**SETIT 2009**
5th International Conference: **S**ciences of **E**lectronic, **T**echnologies of **I**nformation and **T**elecommunications
March 22-26, 2009 – TUNISIA

# New Approach of Speech Phonetic Segmentation Applied at Arabic Language

Houda HOSNI*, Zied SAKKA*, Abdennaceur KACHOURI* and Mounir SAMET*

* *Laboratory of Electronic and Information Technologies (LETI)*
*National engineer school of Sfax B.P.W, 3038 Sfax, Tunisia*

**houda.hossni@gmail.com**

**zied.sakka@yahoo.fr**

**mounir.samet@enis.rnu.tn**

**Abdennaceur.kachouri@enis.rnu.tn**

**Abstract:** In the light of current scientific knowledge, the development of several speech treatment fields has in parallel evolved with a huge amount of researches concerning automatic speech segmentation. In this paper we investigate the development of a precise speech segmentation system, especially applied on Arabic language. The main goal of this work is to implement an algorithm which is able to perform phonetic segmentation of speech, without prior knowledge of the phoneme sequence contained in the signal. We use for that a set of pertinent features to get the better rate recognition normal in the disturbed conditions. The segmentation algorithm achieved above standard quality results compared to what is found in current literature regarding blind segmentation.
**Key words:** phonetic speech segmentation, Arabic language, blind approach

## INTRODUCTION

As far back as the apparition of written texts and pictures, audio signals, simple air vibrations, are the first communication vehicle between peoples. In view of the singular particularity of vocal signal that propelled it to the rank of others signals, it remains the object of a vast number of scientific researches during many decades.

Almost 250 millions speakers of the sixth language spoken all over the world, arabic people don't dispose, until the moment, of a sufficient number of automatic arabic recognition or synthetic systems, due to the lack of a big arabic developed database assuring the contentment of various treatment, training and test process needs and facilitating a host of application related uses. The segmentation of an arabic continuous speech database will be considered as an important step in the construction of a full database. This construction includes the automatic speech segmentation as an essential part.

Automatic speech segmentation requires great importance in various applications. Most common segmentation methods are based on Hidden Markov Models (HMM) using a statistic modeling of phonetic units to line up data with an already known transcription. HMMs have a high-performance talking about temporal alignment, but they suffer from some deficiencies as far as the discrimination is concerned because they don't allow exact phonetic knowledge consideration. More specifically, their use in segmentation process claims a big amount of required data for the training of the system, which leads to an expensive and very slow task. [AJM 04].

Text independent segmentation procedure, also called blind segmentation, is developed to surmount other process's insufficiency.

These methods spot transitions of signal representative parameters through their temporal evolution.

In this paper we propose a system that attempts to identify the phoneme boundaries without relying on the knowledge of the phone sequence, using a set of features including the mean power and Discrete Fourier Transform subtracted from the wave signal.

The rest of paper is organized as follows. Section 1 is devoted to a description of how the segmentation algorithm is implemented. The audio data, on which our system was tested, is presented in section 2. The 3th section includes results and interpretations in the light of existing literature.

This document provides style guidelines which must be respected by authors in order to ensure the uniform appearance of articles published in SETIT 2009 Conference. This document should be used as a model, particularly for the first page, the headers, sub-titles, headings, etc. Articles should be no longer than 12 pages.

## 1. Blind segmentation system: overview

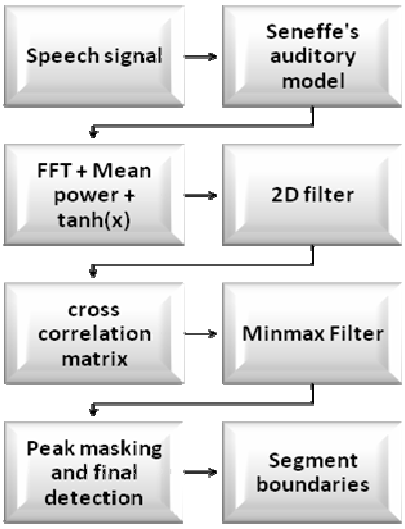


**Fig.1.** *Automatic segmentation system structure*

During its development, the blind segmentation algorithm into phone-sized units hasn't utilized any external information. The main approach principle consisted in spectrum waveform changes detection by comparing times cross correlations. That way, the algorithm automatically places segments boundaries where spectrum changes exceed a minimum threshold level.

We note that the input format signal of our segmentation system is the wave format able to remove overtaking or overflowing risks.

