Automatic Identification of Arabic Dialects

Mohamed BELGACEM, Georges ANTONIADIS, Laurent BESACIER
Université de Grenoble, France Laboratoire LIDILEM & LIG : GETALP
Address
E-mail: mohamed.belgacem@e.u-grenoble3.fr, Georges.Antoniadis@u-grenoble3.fr, Laurent.Besacier@imag.fr

Abstract

In this work, automatic recognition of Arabic dialects is proposed. An acoustic survey of the proportion of vocalic intervals and the standard deviation of consonantal intervals in nine dialects (Tunisia, Morocco, Algeria, Egypt, Syria, Lebanon, Yemen, Golf's Countries and Iraq) is performed using the platform Alize and Gaussian Mixture Models (GMM). The results show the complexity of the automatic identification of Arabic dialects since. No clear border can be found between the dialects, but a gradual transition between them. They can even vary slightly from one city to another. The existence of this gradual change is easy to understand: it corresponds to a human and social reality, to the contact, friendships forged and affinity in the environment more or less immediate of the individual. This document also raises questions about the classes or macro classes of Arabic dialects noticed from the confusion matrix and the design of the hierarchical tree obtained.

1. Introduction and problem

The existence of language's dialects is a challenge to Natural Language Processing (NLP) in general, because it adds another set of dimensions variation from a known standard. The problem is particularly interesting in Arabic and its dialects (J. Meyer et al., 2003). This work aims to highlight the different dialectical phenomena of the Arabic speech and try to build an automatic identification system of Arabic dialects using GMM models. To address this issue we need as first part a vocal corpus. However, the limited work about this topic and the lack of Arabic speech corpora lead us to build one, as a first work.

2. Construction of the corpus

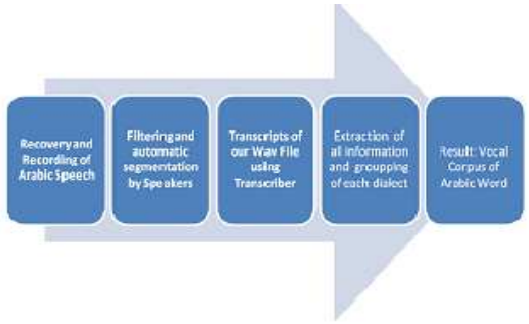
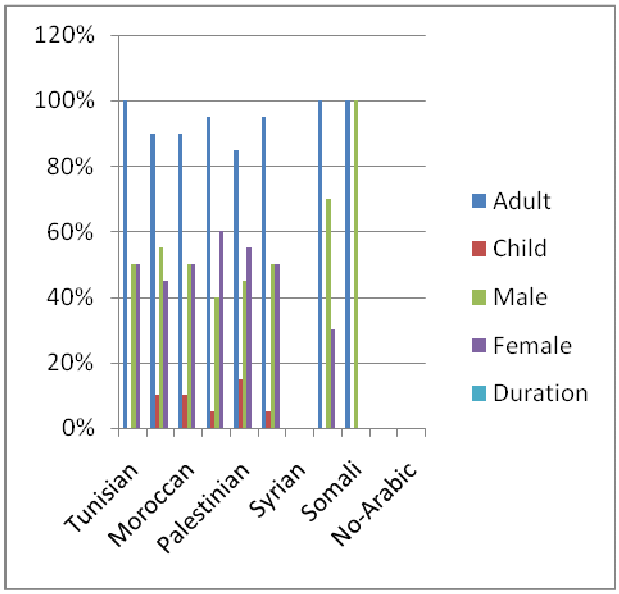


Figure 1: Construction of our dialectal Arabic corpus

2.1 Recording of Arabic Speech

Regarding the collection of vocal data in large quantities for the construction of our corpus of Arabic speech with its different dialects, an interesting approach consists in using broadcast news data.
In the data we have chosen, there is a variety of speakers (Tunisian, Algerian, Moroccan, Egyptian, Iraqi, Armenian, Gulf...) and themes (newspapers, TV shows, political debates, sports, education, social ...). Several people can speak on a given subject successively or simultaneously. The acoustic quality of the recording can change significantly over time.

At this moment, we are particularly interested in news (diary, flash, press reviews, including weather and stock markets, economy, politics, social events ...) in the audio document. Any other form of registration (ads, games, drama ....) will not be transcribed. Following this approach, we recorded the equivalent of 10 hours of Arabic speech with good quality of various dialects from (10 Arabic TV channels and radio: Al-Jazeera, Tunisia, Morocco, Algeria, Egypt, Syria, Lebanon, Yemen…).



* Gulf: This group contains several countries (Iraq, Kuwait, Saudi Arabia, Bahrain, Qatar, Yemen, Oman ...

