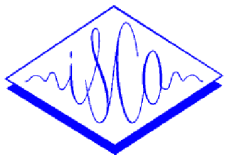

Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

Dialect identification through prosodic information: an experimental approach

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 native Greek adults can identify their mother tongue from synthesized stimuli which contain only prosodic - melodic and rhythmic - information. More specifically we are trying to investigate whether Greek native speakers are able to discriminate their mother dialect form another one, also from Greece, only from prosodic information. In the first section we present the main ideas underlie 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

There have been several theories and studies about the possible identification of a dialect through only prosodic information. Nevertheless, only some have attempted to prove that melody and prosodic information is indeed enough for identifying one's dialect [7,9,10,11,12].
In order to examine our hypothesis we have conducted a pilot study, which includes two perceptive experiments; an identification task and a discrimination one. The utterances that were synthesized and served as stimuli in both experiments, came from recordings in two different regions in Greece, Athens, the capital city and Agiasso, a typical village in the island of Lesvos. In order to eliminate all lexical information, a Text-to-Speech engine, which has been developed at ILSP (Institute for Language and Speech Processing, Greece), was used for producing a prosodically equivalent synthetic stimulus that contained only the phonemes /m/ and /a/, which replaced respectively all consonants and vowels of the original utterances.. In this paper we are presenting the preliminary results of an extended research, which aims to investigate the aforementioned hypothesis. The obtained results from the perception tests are surprisingly positive; nevertheless in order to be able to support the language identification hypothesis with solid evidence, the hypothesis needs to be tested on a larger corpus and with a larger number of subjects, while a prosodic analysis of both utterances and synthetic stimuli could provide a reasonable interpretation of the results.

2. The Process

2.1. Segmentation procedure

The segmentation of the original recordings was carried out manually with the use of the open source program Praat [5]. A phonetician provided the transcription of the audio signals and performed their segmentation into individual phonemes. The transcription of the audio signals was carried out on the basis of the actual uttered speech 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, skipping for example the third one, the transcription and the segmentation of that word was carried out only for the four pronounced ones. For the reason mentioned above, the procedure of the manual segmentation helped us avoid possible errors that might affect the final results. However, an extended testing on larger corpora requires an automatic segmentation process, which unfortunately 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 [6] as it is implemented by him in the Praat environment. The selected algorithm performs quite well with speech signals and it also incorporates mechanisms for voicing detection. 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, which in the synthetic stimuli are transformed into the phone /m/.

2.3. Synthetic stimuli creation

The creation of the synthetic stimuli was performed with the help of the Text-to-Speech engine [7] that has been developed in ILSP, and which 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+". In order to ensure that the synthetic stimuli will sound as natural as possible without much distortion, the target pitch

10.21437/SpeechProsody.2006-12

contour was normalized to fit the pitch characteristics (mean value and bandwidth) of the professional speaker.
After the normalization of the pitch contour, a script was written which by making use of the ILSP TtS engine, it produced a synthetic speech signal, as close as possible to the original recordings, as far as the prosodic characteristics are considered, i.e. the pitch and the phonemes durations, replacing at the same time all vowels with the phoneme /a/ and all consonants with the phoneme /m/. As this is only a pilot study not all prosodic parameters were taken into account for the production of the synthetic signal; the amplitude modulations of the signal was not considered during the production of the synthetic signals, however the importance of it, is something that needs to be further investigated in future research.

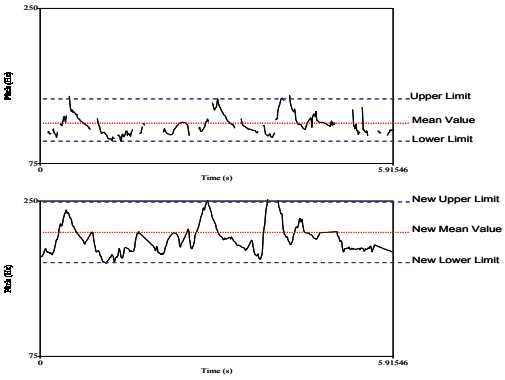

Figure 1: (A) Original pitch contour and normalized pitch contour extracted by synthetic stimulus. (B)