As the performance of speech recognizers often degrades considerably in noise, we have suggested the use of an auditory front end that imitates some aspects of human auditory-nerve and psychoacoustic behavior to decrease noise problem. The waveform signal is filtered by a Seneff's auditory model that tries to capture the essential features extracted by the cochlea in response to sound pressure waves.

This front end, described in the Fig.2, includes a first stage of 40 linear filters, followed by a series of nonlinearities modeling the transformation from basilar membrane movement to auditory nerve stimulus. Such nonlinearities incorporate soft half-wave rectification, a model for short-term adaptation, and a rapid AGC.

Seneff's auditory model had two outputs. "Synchrony" output detects the extent that the nonlinear stage output for a particular channel has energy at the center frequency for that channel. "Mean rate" output is generated by spotting the nonlinear stage output envelope. It approximately corresponds to spectral magnitude information and is rather efficient in producing GSD spectra with a restricted number of well defined spectral lines and also in following the dynamic modifications of speech. Transitions from one phonetic segment to the next are evidently outlined by onsets and offsets in the output representation better represented by the Envelope Detector auditory spectrogram. [SEN 86].

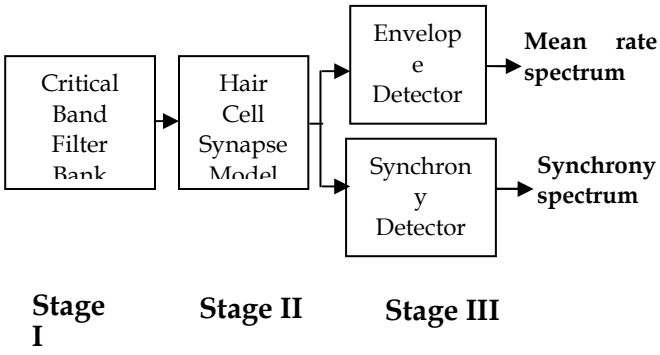


**Fig.2.** *Seneff's auditory model.*

After filtering, we applied a short hamming window (6 ms of length with 1 ms step size) to the amplified signal to get good spectral contrast by choosing window sizes in the order of one pitch-period in length and centering the energy of the glottal pulse in the middle of the window.

As many ways of storing statistic signal information exist, pertinent information is extracted from different properties of speech signal like formant trends, mean power, zero crossing rate, Discrete Fourier Transform and pitch. Using all these cues may perhaps result in some information loss which can be reduced through cautious extraction and mixture. For different necessities, different combinations of these parameters are able to create the wanted results. Therefore, we calculate the Discrete Fourier Transform using the fast Fourier transform algorithm and the mean power from each frame. To emphasize the formant information in the spectrum, every signal frame is also compressed in the frequency-domain with an asymmetric function defined as follow:

$$y(n) = \tanh(\alpha . x(n)) ; \quad \alpha = 0.5 \qquad (1)$$

All spectral frames are stored into a matrix A.

Next, from A, the cross correlation matrix is computed representing the inner coherence of the speech signal as a function of time. Since this step, we can detect some changes in the spectrum. In order to get information from this matrix, a 2 dimensional filter is applied to it resulting in a noisy vector called S showed in the Fig.3. This new representation of the speech signal depicts large negative peaks that indicate spectral variations and mark probable segment limit positions.

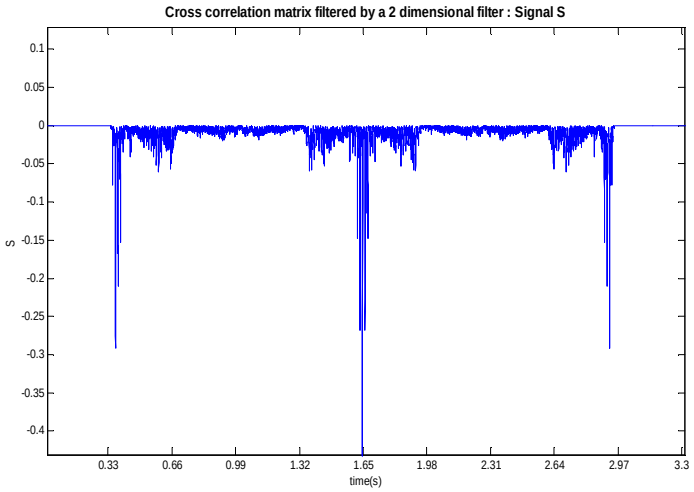


**Fig.3.** *Signal S produced by 2dimonsional filter applied to cross correlation matrix of an Arabic phrase "?hdaa lanaa Zaarunaa maa Saadahu fil Gaabi".*