* Non-native Arabic: English, French, Iranian, Israeli …

Figure 2: Statistics of our dialectal Arabic corpus

2.2 Automatic Segmentation by Speakers

To go faster in the transcript and make clear all information from our speakers, we used an automatic speaker segmentation system (from GETALP team of LIG laboratory). After this automatic segmentation, we did a manual verification to improve performance and assigned to each speaker the necessary informations (name, sex, origin, dialect, studio or telephonic recording ...).

2.3 Transcripts of our Wav File using Transcriber

We describe in this section a set of conventions to structure, annotate and transcribe recordings of radio or broadcast news. The speech of each speaker must be transcribed orthographically. This represents the transcript itself. So far, it is the most important part and therefore maximum attention should be focused on it. The different steps of transcription are: the sound segmentation, the identification of speech parts and the speakers, the identification of thematic sections, the orthographic transcription, and verification. These steps can be conducted in parallel or otherwise applied sequentially over long portions of the signal, depending on the choice of the transcriber.

At this moment, we managed to transcribe only 37% of our corpus (3 hours and a half). Transcribed broadcasts are mixture of Arabic dialects as they are political debates.

3. Automatic identification of dialects

3.1 Introduction

The system designed for this particular task is based on state of the art Gaussian Mixture Models (GMMs) used for Automatic Speaker Verification (SV). We used tools available in "free version" (LIA_SpkDet and ALIZE) (Bonastre, Wils, Meignier. 2005) and developed at LIA. In continuation of this work, we propose here a first study on the extraction of information useful for the characterization of Arabic dialects. We will focus particularly on three sources of informations: (1) complete filtered signal coming from our corpus of different dialects, (2) informations about each speaker (male, female, origin ...), (3) grouping speakers for each dialect. The corpus used in this study and the system of automatic segmentation by speaker are already described in Part One.

3.2 Experiments

The experiments were carried out after "adaptation" of the SV system of LIA. This system (called LIA_SpkDet) is based entirely on ALIZE open platform (Bonastre, Wils, Meignier. 2005) designed and constructed under the Mistral project. LIA_SpkDet is also distributed as free software.

3.3 Protocol

We used in this work, the corpus describes in the first part and to cut out it in two parts:

The corpus of training is composed of 30 one minute sequences of each dialect (Tunisia, Morocco, Algeria, Egypt, Syria, Lebanon, Yemen, Golf' S Countries and Iraq). The sequences of word result from televisual contents in various environments of sound recording (studio, report,

sport but not of simultaneous speakers…).

The corpus of test is composed of 9 one minute sequences of each dialect so that our speakers of the corpus of test are not included in our corpus of training. The sequences result from same televisual contents of the corpus of training.

Set	Dialect	Duration (mn)	Male	Female
Train	9	270	235	235
Test	9	90	45	45
Total	9	360	280	280

Table 1: Division of our corpus: training and test.

3.4 Results

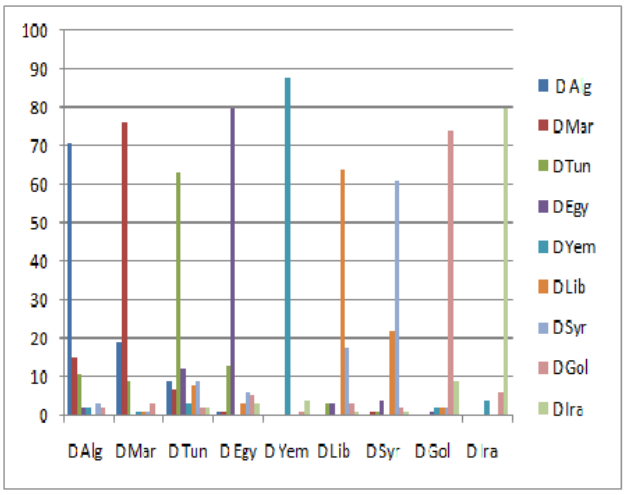


Figure 3: Confusion matrix obtained with GMMs on our Arabic dialects corpus

3.5 Classes of useful informations

Our results show both that our system is able to distinguish the dialects of Arabic-speaking listeners Maghreb and / or the Middle East. It is also interesting to notice that, using a method of ascending hierarchical clustering (see figure 4), we can draw automatically some large families, theoretically and logically related to large geographic areas (these "families" being widely accepted by the linguistic community (Abu-Haidar.F, 1991)). We can distinguish especially the dialects of the east and west, ie the Gulf's Countries and Maghreb (western and eastern Arab world). The first region corresponds to the Gulf countries, Iraq, Yemen and the second is the region of the Maghreb, Egypt, Lebanon and Syria.

