ISCA Archive
http://www.isca-speech.org/archive

First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

Phonetic Study for Automatic Recognition of Arabic

M. Djoudi D. Fohr J. P. Haton

CRIN — INRIA Lorraine
Campus Scientifique
B.P. 239
54506 Vandœuvre-lès-Nancy CEDEX
France

Abstract

We propose in this paper a phonetic study of standard Arabic based essentially on the spectrographic visions of 50 sentences of the DJOUMA corpus we have constituted. This study allowed us to determine the pertinent parameters of continuous speech recognition. For the recognition part, we present the algorithms developed and the results obtained for three important questions :
The segmentation of speech into gross phonetic classes. This process is based on methods adapted from the APHODEX expert system developed in our group [Foh 86], [CHF+86].

The recognition of vowels in continuous speech for multispeakers.

The detection of parameters necessary for identification of plosives and fricatives.

1 Introduction

The automatic recognition of Arabic poses several specific problems due to the characteristics of this language [Sal87], {Hom87]. In particular, due to the fact that Arabic is a consonant language and that the role of vowels is always carried on second range. In order to bring out the pertinent parameters for recognition, we made a spectrographic analysis of standard Arabic phonemes. The recognition process consists in large segmentation and labeling each segment using values of characteristic parameters

2 Phonetic study

2.1 The DJOUMA corpus

For the need of the phonetic study of modern standard Arabic, phoneticians and computer scientists were obliged to develop corpus. Most of these corpora were principally corpus of words or of CV and CVC type which do not verify the criteria of taking into account the real context of productions of phonemes and coarticulation phenomena. We thought that was necessary to have, at first, a read sentences corpus. Our corpus is baptized DJOUMA after /DJOUmal MAqroua/, which means in Arabic "Read Sentences".

2.1.1 Corpus constitution

In order to find the phonetically balanced Arabic sentences, like in the case of French [Com 81], we have searched in Arabic magazines and books to find the correct sentences. We fixed at the outset three criteria :

- the simplicity of the syntactic form.
- the diversity in the form and the content of sentences.
- the reasonable length of the sentence.

Using these basic criteria, we could construct 78 sentences. After that, we did a sort of sounding to three persons who read together the sentences. Our aim is to eliminate the difficult-to-understand or pronounce sentences. The remaining 50 sentences were entirely vocalized so that, according to standard Arabic grammatical rules, the pronunciation will be exact. Afterward, we distributed these sentences in 5 series of 10 sentences, trying to equilibrate phonetically every Series. Quoted below are the distribution of consonants and vowels in the corpus.

Series	Consonants Nb.	Vowels Nb.	Total Nb.
A	159	131	290
B	172	134	306
C	168	135	303
D	169	138	307
E	166	138	304
Total	834	676	1510

2.1.2 Recording

The recording has been done in a calm medium in a magnetic cassette. A set of 7 males and 3 females, pronounced each, a Series of 10 sentences. A two second duration separated the pronunciation of two consecutive sentences. The sentences were sampled at a 16000 Hz frequency and each one was stored in a disk file which has a name in an alphanumerical code such as ANN.AA or ANN :AA where :

- A designates the Series (A, B, C, D or E).
- NN is the sentence number (01 ... 50).

– AA is the name speaker initials (ex : DM).
– The point "." means male and two points " :" female.

We developed software tools of display of the numerical spectrograms, so as to be able to lead spectrographic analysis and in absence of phonetic expert we labeled ourselves manually the set of the sentences using the exact pronunciation.

2.2 Vowels

The Arabic vowel system is generally described as being comprised of six vowel sounds. To each of the short vowel /a/, /i/ and /u/ corresponds a vowel of a longer duration /aa/, /ii/ and /uu/. The temporal opposition short vs. long is fundamental at the grammatical and semantic levels [Bel84]. The relative duration of vowels depends on the environment and how fast an individual speaks. The next table yields in ms the average values of duration of the vowels.

Vowel	Duration
a	79
i	73
u	67
aa	167
ii	158
uu	149

At the final position, short vowels are characterized by a longer duration and the long ones by a less important duration. The study of the the temporal organization of the vocalic quantity is not limited to the measurement of the vocalic duration, it necessitates taking into account the adjacent consonant revelation.