A second special filter "minmax filter" is proposed so that we can remove noise from the signal S and refine its representation. While the filter goes beyond throughout the signal, at each point it takes n successive samples from S (note that we nominate minmax the number n of samples) and concludes the maximum vmax and minimum vmin values of this vector. The filter turns out the difference dmax=vmax-vmin as an output to the point where the minimum value was positioned. [RAS 07]. To understand the functioning of this filter, please look at Fig.4. applied to a speech frame (96 samples), the 40 samples length filter locates existing positive peaks. In our algorithm, the result of filtering is a signal so called S', in which the predictable segment boundary locations are symbolized as simply perceivable positive peaks. Peak heights values are normalized between 0 and 1 to offer a probability classification for each frontier, so, the higher corresponds to the peak and the larger is related to the local spectrum change and the more likely it is that a phone transition has happened. (Fig.5).

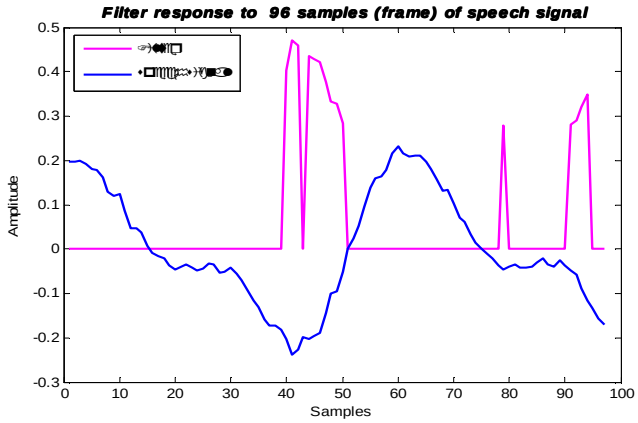


**Fig.4.** *Effect of the second filter on a speech frame (6 ms)*

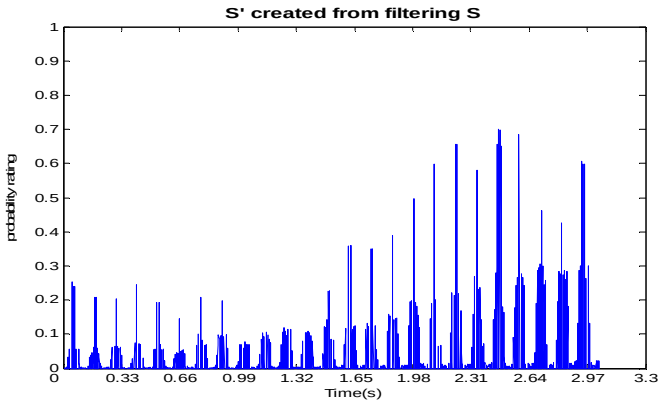


**Fig.5.** *Signal S' produced by filtering the signal S of an Arabic phrase "?hdaa lanaa Zaarunaa maa Saadahu fil Gaabi".*

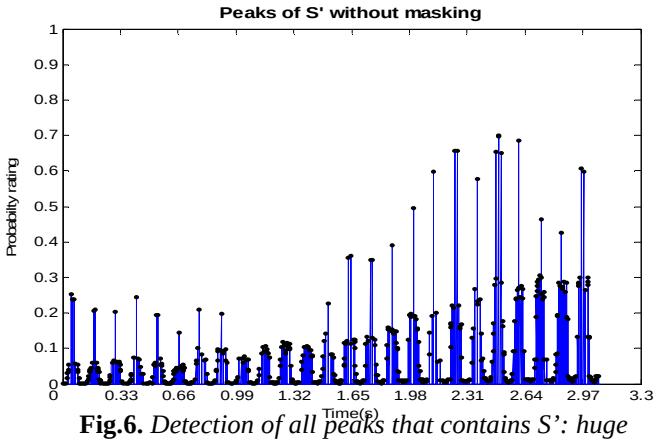


**Fig.6.** *Detection of all peaks that contains S': huge quantity of peaks requiring masking.*

Particularly in the case of long spectral transitions between two separated phones, many closer peaks may exist. (Fig.6). In order to improve the segmentation exactness and to stay away from over-segmentation, we compute the distance in the time domain between each peak crossing the threshold noted Pmin. If two or more peaks are closer than td to each other, we contrast their probability ratings and we only retain the highest one, while its position is slightly adjusted towards the removed peak(s). Then, its new location is between the old peaks being proportional to the ratio of probability ratings of previous peaks. This non linear method is called peak masking because it removes undesirable peaks. Masking distance td and threshold level Pmin are determined after many attempts and tests. The table 1 illustrated the optimum set of parameters used in this work that will be demonstrated later.

