Paper

Unsupervised Phoneme Segmentation Based on Main Energy Change for Arabic Speech

Noureddine Lachachi

Faculty of Exact and Applied Sciences, University of Oran 1 Ahmed Ben Bella, Oran, Algeria

Abstract—In this paper, a new method for segmenting speech at the phoneme level is presented. For this purpose, author uses the short-time Fourier transform of the speech signal. The goal is to identify the locations of main energy changes in frequency over time, which can be described as phoneme boundaries. A frequency range analysis and search for energy changes in individual area is applied to obtain further precision to identify speech segments that carry out vowel and consonant segment confined in small number of narrow spectral areas. This method merely utilizes the power spectrum of the signal for segmentation. There is no need for any adaptation of the parameters or training for different speakers in advance. In addition, no transcript information, neither any prior linguistic knowledge about the phonemes is needed, or voiced/unvoiced decision making is required. Segmentation results with proposed method have been compared with a manual segmentation, and compared with three same kinds of segmentation methods. These results show that 81% of the boundaries are successfully identified. This research aims to improve the acoustic parameters for all the processing systems of the Arab speech.

Keywords—band frequencies, energy changes, formant analysis, phoneme segmentation.

1. Introduction

Phonetic segmentation is the action of dividing the speech signal into its basic language functional units: the phonemes. The accurate segmentation and labeling of speech into phoneme units is useful for diverse purposes, as for example the initialization of speech recognizers, the creation of databases for concatenated text-to-speech systems, the evaluation of the performance of speech recognition tasks, and the health related assessment of speech. In this last point, there are special topics in cognitive communication information that require the segmentation of speech signal into phoneme sized units in the processing of continuous speech. There are many types of applications, where the precise knowledge of phoneme is not important, just the type of the given sound, like vowel, nasal, voiced/unvoiced fricative, stop, etc. In these applications, the linguistic content is not important, just the acoustic characteristics are needed. This kind of segmentation is necessary, when the desired behavior depends on speech timing, like rhythm or the place of voiced sounds. Moreover, such segmentation technique is useful for the visualization of the acoustical parameters of speech in an audio-visual pronunciation training system [1]–[3].

In these issues, automatic alignment tools have been developed (e.g. EasyAlign [4], SPPAS [5]). They offer a consistent and reproducible alignment at reduced cost. The task they perform is known as "linguistically constrained segmentation" or "forced alignment". In these systems, only the time boundaries of the phonemes have to be determined. For this purpose, acoustic modeling based on Hidden Markov Models (HMMs), relying on speech segmentation techniques, has been shown to achieve the best results [6].

As described for example in [7], freely spoken language consists of sequences of various phonemes. Such phonemes can be classified into both voiced and unvoiced sounds. Depending on the manner how these sounds are produced, two different cases can be distinguished. First, voiced sounds such as normal vowels are characterized by a set of several characteristic frequencies that are called formants of the respective phoneme. Second, unvoiced phonemes also show characteristic formants. However, due to the fact that these sounds do not dominantly come from an associated vibration of the vocal folds (rather turbulent and irregular air flows are involved in the corresponding sound production), these phonemes are characterized by broader frequency ranges [8].

Analysis and presentation of the speech signal in the frequency domain are of a great importance in studying the nature of speech signal and its acoustic properties. The prominent part of speech signal spectrum belongs to formants that correspond to the vocal tract resonant frequencies. These are usually referred to as $F1$ indicating the first formant, $F2$ indicating the second formant, $F3$ indicating the third formant, etc. The quality of some of the most important systems for speech recognition and speech identification as well as those for formant based speech synthesis are dependent on how accurate the formant frequencies are determined. The formant defines the range of frequencies that is used for detecting the delimitations of the phonemes in a speech signals. Hence it conduct to the task of segmentation. There are many research works on automatic speech segmentation to classify speech into phonetic classes, but in Arabic language, the segmentation has not been well studied. Therefore, this paper proposes an effective segmentation, suitable for Arabic automatic speech recognition and related applications.

The purpose of this document is to identify segments of phonemes on a frequency range limited to a narrow spectral areas. Presented study is more relevant on the spectral distribution of voice signals where six areas are used.

Using formant analysis of Arabic language, we attempt to detect vowels and consonants that are spoken. Here a standard approach for detect the phonemes in continuous speech is described based on three frequency formants: $F1$, $F2$ and $F3$ to define the range of area frequency. We have investigated the correlations between formants in each phoneme and developed an algorithm to segment speech based on the overlap different vowels in $F1-F2$ and $F2-F3$ planes.

The results, have been compared with a manual segmentation in order to calculate the accuracy that shows the performance, and have been compared with three same kinds of segmentation methods.

2. Supervised and Unsupervised Speech Segmentation

Automatic speech segmentation is the partitioning of a continuous speech signal into discrete, non-overlapping units. Generally, automatic speech segmentation methods are divided in two types.

2.1. Supervised Speech Segmentation

This methods require training on speech material and a priori knowledge [9], [10]. The segmentation algorithm relies on the linguistic knowledge associated with the input speech signal, such as its phonetic transcription or the knowledge of its phoneme sequence as well as by the number of phonemes present. This means that the representation of the utterance in terms of discrete units is known, and pretrained acoustic models of these units are needed for the forced alignment. Thus, the system is only required to locate optimally the boundary locations that best coincide with the phoneme sequence given. The task of the segmentation algorithm is then to locate optimally the phonemes boundaries [11].