2.4. The original utterances

The purpose of this paper is to show that a native speaker can identify his mother tongue from its prosodic proprieties even when this one is compared to another dialect-idiom of the same language. The original sentences used for this experiment were all pronounced by Greek native speakers, three Athenian speakers and three speakers originating from Agiasso. For each region, two masculine and one feminine voice were used. For the recording procedure an MD portable device was used with a multi direction microphone. The speakers from Agiasso are aged between 40-60 years old, they live permanently in Agiasso and in most cases they have an elementary educational level (almost all of Agiasso natives are farmers).Unlike the speakers from Agiasso, those from Athens are aged from 25 to 45 years old. They all have a higher education background and they are originated from Athens in which they have been living in Athens since their birth. The specific utterances used were chosen from a corpus of recordings especially compiled for this research, which includes the recordings of 10 different people from each region. Each speaker was interviewed for about an hour and the interview can be separated into two parts; a free discussion over several issues relevant to the speaker's habits and intersts and a text reading part. As far as the first part is concerned, most of the times the subject of the conversation evolved around the speaker's profession. Some problems were encountered during the text reading part of the interview; the speakers were asked to read an article selected from a recent newspaper as well as some isolated phrases selected by the interviewer. Unfortunately not all speakers, and mainly the speakers from Agiasso, were able to read either due to a vision problem or because they were illiterate.
From our recordings we retrieved utterances of similar length, about 8 – 10 seconds each, including the pauses. We tried to extract these utterances from parts of the recordings were the speech is continuous and affirmative. Likewise we have selected 3 utterances from each one of the three speakers for each dialect. The three utterances of two speakers, one from each dialect, served as introductory cases in order to help the listeners get acquainted with the task as well as with the sounding of the synthetic stimuli, at the beginning of both tests.

3. The experiments

In order to avoid the recognition of specific voice patterns and voice quality, to which are in a way reflected all age and sociological differences of the speakers, instead of the prosodic proprieties of each utterance we decided to use for all synthetic stimuli the voice of another speaker. The used TtS engine is tested and optimized for the specific speaker and the adaptation of the engine to another speaker's voice would demand extra effort. 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: a)not everyone's voice is appropriate to be used for speech synthesis without producing distorted signal and b) the phenomenon of allophones [8,9] was quite dominant in our case where sometimes it was impossible to find 'clear' utterances of both phonemes /m/ and /a/ in one's speech. The synthetic utterances were natural enough to make the listeners focus on the prosodic characteristics of the stimuli and not on the actual signals. Nevertheless, in order to provide the listener with the necessary time-frame for getting used with the sounding of the stimuli, some preparatory stimuli at the beginning of each test were provided to the listeners mainly for this reason. Their grading was not considered in the overall results. As already mentioned, the experiment was consisted of two different acoustic assessments of the audio stimuli; an identification task and a recognition task.

3.1. The first experiment

In the first experiment, the subjects listened to the audio stimuli which were shuffled, and they were asked to identify the Athenian stimuli. The total number of the stimuli was sixteen. The listeners, who were all native Athenians, 8 women and 8 men, of the age between 28 and 45, after having received the same instructions, they listened through a headset in a noise-proof room the stimuli, one after another, with a single beep noise between two sequential stimuli. After each stimulus was played, the listeners were given 3 seconds to decide and write down on the questionnaire their answer. In order to provide the listeners' ear the time to adapt to the nature of the experiment and to the synthetic texture of the audio signals, at the beginning of the test we had inserted 6 stimuli, three from every accent, that were not taken into account for the final results, and they were used as introductory cases. During the experiment and in order to measure the consistency of the listeners' answers, two stimuli were played twice throughout the test. By doing so we

attempted to investigate the degree of difficulty that the listeners were having in completing the test, as well as the concordance of the results.

3.2. The results for the first experiment

We found out that it was rather hard to complete the test, as also most of the listeners said after the test. Only 28% of the listeners were consistent in their answers, while 57% answered differently in the same stimulus and 15% where inconsistent in both test stimuli. This could suggest that the answers could have been put in chance; nevertheless this is not a valid conclusion as it was shown by a t-test analysis of the results.

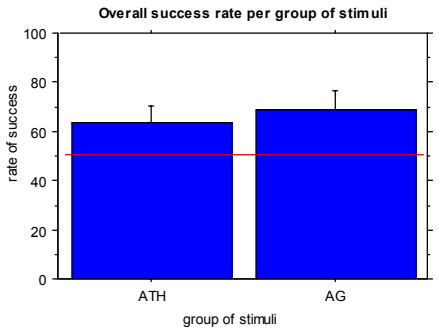

Figure 2: The success rate in recognizing the Athenian accent versus the accent from Agiasso; correct answers in stimuli

The listeners recognized 63.4% of the Athenian accent as well as 69% of the non-Athenian accent (accent from Agiasso). Both results are satisfactory as they give a recognition rate higher of 50%, which means that the listeners' answers were up to a certain degree conscious.