Theoretically, a boundary segments list, noted lf, will be created from the choice of all peaks higher than the threshold level, in other words, among previously selected peaks, only highest ones will be decided as phone limits. Final peak detection is guaranteed by a function that locates and measures signals maximums then returns the value, the position of each phonetic segment and the number of peaks as the final boundaries number. The automatic segmentation of an Arabic phrase is shown at the

Fig.7, where red lines refer to limits generated by the algorithm.

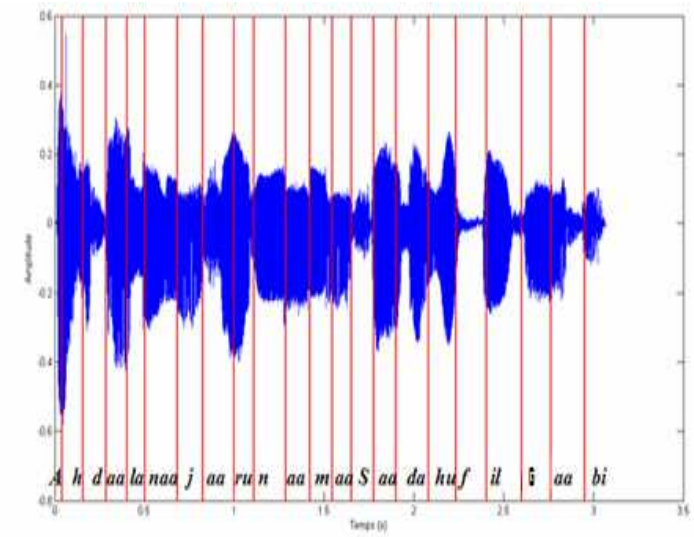


**Fig.7.** *Phonetic segment boundaries automatically generated from an Arabic phrase "?hdaa lanaa Zaarunaa maa Saadahu fil Gaabi".*

**Table.1.** *The best set of parameters used during the development of the segmentation algorithm.*

| *Parameter* | *Explication* | *Best value* |
|---|---|---|
| ***alpha*** | *compression coefficient of tanh(x)* | *0.5* |
| $p_{min}$ | *seuil de sélection des pics* | *0.07* |
| ***minmax*** | *length of filter applied at S* | *40 (samples)* |
| $t_d$ | *masking distance* | *0.065 (s)* |

## 2. Experimental protocols

### 2.1. Databases

The performance of the proposed methods is evaluated using three different datasets.

#### 2.1.1. *Arabic database*

Further to profitable researches regarding Arabic language and to evaluate the phonetic segmentation approach proposed in this article, we have employed a corpus already prepared within our laboratory. 4 words and 2 phrases had been recorded in favourable conditions, and then sampled at a frequency of 16 KHz. Each speaker from 22 (14 men, 5 women and 3 children) had pronounced 10 times continuously all the words and phrases in order to eliminate silences and interferences (hesitation, respiration…), so we have isolated between these words and phrases. The evaluation of automatic segmentation needs a robust reference, for that, we have segmented manually the hole of our corpus to obtain a complete database shown at the next table.

**Table.2.** *Arabic database*

| **Arabic evaluated Corpus** | **Arabic reference Corpus (total phoneme number)** |
|---|---|
| Phrase 1 : *?hdaa lanaa Zaarunaa maa Saadahu fil Gaabi* (22*10) | 22 * 22*10 |
| Phrase 2 : *daH Hanka lawmii fainna allawma iGraa?u* (22*10) | 23*22*10 |
| Mot 1 : *jamiin* (22*10) | 4*22*10 |
| Mot 2 : *jasaar* (22*10) | 4*22*10 |
| Mot 3 : *amaam* (22*10) | 4*22*10 |
| Mot 4 : *xalf* (22*10) | 3*22*10 |

#### 2.1.2. *Noisy database*

The segmentation system robustness is defined by a low error rate even if we are in bad conditions (presence of noise). To be sure of the accuracy of our algorithm, we might evaluate it in a noisy environment. We have then conducted an addition of some sound effects at our Arabic corpus. Two types of noise have been proposed, a white noise (in which the energy of the random signal is uniformly distributed) and a pink noise (a limited spectrum white noise) at different level of SNR = {0 dB, 5 dB, 10 dB and 15 dB}.