Context plays an important role, and vowels vary intensively depending on the place of the articulation of the adjacent consonant, in particular, the spectrographic observations attest a spectral difference related to the opposition emphatic/non-emphatic. Vowels in contact with emphatic consonants are distinguished by a marked lowering of the second formant and a slight rising of the first formant, but it seems that this type of effect is not characteristic of emphatics only. The pharyngeal consonants act on the same way on the formantic structure of the vowels in contact. We give the average values of vowel formants, in different contexts, in the next table.

2.3 Consonants

Arabic contains 28 consonants, to each of them corresponds a particular phoneme. The peculiarity of the system is based on the glottal, pharyngeal and emphatic consonants [Gha87]. The glottal and pharyngeal consonants are distinguished from the rest of the consonants by having distinct vertical places of articulation. A vertical place of articulation is defined as a set of anatomical locations from palate to the glottis, inclusive. In contrast, an horizontal place of articulation is from the lips to the uvula, inclusive [Ani70].

The emphatic consonants are described like having a second place of articulation at the pharynx level [GP82], [Bon77]. According to the manners of articulations, we distinguish :

Plosives which are physiologically characterized by the formation of closure within the vocal cavity by one or more articulators where the driving pressure is blocked and by the sudden release of that pressure (burst).

Context	Formant	a	i	u	aa	ii	uu
	F1	600	295	290	585	289	305
Bilabial	F2	1585	1920	850	1200	2200	775
	F3	2400	2650	2175	2585	2700	2370
b	F4	3513	3413	3059	3510	3360	3016
	F1	579	372	312	552	290	286
Dental	F2	1255	1680	921	1393	2008	881
	F3	2482	2723	2216	2423	2882	2338
t	F4	3312	3522	3216	3370	3479	3150
	F1	615	340	320	609	218	2308
Palatal	F2	1450	2100	790	1405	2140	800
	F3	2411	2730	2420	2438	2753	2456
ʃ	F4	3510	3375	3116	3487	3401	3162
	F1	639	332	312	612	320	346
Velar	F2	1345	1720	891	1403	2098	831
	F3	2510	2752	2236	2501	2722	2254
k	F4	3312	3522	3216	3370	3479	3150
	F1	650	340	320	620	340	350
Emphatic	F2	1200	1600	850	1200	2050	810
	F3	2600	2700	2275	2650	2750	2335
t	F4	3434	3580	3146	3562	3430	3110
	F1	625	342	324	615	331	340
Pharyngeal	F2	1280	1660	810	1300	2100	810
	F3	2630	2700	2289	2673	2700	2309
h	F4	3447	3597	3150	3544	3412	3127
	F1	645	340	313	627	340	350
Glottal	F2	1221	1644	842	1239	2037	814
	F3	2608	2695	2256	2664	2747	2328
?	F4	3402	3578	3152	3550	3419	3100

Fricatives are produced in the vocal by a narrow constriction that causes the airflow to be consistly turbulent.

Nasals are characterized by the formation of one or more oral closures as air flows through the nose.

Trill : A movable articulator is set into vibration by the flow air. The airstream is repeatedly interrupted by this obstruction.

Sonorants posses more acoustic features similar to those of the vowels than any of the other consonantal group. Detailed study of all the phonemes of standard Arabic is available in [Djo89]

3 Segmentation

It is achieved by a set of three modules and consists in the segmentation of the speech signal into large phonetic categories by using non contextual and procedural algorithms. The aim of segmentation is :

– Reducing combinator explosion during the recognition.
– Allowing a centering for automatical labeling.

We have carried three large classes :

– fricatives { /z/,/f/,/θ/, /s/, /ʃ/,/χ/, /h/,/z/, /s/ } and burst,
– vowels { /a/, /i/, /u/, /aa/, /ii/, /uu/ },
– plosives { /t/, /k/, /?/, /b/, /d/, /q/, /t/ }.

we did not take the rest of fricatives and also the plosive /d/ because they present, for all the speakers, the characteristics that lay near of the ones of sonorants.

3.1 Vowels

For the determination of the vowels, we use :

- The energy curve in 250 to 2500 hz frequency band.
- The total energy curve.

The first curve is obtained after summing among the canals corresponding to the frequencies between 250 and 2500 Hz those that reach the visibility threshold on the spectrogram [CHF+86]

The second is obtained by calculation of the energy of the temporal signal.

We search on the two curves the maxima that verify :

- An intensity equals, at least, to the half energy of precedent pick.
- A sufficient right and left valley.
- The presence of voicing.

