From Perceptual Designs to Linguistic Typology and Automatic Language Identification : Overview and Perspectives

Melissa BARKAT & Ioana VASILESCU

Laboratoire Dynamique du Langage ISH
14, avenue Berthelot 69363 Lyon cedex 07

Melissa.Barkat@ish-lyon.cnrs.fr / Ioana.Vasilescu@ish-lyon.cnrs.fr

Abstract

This paper deals with the overview of the methods in perceptual language identification and the suggestion of a new approach based on a two-step methodology integrating to perception "genetic" considerations and resulting into the modeling of perceptually identified discriminative cues. The first study reported here concerns experimental designs for perceptual and automatic identification of the dialectal level of languages that are less represented in the literature although it is spoken on very large geographical area (i.e. Arabic dialectal continuum). The same experimental design is implemented to determine a set of linguistic criteria for the automatic identification of 5 Romance languages (i.e. French, Italian, Spanish, Portuguese and Romanian).

1. Introduction

Since automatic approaches did not really lead to major improvements yet, perceptual experiments in language identification with human subjects provide a useful comparison tool for the evaluation of automatic systems' performances. Indeed, several studies have shown the superiority of the human perceptual performances in language identification task as for (i) the length of the training session [1], (ii) the number of languages to be discriminated [2], (iii) the quality of the speech signal [3] [4] and (iv) the accuracy of the discriminative criteria used for the identification [5] [6] [7]. On the methodological point of view two major trends were developed in time. The study of the human perceptual system started with experimental designs based on natural-speech [8]. These studies, performed on different corpora, were very limited in their scope and in the number of speakers used. The advent of OGI_TS corpus based on 10 languages [2] provided the first reliable tool in order to compare the results obtained by human subjects vs automatic systems. A series of experiments were then conducted using spontaneous speech samples from 1 to 6 seconds and different sets of subjects presenting different linguistic backgrounds (i.e. American-English native speakers vs. other). The results showed that increased exposure to each language and longer training sessions contribute to improve discrimination performances. Besides, it showed that previous exposure to languages entailed better performances, while an *a priori* knowledge of the language helped the subjects to develop their own linguistic discriminative cues as the experiment progressed. However, important issues are not addressed in this type of experiment. First, it concerns the arbitrary selection of languages to be discriminated (i.e. languages belonging to different linguistic families and accounting for the most spoken languages in the United States where the corpus was collected), second the hazy spotting and the limited exploitation of discriminative criteria. Indeed, the author merely pointed out the impressionist description of the subjects' identification cues. More recent researches have tried to make up for the lack of preciseness of the linguistic criteria by using either multidimensional analysis so as to explain the subjects' impressions [9] and/or a geographically and typologically refined choice of languages by selecting specific groups of languages (i.e. from the same geographical area) and using languages attesting comparable rhythmic patterns (i.e. tonal languages). Thereafter, the proliferation of experiments based on modified speech (i.e. prosodic and duration-based information) helped to isolate specific linguistic levels — i.e. mostly the suprasegmental level — [10][6][11][12]. They revealed that prosodic cues are felt to be robust enough for language and dialectal identification by different subjects' populations ranging from adults to new born and primates. But although such experiments have provided *a priori* interesting results (i.e. psychoacoustics proprieties of the pitch, syllabic types, phonotactic particularities, rhythmic patterns, etc.), it seems important improvements could still be considered as for (i) the subjects' origin (i.e. other than native American), (ii) the choice of languages used (i.e. bringing together genetic and/or typological features and considering dialectal varieties), (iii) the analysis of the discriminative cues (i.e. specific acoustic analysis), and eventually their concrete exploitation in the scope of automatic language identification (i.e. statistical modeling). The two studies presented below should be regarded as an attempt to integrate these different elements into a global approach.

2. Case Study One : Towards the Automatic identification of Arabic Dialects

Most automatic recognition systems are based on standard linguistic forms. But it is a common observation to say vernacular varieties of languages can be formally speaking quite distant from the standard language itself. The aim of the first study presented here is to demonstrate Arabic hearers from different part of the Arabic speaking world (i.e. Western vs. Eastern) are able to identify accurately the dialectal origin of a given Arabic speaker even though the degree of mutual intelligibility between them is very low or nil.