#### 2.1.3. *TIMIT database*

Among databases often used for speech applications, certainly, TIMIT is the most widespread. The Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT contains speech from 630 speakers representing 8 major dialect divisions of American English, each speaking 10 phonetically-rich sentences. We have just evaluated the test data that has a core portion containing 24 speakers, 2 male and 1 female from each dialect region. Each speaker read a different set of sentences. Thus the core test material contains 192 sentences for each speaker, each having a distinct text prompt.

### 2.2. Evaluation protocol

The evaluation method of the segmentation algorithm follows a literature convention of manual annotation comparison, in which segmentation boundary locations are compared against the manual annotated boundaries.

First, we have used following abbreviations for all the rest of the paper:

- *Lf*: automatic limit generated by the segmentation algorithm;

➢ *Lref*: manual limit deduced from the reference corpus.

The evaluation protocol consists on taking a limit *Lref,j* from a list of boundaries *Lref* supplied by the reference annotation for the corresponding signal, and then computing distances to each of the boundaries in *Lf* provided by the segmentation algorithm. The closest frontier *Lf,i* found is selected, and its distance to the boundary *Lref,j* is checked. If the distance in time is less than 20 ms, the boundary is considered as matched and *Lf,i* is removed from the *Lf* to avoid several equivalents. If there is no matching *Lf,i* limit for *Lref,j* presented within ±20 ms, it will be considered as a miss.

Every following boundary *Lref,j* in the annotation is tested for a probable match. Finally, we calculate the result as an arithmetical mean of all results from the evaluated signals. (Fig.8) [KVA 93].

Strictly speaking, there is no objective standard for evaluating strictly the errors in the different segmentation methods, because segmentation is very subjective. However, by comparing the automatic segmentation results with the manual segmentation, we can have some segmentation criteria.

Let's denote with *Nhit* the total number of matches limits, the number of boundaries produced by the algorithm are denoted with *Nf* and the number of reference limits are denoted with *Nref*.

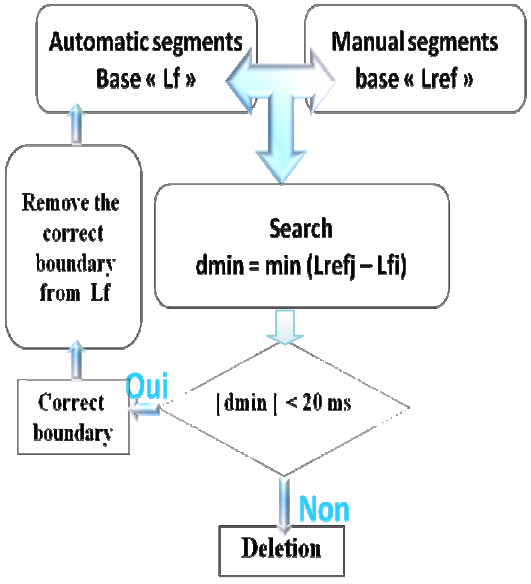


**Fig.8.** *Synoptic of the evaluation method.*

The over-segmentation coefficient and the hit rate are defined as follows ([PET 96]; [AVE 01]):

* The over-segmentation coefficient D' is:

D' = ((Nf/Nref) – 1) * 100 (2)

* The correct detection rate Pc is:

Pc = (Nhit/Nref) * 100 (3)

In lots of domains of research, there are two other criterions for evaluation: precision (4) and recall (5). In the case of speech segmentation, they may be defined as follows ([AJM 04]):

* Precision:

PRC=Nhit/Nf (4)

*Recall:

RCL = Nhit/Nref (5)

From these two coefficients we can estimate a single scalar (0-1) that can be used to approximate overall segmentation quality, the F-measure:

F= (2* PRC * RCL)/(PRC+RCL) (6)

## 3. Results and Discussion

In order to illustrate how the above blind speech segmentation algorithm works, it was at first tested on two clean speech databases; Arabic corpus and Test section of Timit. Then, we evaluate its performance in noisy conditions. The aim was to obtain a fine understanding of the general performance of the algorithm that could be compared to the other results reported in blind segmentation literature, and to determine the general effects of different parameters to the segmentation results.

