sasasa Plat	-,077	,074	,0404	S

Table 3 & Table 4: Test a posteriori PLSD de Fisher; the rate of the certainty degree for each answered stimuli in the three synthetic filters used in the identification task; totality of the given responses versus the correct responses.

4. The conclusions

In overall, we could say that the results of the identification test validate our hypothesis; a native speaker can identify his mother idiom when compared to another idiom of the same language having in disposal only prosodic and not contextual information.

Moreover, in this study we achieved to evaluate an experimental protocol based upon synthetic speech; our statistical hypothesis was validated as well. The listeners tend to identify more easily and more correctly their mother-idiom when they are provided with more prosodic information. Through this affirmation and according to the obtained results and the respective analysis effectuated, we can assume that only rhythmic information, which for the case of this study corresponds to syllabic durations, is not ample enough to provide a solid idea of the language spoken and hence not sufficient for any native speaker to identify it what so ever.

Nonetheless, this rhythmic information in combination with the melodic one, as shown in the results for the second synthesis filter, create a fertile environment for correct identification (Mean score of correct identification =58%); the score obtained for this filter, shows that the listeners actually captured something in the prosodic contour that helped them make their decision, with a certain hesitation though, as the score is still low enough and certainly lower than the one obtained for filter 3.

In addition to the rhythmic and the melodic information of the two precedent filters, phonotactic rules of the every idiom were incorporated in the third filter stimuli. Even though this operation did not cover the whole phonemic system of the native's idiom, was distinctive enough for them to be able to retrieve familiar prosodic patterns. In other words the phonotactic information in the third filter, reflected some aspects of the metric organization of each idiom and hence made apparent for the listeners of the Athenian idiom which sequences were not part or were different enough from their mother-idiom patterns. This procedure resulted to a 69% of correct identification and a 2,018 out of 3 degree of certainty for the listeners who actually answered correctly.

In general, we believe that the nature of the task was not easy; a higher degree of abstraction was necessary during the entire test, which even though lasted a little less than 20 minutes, demanded extra mental effort and concentration. A positive answer in an identification task demands a solid decision on behalf of the listener who is called to identify his mother idiom among sequences of synthetic speech even though his decision was made, sometimes, upon a negative identification.

Future work of this research includes, among others, the phonetic analysis of these results and the investigation for potential correlations between specific phonotactic patterns and the identification rate.

5. Acknowledgements

For the completion of the experiment we feel the need to thank everybody who volunteered to participate in this research as well as M. Dommergues J-Y, professor in Paris 8 University, for his valuable help in the statistic analysis of the results and for his precious guidance during the construction of the entire perceptive experiment.

6. References

[1] Benali, I, 2004, Le rôle de la prosodie dans l'identification de deux parlé algériens: l'algérois et l'oranais, *Actes de MIDL*. Paris,127-132.

[2] Boersma, P, 1993, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings of the Institute of Phonetic Sciences 17*. University of Amsterdam, 97-110..

[3] Boersma, P, 2005, PRAAT: Doing Phonetics by computer (Version 4.3.14), http://www.praat.org, 26/05/2005

[4] Centre de la langue Grecque, 2000, Dialecte et dialectologie du grec modern, in Christidis A.-F.(ed.), *La langue grecque et ses dialectes*, Athènes, Direction des relations internationales.

[5] Dimou, A – L; Chalamandaris, A, 2006. Dialect Identification through prosodic information: an experimental approach, In *Proceedings of Speech Prosody Conference 2006,* Dresden, Germany.

[6] Dommergues, J.-Y, 2006, *Initiation à la statistique appliquée aux sciences du langage et de la parole.* Paris: Reprographie Paris 8.

[7] Ekfonitis+, Institute for Language and Speech Processing TtS http://www.ilsp.gr/ekfonitis

[8] Fodor, J,D, 2002. Psycholinguistics Cannot Escape Prosody, In *Proceedings of Speech Prosody Conference 2002*, Aix-En-Province, France.

[9] Ramus, F; Mehler, J, 1999. Language identification with suprasegmental cues; A study based on speech resynthesis, *Acoustical Society of America*, Vol. 105.

[10] SAS, I, 1992 – 1998, StatView pour Windows. NC.

[11] Κοντοσόπουλος, Γ-Ν, 2001. *Διάλεκτοι και ιδιώματα της Νέας Ελληνικής*. Εκ. Γρηγόρη

[12] Παπάνης, Δ; Παπάνης, Γ, 2004. Λεξικό του Αγιασώτικου Ιδιωματικού Λόγου, Μυτιλήνη.

[13] Πετρούνιας, Ε, 1984. *Νεοελληνική Γραμματική και Συγκριτική (Αντιπαραθετική) Ανάλυση, Μέρος Α': Θεωρία.* Θεσσαλονίκη :University Studio Press.