2.1. Dialectal Geography of Arabic Dialects

The linguistic domain covered by Arabic is widely extended. It goes from Mauritania at the far west to Jordan at the far east. We will show it is yet possible, on the basis of some

phonetic criteria, to set up a linguistic border between Western Arabic dialects and their Eastern counterparts. There are very few researches dealing with the identification of Arabic dialects. The few authors who tackled the issue showed Arabic speakers unconsciously disseminate several dialectal markers (i.e. phonetic, syntactic, lexical, etc.) when producing discourses in Standard Arabic. Our work intends to complete these production-based approaches by studying the way speakers *perceive* dialectal markers.

2.1.1. Corpus and Methods

An acoustic database was elaborated by recording 12 speakers : 6 maghrebine (from Morocco, Algeria and Tunisia) and 6 oriental from Syria, Lebanon and Jordan, who were required to produced a short narration by describing, in their native dialect, a book made of 15 pictures. Recordings were digitized at 22KHz, 16 bits, mono. 96 samples of speech (i.e. complete utterances from 5 to 30 seconds) were extracted and presented at random as stimuli for a perceptual experiment to 18 subjects, native speakers of Arabic from the same 6 countries. The subjects' tasks were (i) to associate the stimuli with specific country as specified above ; (ii) to define the segmental, prosodic and / or lexical cues that allow the subject to identify the dialect in question. In addition to determining the distinctive acoustic cues relevant for the identification of these Arabic dialects, we intended to verify the following theoretical issue : on being exposed to a stimuli, the subjects would be able to identify the relevant dialect area.

2.2. Results

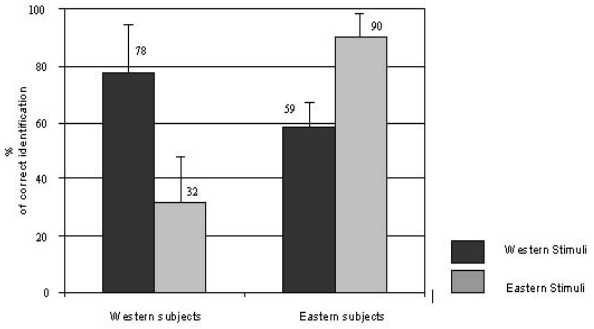

Figure 1: Identification rates as function of subjects' dialectal origin and stimuli's nature (in %)

Significant asymmetries can be observed between the scores obtained by each subjects' category according to the nature of the stimuli. Maghrebine subjects attest 78 % of correct identification for the discrimination of Western stimuli and 32 % for the discrimination of Eastern dialects. The difference between these two rates is statistically significant [$t_{(8)} = 149$, $p < .0001$]. We observe the same disparity between the scores performed by oriental subjects who reach 90% of correct identification for the discrimination of Eastern dialects as opposed to 59% for the discrimination of Western stimuli [$t_{(8)} = 224$, $p < 0001$]. As for cross-area identification task we notice Eastern subjects performed 59 % of correct identification for the discrimination of Western dialects whereas Western subjects obtained 32 % for the discrimination of Eastern stimuli, the difference between these two scores being statistically significant [$t_{(8)} = 4$, $p < .0023$]. These results show the rates obtained for the identification of dialects spoken within the same area than the mother tongue are greater than cross-area identification scores. Besides, our study revealed the influence of external factors on identification abilities : previous exposure to language (maghrebine Arabic in France) increasing subjects' abilities.

2.3. Vocalic characteristics in Maghrebine vs. Oriental Arabic dialects

Acoustic analysis of the vocalic dispersion in maghrebine vs oriental Arabic was performed using spontaneous speech samples. Data of the 6 dialects have been collected in laboratory conditions (22 kHz). Each one the 1500 vocalic segments present in the signal were labeled before an LPC analysis was realized at the middle of steady state of the vowel. As shown on Figure 2 Western dialects tend to develop centralized vocalic segments whereas Eastern dialects tend to privilege the generation of peripheral vowels as far as short and long vocalic segments are concerned.

