Language identification from suprasegmental cues: Speech synthesis of Greek utterances from different dialectal variations.

Dimou Athanassia – Lida & Chalamandaris Aimilios

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

Abstract

In this paper we present the continuation of our research on the ability of native Greek adults to identify their mother tongue from synthesized stimuli which contain only prosodic - melodic and rhythmic - information. In the first section we present the ideas that underlie our theory, together with a brief review of our preliminary results. In the second section the detailed description of our experimental approach is given, as well as the results and their statistical analysis. In the final two sections we provide the conclusions derived from our experiments and the future work we are planning to carry out.

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 two distinct phonemes for replacing every original consonant with the phone /m/ and every original vowel with /a/. The obtained results as far as speech resynthesis is concerned were rather promising, which is the reason why apart from the pilot study we have conducted another perceptive experiment using similar kind of synthetic stimuli.

More specifically, the statistic results of the pilot study showed that Greek adult natives originated from Athens can actually identify their mother tongue, dialect – idiom, when compared to another one of the same language, from prosodic cues at a rate of 63%. In the present paper we describe the results of experiments that have followed the aforementioned researches, which are also very positive and in accordance to our hypothesis. Although a validation of the results onto a larger scale and more elaborate experiment where actually all original phones will be replaced in speech synthesis by distinct phonemic categories is required, the results allow us to be positive in our hypothesis.

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 and in any case, automatic segmentation of idiomatic speech is very hard to achieve.

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 at ILSP, and which is based on time-domain concatenative algorithms. It 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 (voice-talent), 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 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/.

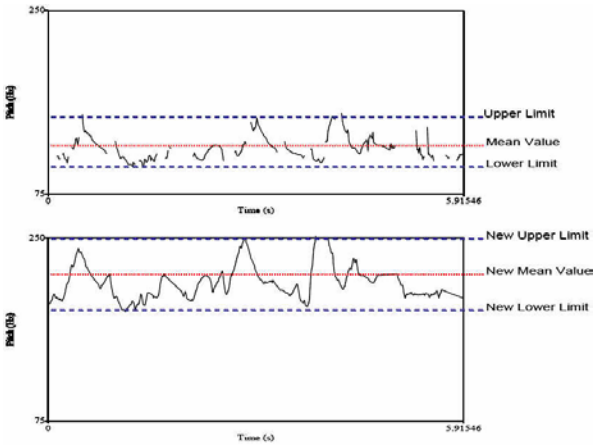

Figure 1: (A) Original pitch contour and normalized pitch contour extracted by synthetic stimulus. (B)

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 all experiments were all pronounced by Greek native speakers, ten Athenian speakers and ten speakers originating from Agiasso. 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.

For the pilot study a smaller number of speakers were used, three from each region; hence a more restricted number of stimuli was produced.

For the second perceptive experiment the utterances were extracted from the recordings of four speakers from each region, two of which were masculine and two feminine. The specific utterances used for all experiments were chosen from this recoded spoken corpus, especially compiled for this research, which includes two parts; a free discussion over several issues relevant to the speaker's habits and interests 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. The total duration of each interview lasted about an hour. From our recordings we retrieved 2 kinds of utterances: long utterances about 8 – 10 seconds each, including the pauses and short utterances about 3-4 seconds each. We tried to extract these utterances from parts of the recordings were the speech is continuous and affirmative. Likewise we have selected 3 utterances of each kind from each one of the four speakers for each dialect.

3. The experiments

We decided to use for all synthetic stimuli the voice of another speaker for the synthetic stimuli. That was mainly 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. The employed TtS engine is tested and optimized for the specific speaker and the adaptation of the engine to another speaker's voice would demand additional 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 all perceptive experiments were provided to the listeners mainly for this reason. Their grading was not considered in the overall results.

3.1. The pilot study

During the pilot study, we have conducted two small experiments, an identification task and a discrimination one, with a limited number of stimuli as well as listeners. At this stage of our research we used only long utterances of 8-10 seconds long that were transformed into non-sense stimuli. A total of 16 stimuli, coming from the dialect of Athens and from the dialect of Agiasso were shuffled in two different ways: as units for the identification task and in pairs for the discrimination one.

Our main objective at the time being was to investigate in which of the two experimental situations the listeners responded to with better results, given the fact that they heard synthetic stimuli, and the underlying reason. In both cases, the subjects, who were all native Athenians, 8 women and 8 men, of the age between 28 and 45 and were asked to identify the Athenian stimuli. The listeners,

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 for the identification task, and in pairs one after another with a double beep noise between two sequential pairs of stimuli. During each task, after each stimulus or pair of stimuli was played, the listeners were given 3 and 4 seconds respectively 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 each test they listened to six stimuli, three from each region, the score of which was not taken into account.

3.2. The results for the pilot study

According to most of the listeners it was rather hard for them to complete both tests. Between the two tasks, the one that was even more difficult to answer was the first one, the identification task. The statistic results confirmed listeners' opinion, as for the identification task the success rate for recognizing the Athenian dialect, came up to 63% while in the discrimination task it reached 71%. 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. In order to be able to attribute any interpretation to these results, and due to the small number of participants (N<20) we have effectuated a Chi-2 test; that is in order to investigate whether the distributions of the correct responses for the Athenian stimuli of the listeners for the two types of perceptive experiences 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 both tasks 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 Athenian accent is concerned, the scores of 63% and 71% 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.