### 3.1. Segmentation of the Arabic material

The first results, presented in table 3, contain the evaluation of a basic segmentation of the Arabic database with optimum parameters.

By accepting higher values of over-segmentation, higher correct detection rates can be obtained.

For instance, with the two phrase we have reported respectively a correct detection rate Pc = 83.87 % with an over-segmentation value D'= 60.53 % and Pc = 78.43 % with a negligible over-segmentation equals to 1.58%. With a female speaker, we have obtained a correct detection rate Pc = 88.36 (D' = 70.72 %) and Pc = 78 .82 %. (D' = 1.87 %). The efficiency of all system is recapitulated on the F-measure = 0.778.

These results obtained are well in line with the other results reported in literature concerning blind segmentation algorithms. For example, [AVE 01] reported a correct detection rate of Pc = 73.58 % with an over-segmentation value D' = 0 %, or by allowing excess over-segmentation, they gained Pc = 90 % and D' = 63 %. lately, [EST 07] have reported a correct detection rate of Pc = 76.0 % with non-existent over-segmentation D' = 0 % while at Pc = 90.3 %, D' = 75.5 %. Three years ago, [ZIB 05] have obtained an F-measure equals to 0.729.

**Table.3.** *Segmentation results for the Arabic corpus with different amounts of over-segmentation.*

| Data | Pc | PRC | RCL | F | D′ |
|---|---|---|---|---|---|
| **Phrase1** | 83.87 | 0.56 | 0.83 | 0.67 | 60.53 |
| **Phrase2** | 78.43 | 0.73 | 0.78 | 0.75 | 1.58 |
| **Word1** | 92.16 | 0.56 | 0.92 | 0.7 | 63.63 |
| **Word2** | 90.22 | 0.82 | 0.9 | 0.86 | 11.69 |
| **Word3** | 92.76 | 0.77 | 0.93 | 0.84 | 22.67 |
| **Word4** | 89.32 | 0.81 | 0.89 | 0.85 | 14.87 |
| **TOTAL** | Pc = 87.8 | | F = 0.778 | | |

### 3.2. Segmentation of the English material (TIMIT)

The results of the test section segmentation are shown in table 4. These results are very satisfactory comparing with the results reported by okko Rasannen in 2007, [RAS 07], when segmenting the same material; Pc = 83.79% with D'= 42.10%.

The Fig.9 shows the automatic (red lines) and the manual (black lines) segmentations of the phrase « She had your dark suit in greasy wash water all year » pronounced by a « New England » dialect female speaker.

**Table 4.** *Evaluation of the section test segmentation with different over-segmentation values.*

| Data | Pc | D′ | PRC | RCL | F |
|---|---|---|---|---|---|
| **Section test of TIMT corpus** | 78.91 | 1.06 | 0.78 | 0.79 | 0.78 |
| | 83.17 | 38.96 | 0.62 | 0.84 | 0.71 |

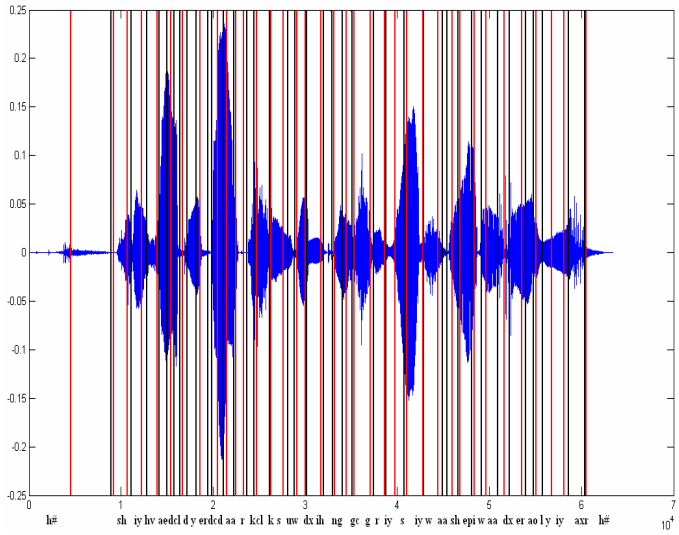


**Fig.9.** *Segmentation of the phrase « She had your dark suit in greasy wash water all year » (D′=19%).*

We have noticed that the segmentation algorithm gives better results with Arabic speech than with TIMIT material. This is an expected result even if general phonetic differences between English and Arabic are ignored, since the Arabic material is spoken more slowly and each syllable is being stressed with greater care.