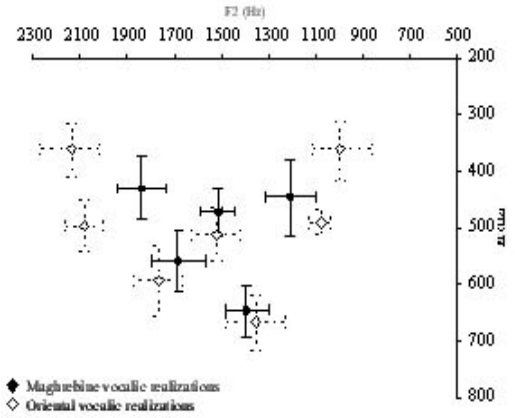

Figure 2 : Vocalic dispersion in western vs oriental Arabic

In Arabic duration opposition is phonological : to each short vowel corresponds a long vocalic segment of the same timbre. Considering the realization of duration in dialectal Arabic (table 1 below) each dialectal variety realizes a contrasting duration ratio (i.e. duration ratio R = long vowel/short vowel). Statistically duration ratios present significant differences from one dialectal area to the other, Western dialects attesting lower values [$t_{(3, 2.35)} = 2.50$, $p = .04$].

Table 1: Mean duration ratio in western vs oriental Arabic

Duration Ratio			
Western area		Eastern area	
Morocco	1.8	Syria	2.3
Algeria	2.0	Lebanon	2.6
Tunisia	2.0	Jordan	2.1
Mean ration / area			
1.9		**2.3**	

2.4. Towards the Automatic Identification of Arabic Dialects

Experiments are performed with data from 20 speakers from various geographical origins. Each speaker has recorded 4

repeats of the text "*The Wind & the Sun*" (International Phonetic Association) pronounced in their own dialect. The mean duration of one repeat is about 30 secondes. 5 speakers are gathered for each dialectal area as training sets. The 10 other speakers are considered for the test. Both training and test procedure involve the four repeats for each speaker (i.e. 40 tests are performed). For each recording, vowels are parameterized [13]. Four vocalic system modeling (i.e. VSM) topologies are studied, with a number of Gaussian component ranging from 5 to 20. The objective is (i) to test if an *automatic* discrimination between the two areas is possible and (ii) what model size leads to the best description of the vowel system of each dialectal area since this optimal size may be related to the vowel system complexity.

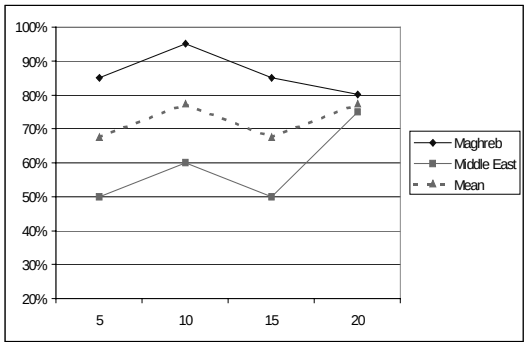

Figure 3 : Correct identification rate for the test set according to the model size. Parameters are 8 MFCC + Duration.

With complex models (20 Gaussian components), the correct discrimination rate reaches a mean value equal to 78 %, and the identification rate of each area are not significantly different. Another set of experiments using longer samples of speech (i.e. 2 minutes) was performed in order to measure the influence of samples' duration on identification performances : the scores reach 90 % of correct identification.

3. Case Study Two : Towards the Automatic Identification of Romance languages

The aim of the second case study is to determine which perceptual categories are salient in Romance languages identification and discrimination. We selected 5 languages from the same family and sharing common linguistic features inherited from Latin, the language mother, to make more complex the discrimination task, thus to favor the emergence of potentially robust discriminative cues. 4 different sets of listeners are tested, consisting in listeners with different mother tongues (French, Romanian, Japanese and American English native speakers).