This procedure allows the performance of an average vocalic duration, that give us an indication of the speed of elocution.

3.2 Fricatives

We calculate two curves :

- Zero crossing curve on the signal filtered by a high pass filter at 800 Hz.
- A gravity center curve, calculated on the visible part of numerical spectrograms.

A fricative is detected if we put in evidence a local maximum on the two curves.

3.2.1 Plosives

We calculate an energy curve on the temporal signal preaccentuated and filtered by a high pass filter at 600 Hz. the plosives correspond to a local minimum on this curve.

4 Classification and labeling

4.1 Vowels

From the vocalic segments found during segmentation, on each sample, we compute the first three formants as being visible picks in the frequency bands [250-850], [750-2300] and [1800-2800] Hz, respectively for F1, F2 and F3. These picks are performed from the sampled temporal signal LPC coefficients at 16000 Hz. For each vowel of Arabic, we estimate a distance ratio between the formants of the segment and the average values of formants of vowel. These last ones are determined experimentally during the phonetic study. Afterwards, we recover the value of the average vocalic duration to use it at a fuzzy threshold between short and long vowels. According to the difference between values of formants and duration, we assign a score for each Arabic vowel.

4.2 Fricatives

At the actual state of the system, for identification of the fricatives, we use the parameters :

- Voicing degree of segment,
- Low limit of friction noise,
- Gravity center of noise,
- and transition of formants F1, F2 and F3 of the next vowels.

4.3 Plosives

For the plosives [DDLO83], we use the following parameters :

- Voicing degree of the segment,
- The presence of the burst [Djo86],
- The value in Hz of gravity center of burst,
- The concentration degree on the burst,
- and transition of formants F1, F2 and F3 of the next vowels.

4.4 Fuzzy information processing

For the evaluation of the score of each phoneme in function of the characteristic parameters, we have fixed an initial score for each parameter in function of its pertinence. The combination of scores to obtain the final score, using the following fuzzy functions is as follows :

$super(x, d, f)$: that gives 1 if $x > f$; 0 if $x < d$ and linearly a value between 0 and 1 if x is between d and f,

$infer(x, d, f)$: that gives 1 if $x < d$; 0 if $x > f$ and linearly a value between 0 and 1 if x is between d and f,

$in(x, d, f)$: that gives 1 if x is between d and f ; 0 else.

$middle(x, d, f)$: that give 1 if x is in the middle of the segment df ; 0 if $x < d$ or $x > f$; and linearly, a value between 0 and 1 if x is between d and the middle or x is between the middle and f.

5 Results

We have tested the segmentation algorithms on the corpus DJOUMA, the results we indicate are calculated by comparison with the manual labeling for 3 male speakers.

Class	Present Nb.	Found Nb.	Inserted Nb.
Fricatives	200	185 (93%)	8 (4%)
Plosives	288	279 (97%)	9 (3%)
Vowels	676	642 (95%)	27 (4%)

The relatively important insertion rate of plosives is explained by the fact that many /f/ are labeled at the same time plosives and fricatives, and because /m/ is labeled many times as plosive.

the omissions of vowels produce in the context "vowel-sonorant-vowel" where the energy variations are very weak. In this case, the system finds only one vowel of two or gathers this in a unique large group.

The omissions of fricatives are due to /f/ and /h/. For the identification, the three phonemes that totalize the best scores are held as being labels of the segment. The percentage of correct labeling is given in the next table.