### 3.3. Segmentation of the noisy database

The noise robustness of the algorithm was tested by adding white and pink Gaussian noises to the Arabic speech signals before evaluating the segmentation quality. The correct detection rate first started to decrease as noises were added to the original speech. Also, at very poor SNR-levels the hit rate *Pc* seemed to converge asymptotically to a specific value, as did the over-segmentation *D'*.

We have reported accepted results mainly in the case of pink noise, the F-measure that reflect the real performance of the algorithm reaches a value of 0.73 with SNR = 15db.

Results shown in the following table indicate that our segmentation system is robust and gives good outcomes.

**Table 5.** *Evaluation of the segmentation of the Arabic noisy database.*

| Type de bruit | SNR (dB) | Pc | PRC | RCL | F |
|---|---|---|---|---|---|
| **Blanc** | 0 | 61.72 | 0.51 | 0.62 | 0.55 |
| | 5 | 66.67 | 0.6 | 0.67 | 0.63 |
| | 10 | 71.86 | 0.65 | 0.72 | 0.68 |
| | 15 | 74 | 0.69 | 0.74 | 0.72 |
| **Rose** | 0 | 63.08 | 0.58 | 0.63 | 0.61 |
| | 5 | 70.14 | 0.63 | 0.7 | 0.66 |
| | 10 | 73.82 | 0.69 | 0.74 | 0.7 |
| | 15 | 75 | 0.72 | 0.75 | 0.73 |

For SNR = 0db, we have reported a hit rate Pc equals to 61.72 % and an F-measure F = 0.55. For a high value of SNR (15 db), Pc = 74 % and F = 0.72 in the case of white noise.

### 3.4. Parameter dependency

Many factors in the segmentation algorithm can be adjusted to turn out different results for speech segmentation. We have noticed during the evaluation of this algorithm that when accepting high values of over-segmentation, we can obtain higher values of hit rate. The most efficient method to do this is to regulate the length of the Minmax filter, the probability threshold *pmin* of the peak detector or the masking distance *td*. (Fig.10).

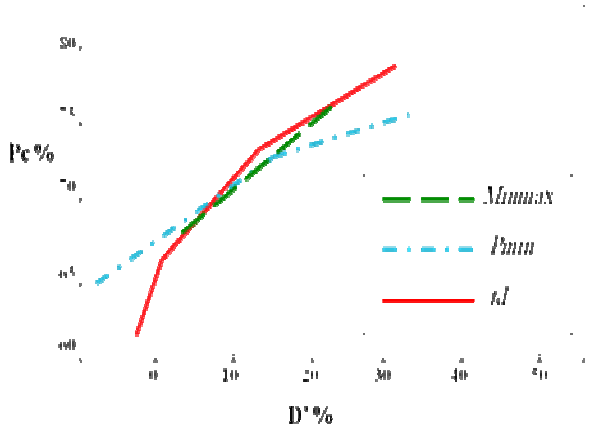


**Fig.10.** *Illustration of the relation between the hit rate Pc and the over-segmentation D', and the influence of some parameters on the performance of the algorithm.*

The relation linking the hit rate Pc and the over-segmentation D' may be due to the fact that the augmentation of D' gives a biggest probability of coincidence between manual boundaries and automatic ones that will be accordingly more numerous that the speech is segmented. Fig.11 shows that the over-segmentation parameter is inversely proportional to the masking distance *td*.

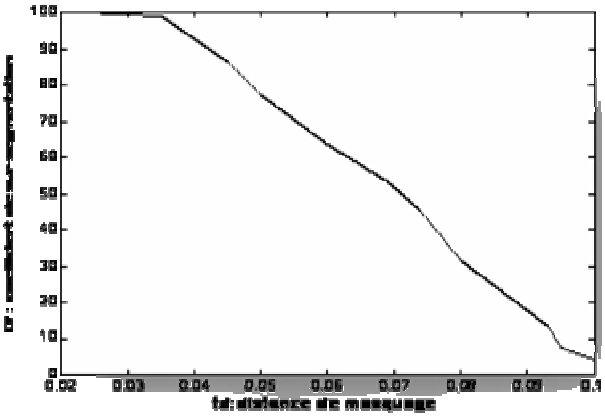

**Fig.11.** *The reduction of the masking distance td leads to the augmentation of the over-segmentation D'.*

In any case, this observation leads to a conclusion that the segmentation results that are reported with relatively high over-segmentation values in the literature say very little about the true nature of the algorithm. Our purpose is to maximize the correct detection rate without letting the over-segmentation raise. So, we have investigated the F-measure that gives more true information about the performance of the segmentation system.

