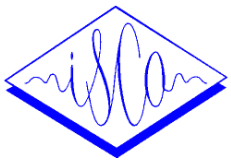

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

The SAPHA Acoustic-Phonetic Decoder System for Standard Arabic

M. Djoudi J.P. Haton

CRIN — INRIA Lorraine
France

Abstract

We present in this paper the SAPHA system is developed for the purpose of the acoustic-phonetic decoding of standard Arabic.
First of all, we give a description of the general architecture of the system and the various modules which compose it. Afterwards, we develop the principal stages of the Arabic decoder, namely :

- segmentation of the speech signal into large phonetic classes.
- automatic extraction of the parameters that are pertinent to phonetic recognition of standard Arabic.
- multispeaker identification of phonemes in continuous speech by using an expert system based on production rules.

The results obtained in broad phonetic classfication and phoneme labelling for three male speakers are also presented and discussed.

1 Introduction

We describe an acoustic-phonetic decoding system of standard Arabic, which makes possible the analytic recognition of phonemes in continuous speech for multi-speakers. Combining other linguistic information (lexicon, syntax, semantics and pragmatics) the system can be considered as an important step towards a system for oral dialogue between man and machine [8]. It also acts as a module of a dictation machine controlled by a voice in standard Arabic [4]. The system receives the speech signal previously digitalized and as a result sends back a phonetic lattice. As for the recognition module itself is realized in the form of an expert system based on production rules [3] [6]. Knowledge is acquired after a phonetic study of Arabic [1] [2] [7] carried out on DJOUMA corpus, which is composed of 50 sentences pronounced by 10 speakers (7 males and 3 females). This study also allowed to adopt a strategy for signal segmentation into large phonetic classes and to determine the distinctive values of the parameters used during recognition [5]. In the same way, the manual labelling of the corpus sentences allowed to test the performances of the system.

2 ARCHITECTURE OF THE SYSTEM

The SAPHA decoder is made up of a set of modules corresponding to the acoustic, phonetic and phonological levels [6]. These modules are (cf figure 1):

Acquisition module

Speech acquisition is made from a microphone or a tape, or else a speech file ready for use on a magnetic disc. We can also listen to the recorded speech signal, change the value of the sampling frequency (by default: 16 kHz) and interactively display and manipulate the signal.

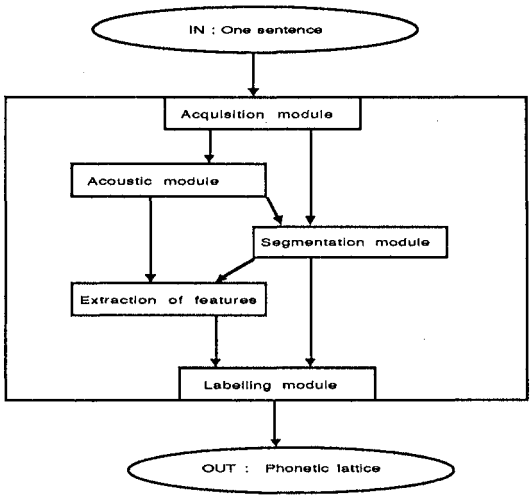

Figure 1: Architecture of SAPHA

Acoustic module

This module extracts the following acoustic parameters from the temporal signal,

- signal energy which can be used to discriminate between speech and non-speech, between voiced and unvoised sounds. It also provides data on intonation;
- number of zero-crossing per second: this parameter, easily computable, is also used to disciminate between speech and non-speech, and between voiced and unvoiced sounds;
- LPC coefficients, which allow to detect the peaks corresponding to formant frequencies;
- frequency features computed by FFT directly from the signal or from the LPC coefficients.

Segmentation module

This module performs the segmentation of the speech signal intobroad phonetic classes by using non-contextual algorithms based on simple criteria. The essential goal of the segmentation is to reduce the combinatorial explosion during the recognition and to allow a centering for an automatic labelling.

Extraction of features

The extraction of pertinent phonetic features is a very important stage in the process of phonetic decoding. The values of these indices will be used at the moment of the activation of a rule for the labelling module (cf section 4).

Labelling module

It is at this level that the proper decoding takes place. From the segments provided by the segmentation module, this module tries to find out the prounounced phonemes by using the indices extracted at the previous stage and the rules of the knowledge base.

3 Segmentation into broad phonetic classes

This segmentation is carried out by a set of three modules and consists of the segmentation of the speech signal into three large phonetic classes, ie :

- vowels { /a/, /i/, /u/, /aa/, /ii/, /uu/ }
- plosives { /t/, /k/, /ʔ/, /b/, /d/, /q/, /ṭ/ }.
- fricatives { /ʒ/,/f/,/θ/, /s/, /ʃ/,/χ/, /ħ/,/z/, /ṣ/ } and burst,

The rest of phonemes is put into the class of unknown.