Class	Percentage
Vowels	95
Fricatives	87
Plosives	83

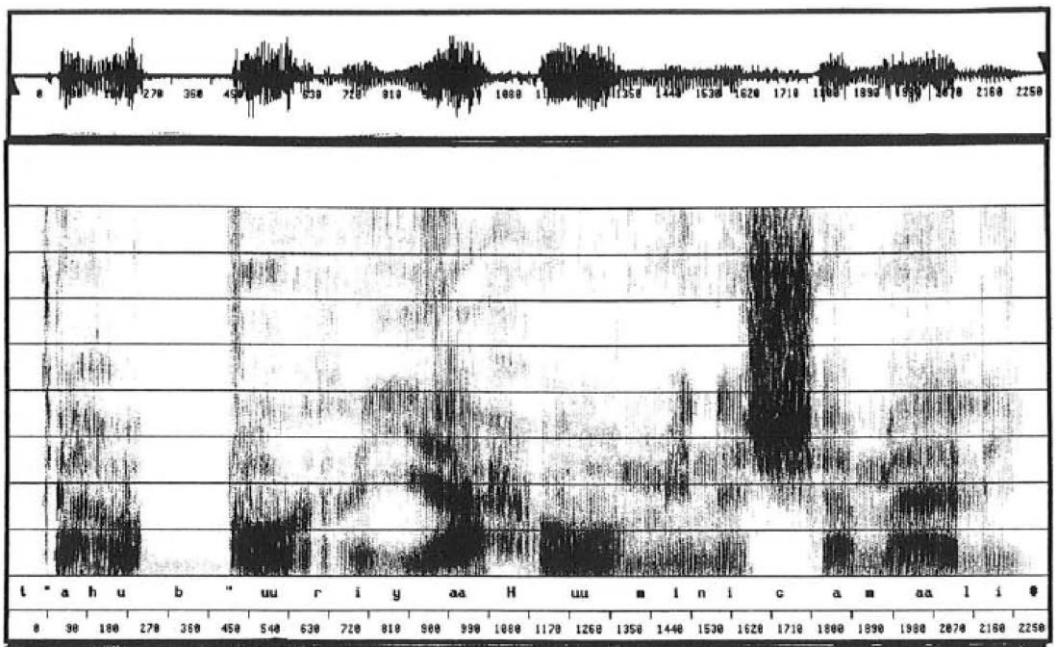

FIG. 1: Temporal signal And spectrogram
Sentence translation of *The winds blow from the nort*

6 Conclusion

We have presented a phonetic study of standard Arabic which allowed us to compute the pertinent parameters for continuous speech recognition. The work takes its originality from the taking into account of the vowels in all contexts including the emphatic consonant neighbourhood. The perspectives to give to this work is to develop the labeling algorithms for the other phonetic classes and the improvement of recognition scores by the introduction of expert knowledges.

References

[Ani70] S. H. Al. Ani. Arabic Phonology. An Acoustical and Physiological Investigation. Mouton & Co N.V., 1970.

[Bel84]Y. Belkaid. Les voyelles de l'Arabe littéraire moderne. analyse spectrographique. Technical Report 16, Travaux de l'institut de phonétique de Strasbourg, 1984.

[Bon77] J. F. Bonnot. Recherche expérimentale sur la nature des consonnes emphatiques de l'Arabe classique. Technical Report 9, Travaux de l'institut de phonétique de Strasbourg, 1977.

[CHF+86] N. Carbonell, J. P. Haton, D. Fohr, F. Lonchamp, and J. M. Pierrel. APHODEX, design and implementation of an acoustic-phonetic decoding expert system. IEEE International Conference onAcoustics, Speech and Signal Processing, 1986.
[Com81] P. Combescure. Vingt listes de dix phrases phonétiquement équilibrées. Revue d'Acoustique, 14(56), 1981.

[DDLO83] P. Demichelis, R. DeMori, P. Laface, and M. OKane. Computer recognition of plosive sounds using contextual information. IEEE Trans. Acoust., Speech, Signal Processing, ASSP-31(2), 1983.

[Djo86] M. Djoudi. Détection et localisation de la barre d'explosion en parole continue et dans un contexte multilocuteur. Rapport de D.E.A, Centre de Recherche en Iinformatique de Nancy, 1986.

[Djo89] M. Djoudi. Etude phonétique de l'Arabe standard. Technical Report 89-R-057, Centre de Recherche en Informatique de Nancy, 1989.

[Foh86] D. Fohr. APHODEX : Un système expert en décodage acoustico-phonétique de la parole continue. Thèse de Doct. Univ. de NANCY 1, 1986.

[Gha87] S. Ghazali. Elements of Arabic Phonetics. In Applied Arabic Linguistics and Signal & Information Processing, pages 51–58. Hemisphere publishing corporation, 1987.

[GP82] A. Giannini and M. Pettorino. The Emphatic Consonants in Arabic. Giardini editori e stampatori, 1982.

[Hom 87] J. M. Hombert. Acoustics phonetics. In Applied Arabic Linguistics and Signal Information, pages 27--58, 1987.

[Sal87] A. Hadj Salah. Arabic Linguistics and Phonetics. In Applied Arabic Linguistics and Signal & Information Processing, pages 3–22. Hemisphere publishing corporation, 1987.