ALGASD PROJECT: Statistical Study of Vocalic Variations according to Education Levels of Algiers Speakers

G. Droua-Hamdani (1), S. A. Selouani (2), M.Boudraa (3) B. Boudraa (4)
(1) Speech Processing Laboratory (TAP), CRSTDLA, Algiers, Algeria. Email: gh.droua@post.com.
(2) LARIHS Laboratory. University of Moncton, Canada selouani@umcs.ca
(3-4) Speech Communication Laboratory, USTHB, Algiers, Algeria. mk.boudraa@yahoo.fr

I. ALGASD Database

I.1 Description

Algerians' official language is Modern Standard Arabic (MSA). They learn it at school from the primary to the secondary levels. As elsewhere, written MSA differs from the adopted spoken language (mother tongue).

72 % of Algerian people speak in their daily life the *Darija*, "Algerian Arabic dialects". They are variants of the MSA stemming from the ethnic, geographical and colonial occupiers influences as: Spanish, French, Turkish, Italian, etc. That, within Algerian Arabic itself, there are significant local variations (in pronunciation, grammar, etc.) observed from town to town even they are near to each other.

In addition to the *Darija*, 28% of inhabitants speak Berber language. However, some of Algerians use in their life French language.

Our project consists on conception and realization of Algerian speech database with (MSA) as substratum. All selected speakers are chosen among Algeria's regions principally in areas presenting important regional variations in pronunciation. So, we divided statistically a number of 300 Algerian native speakers on 11 areas. 8 of them are located in the North of the country as it is the most populated area: 3 regions from the center (Algiers, Tizi Ouzou and Medea), 3 regions from the East (Constantine, Annaba and Jijel) and 2 regions from the West

(Oran and Tlemcen). We choose from the South 3 other regions which are (Bechar, Ghardaïa and El Oued).

In order to get all social levels among the population, we elaborated a speaker profile which considers the age and the academic level of each speaker of the database.

I.2 ALGASD Texts Material

ALGASD text material is elaborated from set of 200 Phonetically Balanced Arabic Sentences called Reference Corpus. From the latest, we conceived three different sub-sets of texts material which are distributed on the 11 regions. Every sub-corpus aims to provide us a specific acoustic-phonetic knowledge:

-Common corpus (Cc) composed by 2 sentences was read by all the speakers (300). It serves to brew the regional varieties of pronunciations

-Reserved corpus (Cr) included 30 sentences is read by 86 speakers. Its aims to contain all Arabic phonemes and observed oppositions in the Arabic language.

-Individual corpus (Ci) is constituted by 168 sentences of the reference corpus. These ones are used to gather a maximum of allophonic contextual differences. They are read by 222 speakers.

Total number of ALGASD voice bank achieves to 1080 utterances. 55.5% are obtained from Cc texts' recordings, 24% from Cr texts and finally 20.5% from Ci texts.

From ALGASD, we realized an acoustic analysis on the correlation between a daily spoken MSA and education levels of Algerian speakers. This study aims to show different particularities observed in their vocalic system.

II. Statistical Analysis of Daily Spoken SA and Education Levels Influences on Algiers Vocalic Pronunciation

To learn MSA, Algerians go to school for approximately 13 years. This period is divided on three academic levels: primary, middle and secondary school. Afterward, some of them go to universities where they prepare graduation or post-graduation diploma.

In this study, we took of the complete ALGASD only sound corpora relative to Algiers which corresponds in our code to R1. The acoustical analysis is realized for 50 speakers from 80 of R1 (50% male 50% female).

We divide R1 sound files into 2 categories:

a) - Average Category (C1) which includes sound files of 16 speakers whom have studied until secondary level. Usually, these speakers don't use MSA in their daily and professional life.

b) - High category which is composed by 34 speakers. They have followed the university studies (graduation and post-graduation). Afterward, we divided this set in two sub-categories:

- Those using daily MSA in their professional lives such as teachers in human sciences universities, journalists of the Arabic press and lawyers. This category (C2) is constituted by 19 speakers.

- Those that do not use MSA in their daily life or very rarely such as: teachers of technical universities, French press, Doctors, etc. They employed in their professional occupations French language rather than Arabic. Total speakers of this category (C3) are 15 ones.

III. Results

The study is based on statistical analysis of formants values (F1, F2) and durations of all vowels (brief, long). We check by that elaborating of qualitative and quantitative approaches of the vocalic distribution according to different speakers categories.

We estimated in the first time ratios relative to duration's diminution or lengthening for two types of vowels (brief, long) for all 3 categories, and in the second time we proceeded to discriminative analysis which leads to different statistical results. These latest reveal a significant interaction between academic levels of Algiers speakers and their daily use of the Arabic language.

This Influence is principally observed in their vowels quantity (duration) rather than in their vowels quality (formants). Figure.1 shows the interaction between vowels' length and academic levels of Algiers speakers.

Figure .1: Interaction between vowels durations and academic categories