However, and even though the global statistic results for both tasks are rather encouraging the small number of items taken into account for the t-tests (N=8) is probably the reason why in the identification task the correct identification rate for the Athenian stimuli is up to 63% but is inferior of the respective 69% for the success rate for the stimuli from Agiasso. The effectuated t-test showed that the difference between the two success rates is not significant. We suppose that with a larger number of items the recognition rate will be superior.

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 1: Statistic results from the 1st experiment

In the bottom line, the Athenian listeners identified correctly in both cases the stimuli of the Athenian accent. We should note that the first task, for which the listeners attained the score of 63% for correctly identifying the Athenian accent, had an additional difficulty as the listeners were asked to make a solid decision and identify their mother tongue among sequences of synthetic speech. Whilst in the discrimination task, in a sort of negative identification, anything that does not seem familiar can be more easily characterized as non-Athenian.

3.3. The second experiment

By maintaining the exact experimental framework, we performed another more elaborate identification task, and a rhythmic analysis of the utterances used, in order to evaluate the efficiency of the method and the stimuli used.

The main differences with the identification task of the pilot study are the following:

1) The utterances for this experience were of two kinds:
    a. long utterances 8-10 seconds long
    b. short utterances 3-4 seconds long
2) In total forty sequences, twenty from each dialect were used, instead of 16 in the pilot study.
3) Among these forty sequences two long ones and two short ones utterances, two from each dialect, were repeated in order to measure the consistency in the listeners' answers.

3.4. The results of the second experiment

The obtained results from this experiment were ambiguous. The Athenian listeners (N=20) regognised correctly the Athenian accent at only 50%, and the accent of Agiasso at 57%. The difference between these two mean values is not significant as the t-test for independent group depicted.

The most interesting finding though of this experiment is the high rate in the consistency scores: the short Athenian stimulus was correctly identified at 74% and 68% both times that appeared in the test while the long stimulus from Agiasso was correctly identified as non-Athenian at 68% and 63%.

3.5. Is it a question of fast speech rate?

In order to proceed with our analysis, we measured the words pronounced per minute in all utterances used in our experimental protocol.

The overall statistic results showed that there is a significant difference (p=0.0023) between the average number of words per minute pronounced by the two groups of speakers; (197 words/min for the Athenian speakers versus. 155 words per minute for the speakers from Agiasso).

The word rate pronounced of each of the four types of utterances used (short Athenian, long Athenian, short Agiassian, long Agiassian) is illustrated at the following table. The effectuated t-tests showed that Athenian speakers tend to pronounce more words per minute in short utterances than in long ones as the difference between the words per minute in short and in long Athenian utterances is significant (p=0.008). On the contrary, the speakers from Agiasso tend to use a similar word rates in short as well as in long utterances, since the difference words per minute in the two types of utterances is not significant.

	ATH_S	ATH_L	AG_S	AG_L
% Correct Identification	60	39	34	84
words/min	225	168,8	169	141

Table 2: Correct identification scores and words per minute rate for all four stimuli categories of the identification task.

As shown in table 2, the stimuli that failed to be identified had similar word rate even though they belonged to different dialects. We effectuated a t-test for the number of words per minute pronounced in these two stimuli categories, placing a theoretical value at 168 which is the mean value for words per minute for both categories.We found that the distribution for the long Athenian stimuli is entirely overlapped by the distribution of the short stimuli from Agiasso.

4. The conclusions

In general, and for the given pair of Greek dialects, we could assume that language identification can indeed be achieved through only prosodic information.

As far as the pilot study is concerned, the listeners performed better in the discrimination task than in the identification one. A possible interpretation is that in a hesitation moment a listener is prone to a wrong identification of the stimulus given one has not yet shaped an acoustic profile of how his or her mother tongue could sound in a synthetic environment. On the contrary, in a discrimination procedure one is given the possibility to choose the stimulus that best fits in this profile.

The results of the second experience clearly showed that the experimental protocol was in a way biased by the rhythmic parameter of speech elocution. The listeners could not identify correctly the two dialects when their word rates were similar but could identify them when the respective rhythmic patterns were not overlapped.

Further work is needed and is orientated in the two following directions: a). in order to avoid rhythmic interference, the phonological structure of both dialects needs to be examined and applied to the speech synthesis protocol, by using six distinct phonemes for replacing their respective phoneme types (all voiceless plosives with /t/, all voiced plosives with /b/, nasals with /m/, liquids with /l/, fricatives with /s/ and all vowels with /a/). We believe that through such an operation the synthetic stimuli will sound more familiar, even though they will remain non-sense, as micro prosodic information will be kept. By doing so it is possible that the listeners will need to make a small effort for the necessary abstraction in order to assimilate these sounds to the sounds of human language. b) A larger scale experiment is needed in terms of items and participants, as well as in terms of other types of speech such as theater plays for which an existing corpus of recordings of the same two dialects is being already analysed.

5. 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 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. We also need to thank Dr. S. Raptis. P Tsiakoulis and S. Karabetsos, who also constitute the core of ILSP speech synthesis team and we want also to thank all the speakers, from Athens and from Agiasso, who gladly 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.

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 : 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.