3.1. Experiments

Data of the five languages have been collected in laboratory conditions (22 kHz, quiet environment, 16 bits, mono) and an intensity normalization is performed. For each language, two speakers (one male and one female) were involved in the learning phase and two other speakers were considered for the test. Speech utterances consist either in read speech or in story-telling speech. 4 sets of subjects were considered. Two sets correspond with native speakers of a Romance language (French and Romanian), while the subjects of the third and fourth sets are Japanese and English American speakers. Each set consists in 20 listeners (male and female subjects). The previous exposure of the subjects to Romance languages is quite homogeneous for each set. The experimentation was divided in two phases: (i) A training phase during which subjects are familiarized with speech samples in each of the five languages, *via* two excerpts of 10-second duration *per* language. Stimuli are presented at random and each stimulus is pronounced by a different speaker ; (ii) A test phase during which subjects listen to 50 stimuli on an AB model: Each stimulus consists in two 6-second utterances separated by a short bell sound. For each stimulus, the subject benefits from a 2-second delay to answer whether A and B are excerpts of the same language or not. No utterance is repeated in the test, and each Li–Lj combination, with $\{i,j\} \in [1,\ldots,5]^2$ is presented twice.

3.2. Results

The percentages of correct answers given by each set of subjects have been analyzed to assess their significance in order to evaluate the relevance of the subsequent multi dimensional analysis. Univariate t-test have been performed to evaluate if the results are significant ($p < 0.001$) or not. In most of the cases, the answers provided by the Romanian, the French and the American subjects are significant On the contrary, answers from the Japanese subjects are not significant. Therefore, a multi dimensional analysis of the Japanese answers is not relevant. Consequently, the analysis is performed for the French, Romanian and American experimental groups of subjects. Both same language and different language stimuli are considered in this study. A 2-dimension display of the answers of the **French subjects** is necessary to take the inter-language differences into consideration (Figure 4). In the (D1/D2) plane, languages cluster in three groups (mother tongue, familiar languages and unknown languages). The first dimension splits mother tongue (French) from foreign languages, while the second dimension distinguishes between familiar languages (Italian and Spanish) and unknown languages Portuguese and Romanian). French subjects seem to be unable to differentiate two unknown Romance languages with only a few seconds of training period. Phonological considerations may also be considered since languages cluster along D2 according to the vowel systems: a three contrast opposition with nasal and central vowels respectively is present in Portuguese and Romanian and not in Spanish and Italian. The analysis of the answers of the **Romanian subjects** receives also a 2 dimensional solution. It leads us to interpret the first dimension in the same way as for French subjects, since it isolates Romanian (mother tongue) from the rest of the languages (Figure 5). The second dimension is interpreted as the complexity of the vocalic system since languages with front/back oppositions in their vocalic systems (Italian and Spanish) are isolated from languages with more than two oppositions and nasal vowels (French and Portuguese). Finally, the same 2-dimensional display was used to explain **American subjects'** result (Figure 6). The first dimension is associated with acoustical cues and 2 clusters are obtained: Romanian, Portuguese, French vs. Spanish, Italian. We can associate this dimension with the vocalic complexity which correspond to each cluster of languages : languages with two oppositions (Italian, Spanish) in their vocalic systems are opposed to languages

possessing three contrast oppositions in their vocalic systems, the third one being central (Romanian), nasal (French, Portuguese) or front rounded. The second dimension opposes familiar languages (French, Spanish) to non familiar languages (Italian, Romanian, Portuguese).