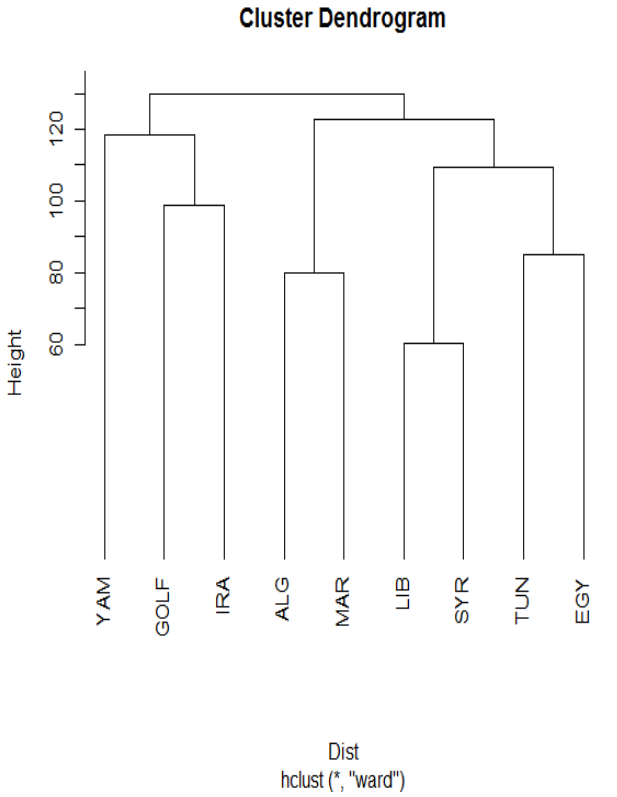


Figure 4: Classification of Arabic dialects using the method of ascending hierarchical clustering

The previous figure shows the concept of classes and subclasses obtained. We can divide Arabic dialects in 2 classes:

- Middle East Dialect: Currency in 2 subclasses Golf's Countries dialect and Yemin, Iraq dialect.
- The second class Dialect is divided into three subclasses, Morocco and Algeria represent the far left, Egypt and Tunisia the right end and Syria and Lebanon are an intermediate zone, but also shows that Syria and Lebanon are closer to the dialects of the Maghreb as the Middle East.

4. Discussion

In this paper, we propose our system of automatic identification of Arabic dialects. This study was conducted using our vocal corpus of different dialects and the SV system of LIA called LIA_SpkDet and based entirely on ALIZE free platform.

From an experimental point of view, the results and the concept of classes obtained by our system seem most interesting in this study, in that they can confirm almost automatically the theoretical results of specialists from Arabic that show the 2 classes in Arabic dialects with the Middle East representing the eastern dialects (Gulf countries) and as second-class dialects of North Africa, Egypt, Lebanon and Syria. Although this division is an oversimplification of Arabic dialectology (Barkat.M, 1999), it is widely accepted by the linguistic community and can be supported by linguistic and geographic data. (Rjaibi-Sabhi.N, 1993).

5. Conclusion

This work presented a system that automatically identifies the Arabic dialects. The difficult task of properly identifying various Arabic dialects was examined. Since the Arabic language has many different dialects, they must be identified before Automatic Speech Recognition can take place. Due to the limited availability of Arabic speech databases, it was necessary to create new data for dialects.
A new model has been presented in this work based upon the features of Arabic dialects; namely, a model that recognizes the similarities and differences between each dialect. The model utilized in this work was the Gaussian Mixture Models (GMM) (Reynolds, Quatieri, Dunn. 2000).
Therefore this new initialization process is used and yields better system performance of 73.33%.

6. References

Abu-Haidar.F. (1991) « Variabilité et invariance du système vocalique de l'arabe standard ». Thèse de Doctorat en Sciences du Langage, Université de Besançon.

Rjaibi-Sabhi.N. (1993) «Approches Historique, Phonologi-que et Acoustique de la Variabilité Dialectale Arabe : Caractérisation de l'origine Géographique en Arabe Standard», Thèse de Doctorat, Université de Besançon.

Barkat.M, Pellegrino.F & Ohala.J.J. (1999) « Prosody as a Distinctive Feature for the Recognition of Arabic Dialects ». Dans Proceedings of Eurospeech'99, 395- 398, Budapest (Hongrie).

Reynolds D. A, Quatieri T.F, Dunn R.B. (2000) « Speaker verification using adapted Gaussian mixture models, digital signal processing (dsp) ». In a review journal - Special issue on NIST 1999 speaker recognition workshop 10 (1-3), pages 19–41.

J. Meyer, F. Pellegrino, M. Barkat-Defradas & F.Meunier. (2003) « The Notion of Perceptual Distance: The Case of Afro-Asiatic Languages». Dans Proceedings of the 15th International Congress of Phonetic Sciences, Barcelone, 3-9 September 2003.

Bonastre J.-F., Wils.F, Meignier.S. (2005) « Alize, a free toolkit for speaker recognition ». In ICASSP-05, Philadelphia, USA, volume 39, pages 430–451