T_test (Indipendent group) for Correct Identification of the type of Stimulus (ATH vs AG)

	Ecart moyen	DDL	t	p
ATH, AG	-5,250	14	-,506	,6209

Table 2: Statistic results from the 1st experiment

However, a t-test for independent groups for correct identification of the type of stimulus to be identified shows that the difference between the two recognition rates (for Athenian accent – 63.4% vs. accent from Agiasso – 69%) is not important In order to be able to give any interpretation of this rather unsatisfactory result, and due to the small number of participants (N<20) we have effectuated a Chi-2 test; with this test we wanted to see whether the distributions of the responses of the listeners for the two types of stimuli are different from a theoretical distribution (50) and therefore interpreted as a choice made by chance. The results of the Chi-2 test are rather encouraging; the distributions for the two types of stimuli are significantly different from the theoretical one ($p < 0.0001$ & $Chi2 = 52.191$). One possible interpretation of these results could be that as far as the recognition rate of the two accents is concerned, the scores of 63% and 69% are significantly higher than the mean random value of 50% at a 0.05 level; the trend in both tests is to show that the choice of the listeners was not made by chance. Additionally, the rather small number of items taken into account for the t-tests (N=8) is probably the reason why the correct identification rate for the Athenian stimuli of 63% is inferior of the 69% of correct identification rate for the stimuli from Agiasso. We suppose that with a larger number of items the recognition rate will be superior.

3.3. The second experiment

The second stage of the experiment was consisted of the comparison of two different stimuli, one with Athenian accent and one with the accent of the Agiasso dialect, and the listeners were asked to decide which stimulus had Athenian accent and which had not. This test was aiming to bypass the potential difficulty of the first test, as we believed that it would be easier to compare two sequential stimuli and decide which one sounds closer to the Athenian accent. For this second experiment we used the same test-set used earlier in the first experiment. The test consisted of the same 20 audio stimuli which this time they were grouped in pairs containing one sample from both accents each. By doing so we ended up with 9 pairs of audio stimuli that were shuffled, while, as also done in the first experiment, one pair was repeated once in the test-set, in order to investigate the concordance of the answers. The listeners had 4 seconds after every audio sample to decide and write down on a questionnaire which of the two stimuli they believed resembled more to the Athenian accent. In order to provide the listeners' the time to adapt to the nature of the second task at the beginning of the second test we inserted the same 6 stimuli as in the first experiment, in pairs of two, which were also used as introductory cases. The listeners' responses for this part of the test were not taken into account for the final results.

3.4. The results of the second experiment

The overall success rate in discriminating correctly the Athenian accent from the one from Agiasso was 71%. A t-test analysis of the data leads us to the conclusion that the mean value is statistically significant. In this case again then 95% confidence interval lies entirely above the 0.0.

	N	Mean	Std. Dev.	Std. Error Mean	
Test2	16	.71	.131	.031	
	Test Value = 0.5				
	t	df	Sig. (2-tailed)	Mean Diff	95% Conf. Int. of the Diff.
					Lower / Upper
Test2	6.8	15	.000	.21	.14 / .27

Table 3: Statistic results for the 2nd experiment

The mean rate of identifying the Athenian dialect, which was the main object of research in both tests, rise up to 71% in the second test versus 63% when the listeners were asked to identify them in isolated environment.

4. The conclusions

In overall, we could say that the results of both tests are promising and encouraging, as far as the task of language identification from suprasegmental cues is concerned. The

Athenian listeners identified correctly in both cases the stimuli of the Athenian accent. In the first experiment the listeners attained the score of 63% for correctly identifying the Athenian accent; we should note here that the task in this first experiment was more difficult as the listeners were asked in a way to recognize the identity of each stimulus. A t-test on the answers of the Athenian listeners for correct identification of the Athenian accent and the accent from Agiasso, gives us a clearer image of the identification process. The stimuli that were more difficult to identify were those of the Athenian accent as they were recognized up to 63% correctly, while the stimuli from Agiasso, recognized as non Athenian, were identified as such at 69%.