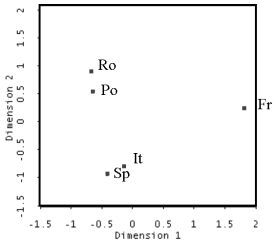

Figure 4 : French subjects (D1/D2)

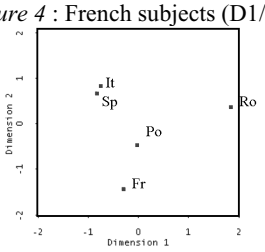

Figure 5 : Rumanian subjects (D1/D2)

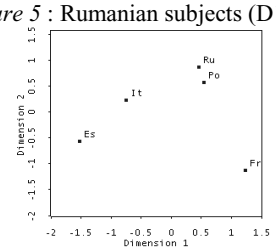

Figure 6 : American subjects (D1/D2)

The results suggest that naive listeners perceive distinctiveness among unfamiliar languages by employing different strategies in language differentiation [14]. These strategies are mother tongue dependent and are based on several types of information, linguistic (segmental) and/or extra linguistic (familiarity). Furthermore, a previous exposure to the Romance languages and the knowledge of at least one of them are fundamental factors in the language discrimination task. Truly naive subjects (Japanese listeners) are unable to extract salient patterns after a brief training period. Our findings suggests that the most relevant linguistic patterns could be characterized as [± vocalic complexity], since two populations (Romanian and Americans) seem to be particularly sensitive to it.

4. Conclusions

The present paper intends to show how automatic systems could draw advantage of human perceptual abilities for language identification in order to improve their performances. Indeed, we demonstrated how perceptual designs could help researchers to target better some of the discriminative cues that are used by human subjects during a language identification task. Besides, our study reveals several external factors — such as previous exposure to language and/or linguistic proximity between the mother tongue and the languages to identify — strongly interfere with subjects' abilities. The acoustic analysis reported in the first case study (i.e. discrimination of Arabic dialects) accounts for a way to verify experimentally subjects' auditory impressions. Results assessed the importance of a precise analysis of the discriminative criteria in order to exploit them properly in the scope of Automatic Language Identification. Eventually, we performed automatic experiments so as to measure the reliability and robustness of discriminative cues and to illustrate the two-step methodology we suggested at the beginning of this work (i.e. from perceptually identified discriminative cues to automatic identification).

5. References

[1] Ohala, J.J. & Gilbert, J. B., " Listeners ability to identify languages by their prosody ", in P. Léon et M. Rossi (Eds), 79, Problèmes de Prosodie, vol II, pp. 123-131.

[2] Muthusamy, Y.K., (1993), "A Segmental Approach to Automatic Language Identification", PhD thesis, Oregon Graduate Institute of Science & Technology.

[3] Lippmann, R. P., (1997). "Speech recognition by machines and humans." Speech Communication 22: 1-15.

[4] Pols, L.C.W., (1997), " Flexible, Robust, and Efficient human speech recognition », Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21, 1-10.

[5] Navratil, J., (1998). A perceptual experiment in language identification. The 8th Czech-German Workshop "Speech Processing", Prague.

[6] Ramus, F., (1999), "Rythme des langues et acquisition du langage", Thèse de Doctorat en Sciences Cognitives, EHESS, 230 pages, Paris.

[7] Barkat, M., (2000), Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes, PhD dissertation, University of Lyon 2, 300 pages, April 2000.

[8] Sugiyama, M., (1991), Automatic Language Recognition using acoustic features. Technical Report TR-I-0167, ATR Interpreting Telephony Research Laboratories.

[9] Stockmal V., Muljani D., Bond Z., (1996), " Perceptual Features of Unknown Foreign Languages as Revealed by Multidimensional Scaling", ICSLP.

[10] Maidment, J.A., (1983), Language Recognition and Prosody : Further Evidence, Speech, Hearing and Language : Work in Progress, U.C.L no 1.

[11] Foreman, C. G., (1999). "Dialect identification from prosodic cues", Proceedings of ICPhS99, San Francisco.

[12] Barkat, M., Ohala, J-J. , Pellegrino, F., (1999), Prosody as a distinctive feature for the discrimination of Arabic dialects, Proceedings of Eurospeech 2000, 395-398,vol. Budapest.

[13] Pellegrino, F. & Barkat, M., (1999), "Investigating dialectal differences via vowel system modeling : Application to Arabic, in Proceedings of ICPhS99, San Francisco, pp. 297-300.

[14] Vasilescu, I., (2001) Contribution à l'identification automatique des langues romanes, PhD dissertation, University of Lyon 2, 250 pages, April 2001.