Segmentation of vowels

To select the vowels, we use the energy curve in a frequency band ranging between 250 and 2500 Hz (where the first two formants are located) and the total energy function. The first parameter is obtained by adding up among the channels in the band 250-2500 Hz those which reach the threshold of visibility on the spectrogram. The second is obtained by computing the energy of the temporal signal. Then, on the variation curves of these two parameters we locate the maxima that verify

- an intensity at least equal to half the energy of the previous peak,
- a sufficient right and left valley,
- the presence of voicing.

This module allows the performance of an average vocalic duration, that gives an indication of the speed of elocution.

Segmentation of plosives

For the determination of plosives, we compute an energy curve on the temporal signal filtered by a high pass filter at 600 Hz. Plosives correspond to local minima on this curve.

Segmentation of fricatives

To select fricatives, we compute two curves : a zero crossing curve on the signal filtered by a high pass filter at 800 Hz, and a center of gravity curve computed on the visible part of spectrograms. A fricative is detected if a local maximum on these curves is identified.

Processing of inclusions

Each of the three procedures described above returns a list of detected segments (beginning, end, position of extremum, coefficient of certainty). The three lists must then be merged in order to obtain a first segmentation under the form of lattice. We first deal with the problem of inclusions. If two segments are included into each other according to the position of the maxima, either we generate a segment with two features (for example fricative and plosive for an /f/) or we generate two segments (for example plosive and fricative when the plosive has a fricative burst /ti/). On this second assumption, the boundaries of the segments are calculated by spectral difference.

4 Extraction of features

We will describe the way phonetic features are extracted from the speech input.

Segment duration

It is simply the segment length converted into milliseconds.

Degree of voicing

It is the ratio between the number of voised samples and the total number of the samples of the segment.

Position of the burst

Used to identify the plosives, the burst is an explosion of energy of short duration. The basic idea is to detect the burst in various frequency bands as being the position where partial energy is at a maximum. The final position of the burst is the sample which is most often detected in the frequency bands and which corresponds well to its visibility on the spectrogram.

Characteristics of the burst

When a burst is detected, we compute the ratio between the maximum and the average energy and we project its value in the [0,1] range. Moreover, we calculate the position of the center of gravity.

Formant tracking

The formants of the vowels play a very important role for the vowel identification task. For each sampling of the vocalic segment from the LPC coefficients, we compute the first relevant peaks which will be applicants for the positions of formants. Then, we calculate the first three formants as being visible peaks in frequency bands [250-850], [750-2300] et [1800-2900] Hz respectively for F1, F2 and F3.

Formant transitions

To take into account the phenomena of co-articulation, we study the formant transitions CV or VC at the limit between a vowel and the adjacent consonant. Then, we decide for each formant if the transition is upward, downward or flat. Therefore, we take a space at the limit of the vowel and we pass the values of the formant through a procedure of linear regression.

Lower limit of noise

This parameter is very important to discriminate between fricatives. To calculate this limit, we first estimate the lower threshold of visibility on each segment sample, then, we calculate the average on the entire segment.

Center of gravity

For a given segment, we compute the frequency which corresponds to the average center of gravity computed on each sampling. This parameter is especially used for the recognition of fricatives.

5 Phonetic labelling

The SAPHA labelling module is an expert system based on production rules. It consists of a knowledge base of phonemes identification and in an inference engine. This system has already been used for the phonetic of French in the framework of the APHODEX project [3]

Knowledge base

The acquired knowledge is coded under the form of production rules. There is an example of rule:

```
R135
C Rule for /t./ context i ii C
CONTEXT_RIGHT [ i ii ]
IF
burst-present_ACT &
^(burst-freq_ACT 5200 6800)
THEN [ t. 8 ]
```

The knowledge base is presently made up of 193 rules.

Measures base

To each segment is assciated a measure base, which contains the results of procedures of signal processing applied to this segment.

Inference engine

The inference engine allocates a list of one or several phonemes to each segment detected by the segmentation module. The activation of a rule depends on the conditions expressed in left and right contexts, on the conclusion and on segmentation. A plausibility is assigned to each hypothesized phoneme. The conclusion of a rule may also be an action carried out or a list of phonemes. To process the fuzziness that can be found in the rules, the engine uses a reasoning mechanism based on fuzzy logic.