Even though the difference between the two recognition rates of each group is not significant, what is more important to say about this difference is that in the case of the Athenian accent, the recognition rate is not superior of 50% ($p < 0.0001$). We believe that this result is due to the nature of the task of identification; a positive answer in an identification task demands a solid decision on behalf of the listener who is called to identify his mother tongue among sequences of synthetic speech. This is a more difficult decision to be made than the negative identification; in this second case, anything that does not seem familiar can be more easily characterized as non-Athenian. In a hesitation moment a listener is prone to a wrong identification of the stimuli because one has not yet shaped an acoustic profile of how his mother tongue could sound in synthetic environment. But in order to be able to support such an interpretation, a prosodic analysis of both utterances and synthetic stimuli is required. Such analysis is planned at a second stage of this research so as to see if the identification positive or negative is based upon the recognition of specific prosodic proprieties.

In the second experiment admittedly the task was easier for two reasons. Firstly, because the listeners were asked to choose between two stimuli in order to identify the Athenian accent. They had therefore a variety of prosodic information, so they could match their answer to the stimulus that resembled more to what they innately perceive as their mother language and reject the stimulus that less fitted to the same criteria. The higher rate of success in the second experiment can be also attributed to the nature of the experimental procedure; the discrimination task came second after the identification one, during which the listeners had actually the time to get acquainted to the sounds of the synthetic stimuli and perform the necessary abstraction in order to assimilate these sounds to the sounds of human language. This second argument however, does not reduce the importance of the success rate in the second experiment, which is elevated at 71% vs. 63% of the first experiment and which as proved by the t-tests is far from being chosen by chance ($p<0.05$).

5. Future Work

As mentioned before, this paper describes the results of a pilot study, which aims to investigate the possibility of identifying a dialect from only prosodic information. The preliminary results showed that up to a certain point, this theory stands on solid evidences; nevertheless the scale of the experiment does not allow us to reach to general conclusions. Hence our immediate future work involves a generalization of the results through a larger scale experiment.

6. Acknowledgements

Part of this research was carried out at the premises of the Institute for Language and Speech Processing in Greece. We would like to thank Dr. Athanassios Protopapas for his valuable help in the statistical analysis of the acquired data and Ms. Jean-Yves Dommergues, professor at the Paris 7 University – Denis Diderot, for his valuable advice on statistical issues and the methodological approach that we have used, S. Raptis. P Tsiakoulis and S. Karabetsos, who are the core of ILSP speech synthesis team and we want also to thank all the speakers, from Athens and from Agiasso, who accepted to be recorded for the use of this study as well as the staff of ILSP who gladly participated in the experiment as listeners.

7. 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 : 127-132, Paris.

[2] BOERSMA P. (2005), PRAAT: DOING PHONETICS BY COMPUTER (VERSION 4.3.14), http://www.praat.org, 26/05/2005

[3] 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:, 97-110. University of Amsterdam.

[4] BOTINIS A., BANNERT R., FOURAKIS M., PAGONI – TETLOW S. (2002), "Cross-linguistic segmental durations and prosodic typology", *Fonetik,* Vol.44.

[5] 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.

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

[7] FODOR J.D. (2002), "Psycholinguistics cannot escape prosody", In *Proceedings of the Speech Prosody 2002 Conference, Aix-en-Provence, France.*

[8] KONTOSOPOULOS N.G. (2001), « Διάλεκτοι και ιδιώματα της Νέας Ελληνικής », Αθήνα, Εκδόσεις Γρηγόρης, 84-108.

[9] OTAKE T. & CUTLER A. (1999), "Perception of suprasegmental structure in a non- native dialect", *Journal of Phonetics* 27, 229-253.

[10] RAMUS F. (1997), « Le rôle du rythme pour le discrimination des langues », Actes des JIOSC 97 : 225-229, Orsay.

[11] RAMUS F. (1999), «La discrimination des langues par la prosodie : Modélisation linguistique et études comportementales», in Pellegrino F.(ed.), De la caractérisation à l'identification des langues, Actes de la 1ère journée d'étude sur l'identification automatique des langues, Lyon, Editions de l'Institut des Sciences de l'Homme: 186-201.

[12] RAMUS F., MEHLER J., (1999), "Language identification with supragegmental cues: a study based on speech resynthesis", Journal of Acoustic Society of America, vol.105, No.1: 512-521.