We have noticed that even if for high values of RCL, the F-measure is not always satisfactory despite the fact that RCL is the same as the correct detection rate (RCL=Pc/100). The masking distance td, being responsible of the augmentation of the over-segmentation value and then the correct detection rate Pc, has an important effect on the variation of the RCL value. In order to deduct the optimum value of td giving a high value of F-measure, we have looked for the relation connecting td, RCL and PRC. So, we have adjusted the masking distance value and observed the evolution of last mentioned coefficients for the first phrase in our Arabic database. F-measure is increasingly high if PRC and RCL have high close values. The Fig.12 shows that the most appropriate value of masking distance is about *td* = 0.065s.

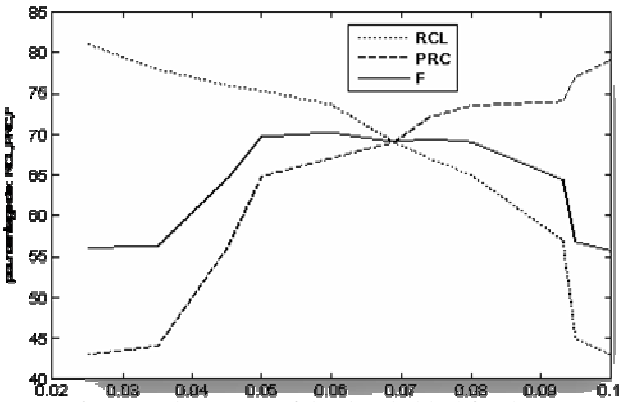


**Fig.12.** *Adjustment of the threshold td for the optimization of the PRC and RCL values and the maximization of the F-measure.*

## 4. Conclusion

For the purpose of reducing the manual intervention in a segmentation task, a new process for speech segmentation is accessible in this work. The method is based on a new preprocessing approach that investigates the amelioration brought by an auditory model able to extract robust acoustic features which capture information about phoneme transitions. Besides, it utilizes cross-correlation estimates of short window length FFT-spectrums in order to detect potential phonetic boundaries. The interest of the text independent segmentation algorithm is its accuracy and ability to realize phonetic cutting of any language and at any conditions. Then, we have improved our Arabic continuous speech database to can be used as a front-end for any speech application which requires segmentation.

The results attained by the present version of the algorithm are analogous to those found in current literature ([AVE 01], [EST 07], [RAS 07]), The evaluation method used in this paper is the strictest possible version of the search region evaluation, and the reached results with it are very close to other greatest blind algorithms. However, we should remain in mind that the reference comparison is not the final objective of speech segmentation that depends on the complete speech processing system, in which it is implemented, and the most significant assessment method would be in that case to examine the functionality of the whole system.

We have to notice that the long-term goal of this work is to develop an Arabic learning speech recognizer that can be used to develop language skills of infants.

## REFERENCES

[AJM 04] J. AJMERA, I. MCCOWAN, H. BOURLARD, « Robust Speaker Change Detection », *IEEE Signal Processing Letters*, Vol. 11, No. 8, 2004.

[AVE 01] G. AVERSANO, A. ESPOSITO, A. ESPOSITO, M. MARINARO, « A New Text-Independent Method for Phoneme Segmentation », *Proceedings of the IEEE international Workshop on Circuits and Systems*, Vol. 2, pp. 516-519, 2001.

[EST 07] Y.P. ESTEVAN, V. WAN, O. SCHARENBORG, « Finding Maximum Margin Segments in Speech », *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[KVA 93] K. KVALE, « Segmentation and Labeling of Speech », PhD Dissertation, The Norwegian Institute of Technology, 1993.

[PET 96] B. PETEK, O. ANDERSEN, P. DALSGAARD, « On the robust automatic segmentation of spontaneous speech », *in proceedings of ICSLP*, 1996.

[RAS 07] O. RASANNEN, «Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture », helsinki university of technology, 2007.

[SEN 86] S. SENEFF, « A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research », *Proceedings oflCASSP-86*, April 1986.

[ZIB 05] J. ZIBERT, F. MIHELIC, « Development, Evaluation and Automatic Segmentation of Slovenian Broadcast News Speech Database », Faculty of Electrical Engineering University of Ljubljana, 2005.