6 Experimental results

We have tested the segmentation algorithms on the DJOUMA data base that was manually segmented and labelled. The results given below are computed by comparison with this manual labelling for three male speakers.

class	number present	number found	number inserted
vowels	676	642 (95%)	27 (4%)
plosives	288	279 (97%)	9 (3%)
fricatives	200	185 (93%)	8 (4%)

Omissions of vowels occur in contexts VCV (where C is sonorant), in which energy variations are very low. In this case, the system finds out only one of the two vowels or puts together the two vowels in one large group. The relatively important rate of the insertion of plosives can be explained by the fact that a large number of /f/ are labelled as both plosive and fricative and above all by the fact that /m/ is labelled several times as a plosive. For the recognition part, the percentage of correct labelling (three labels by segment) is given in the following table.

Class	rate
Vowels	94
Plosives	80
Fricatives	83
Unknowns	71

The good performance for vowels comes from the simplicity of the vocalic system of Arabic and the absence of nasal vowels. The relatively moderate rate of plosives and fricatives is explained by the fact that the knowledge base is not yet totalled assessed and completed.

7 Conclusion

We have presented in this papaer an acoustic-phonetic decoding system of standard Arabic using knowledge-based techniques. The prospects of the present work are to segment more finely the unknown class, to design for a new form of the knowledge representation so that the context is actually taken into account and to work in collaboration with an expert phonetician in order to obtain more reliable and consistent knowledge.

References

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

[2] J. Cantineau. *Cours de phonétique arabe.* Librairie Klincksieck, 1960.

[3] 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 on Acoustics, Speech and Signal Processing*, 1986.

[4] M. Djoudi, D. Fohr, and J. P. Haton. MARS: un système de reconnissance de l'arabe moderne. *Actes des 18ème Journées d'Etudes sur la Parole*, 217–221, Mai, 1990.

[5] M. Djoudi, D. Fohr, and J. P. Haton. Phonetic Study for Automatic Recognition of Arabic. *Proceedings of European Conference on Speech and Technology*, 2:268–271, September 1989.

[6] M. Djoudi, D. Fohr, and J. P. Haton. SAPHA: un système expert pour le décodage acoustico-phonétique de l'arabe standard. *Première Conférence Maghrébine sur le Génie logiciel et l'Intelligence Artificielle*, Septembre 1989.

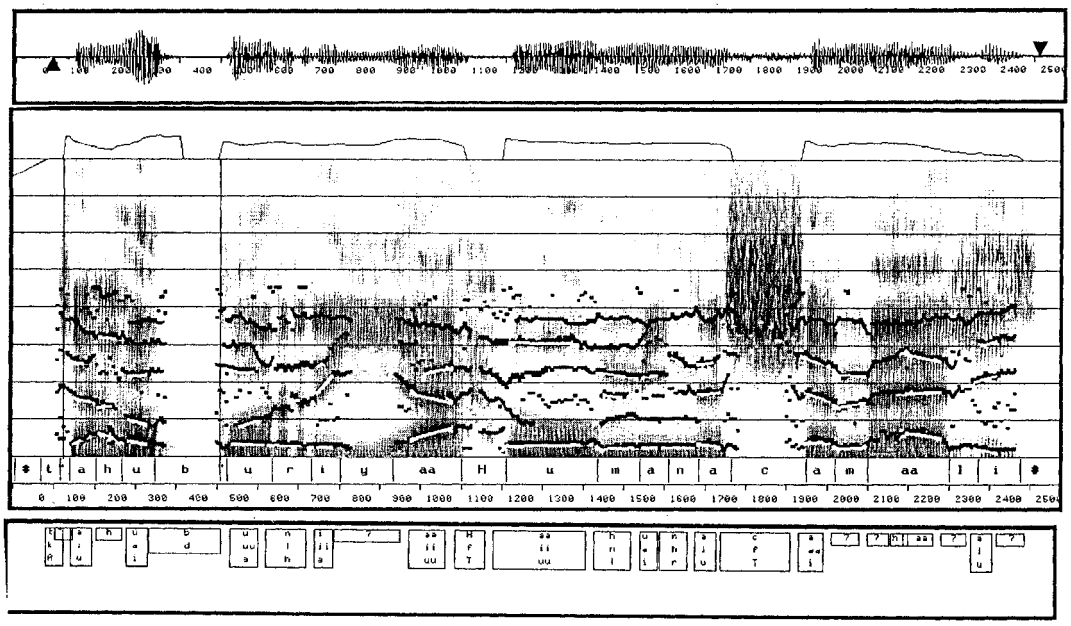

Figure 2: Temporal signal and spectrogram

Sentence: translation of:
'The winds blow from the north'