2.2. Unsupervised Speech Segmentation

These methods do not require training data to segment speech signal [12], it uses a set of rules derived from the decoding of human knowledge issued of the nature of the floor to make the operation of segmentation. Indeed, the segmentation algorithms are designed without any prior linguistic knowledge about the phoneme sequence of the input speech signal. The system blindly determines the best estimate of the number of phonemes along with their boundary locations, based on the acoustic cues extracted from the speech signal.

Acoustic (rate of) change (see [13] for early work on unsupervised automatic speech segmentation and below for more recent work) is an example of prior human knowledge that is used to solve the speech segmentation task. The task for an unsupervised segmentation algorithm is based in two point. The number of segments in the speech signal needs to be determined and the position of the boundaries determined on the basic characteristics of the acoustic signal. The unsupervised methods yield a desirable and more flexible framework for the automatic segmentation of speech and their algorithms are generally simpler than used in supervised methods [14].

2.3. Unsupervised Speech Segmentation Application

Some applications of the unsupervised speech segmentation include:

- Speaker verification systems. To achieve a phoneme level segmentation (without orthographic information) of a user selectable password in a text-dependent speaker verification systems.
- Speech recognition systems. To obtain phoneme level segmentation (level modeling phoneme) in a low-to-medium size vocabulary speech recognition systems, with user-defined vocabulary (such as, in voice dialing applications).
- Language identification systems. To find a phoneme level segmentation for multilingual un-transcribed corpus applied to automatic language identification.
- Speech corpus segmentation and labeling. To obtain a great level of phoneme segmentation of a speech corpus. This can be used as seed values to aid the subsequent manual process of phonetic transcription.

3. Modern Standard Arabic

The Arabic language has a standard pronunciation, which basically is the one used to recite the Quran. The same pronunciation is used in newscasts, discourses and formal actuations of all types [15]. Spoken in the Middle East and North Africa, Arabic has different dialects where some letters are pronounced in different manner [16], [17]. However, the literary Arabic also called Modern Standard Arabic (MSA) or Al-fus-ha. One of the differences between the spoken and written Arabic is the presence of diacritics marks (spoken segments that not present in the written form). The complexity of this language is due to the unusual morphology: words are formed using a “root and pattern” scheme, where the root is composed of 3 consonants, leading to several possibilities using one root.

3.1. Arabic Phonology

Phonetically, MSA has 34 basic phonemes of which six are vowels (short vowels **/i/**, **/u/**, and **/a/** and long vowels

/i:/, **/u:/**, and **/a:/**), and 28 are consonants. The Arabic alphabet only consists of letters for long vowels and consonants. Other pronunciation phenomena, including short vowels (*harakat*), nunation (*tanwin*) and consonant doubling (*shadda*), are not typically written. However, they can be explicitly indicated using diacritics. Vowel diacritics represent the three short vowels: **a**, **i**, and **u** (*fatha*, *kasra* and *damma*) or the absence of a vowel (*sukun*).

Additionally, pharyngeal and emphatic phonemes comprise two distinctive classes of phonemes in Arabic. These two classes are found only in Semitic. These phonemes can be grouped according to the articulation of the lips and tongue during speech [18].

3.2. Arabic Syllables

The syllable types allowed in MSA are **CV**, **CVC**, and **CCV**, where **V** indicates a (long or short) vowel and **C** indicates a consonant. Arabic sentences must start with a consonant [19], and all Arabic syllables must contain at least one vowel. In addition, while such vowels cannot occur in word initial position, they can occur between two consonants or in word-final position. This is in contrast with other major languages, i.e. English, Japanese. In Japanese language, vowel can occur at any position of a word and most of the Japanese words end with vowel like pronunciation. Arabic syllables can be classified as short or long. The **CV** syllable type is a short syllable while all others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant. For Arabic, a vowel always forms a syllable nucleus, and there are as many syllables in a word, as there are vowels in it [20].

3.3. Formant Analysis in Arabic Speech

It has been noted that generally most of the energy of vowel lies below 2 kHz and in case of voiced consonants lies below 3 kHz as shown in Fig. 1 [21]. Vowels are lower-frequency components of speech and create the sound volume of speech.

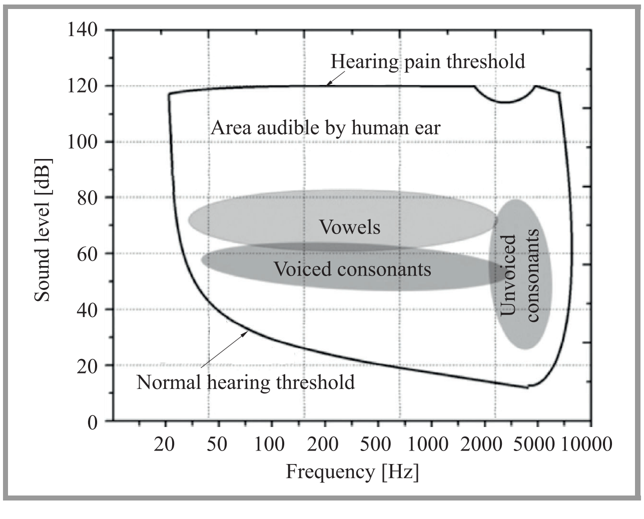

Fig. 1. Normal hearing frequency distribution of human speech.

Vowels are among the essential components of any spoken language. The analyze and the study of vowels in Arabic is very important designing reliable and robust speech processing systems due to the fact that almost 60 to 70% of Arabic speech is vowels [22].

Table 1
The relationship between the vocal tract characteristic and the two formants $F1$, $F2$

Vocal tract characteristic	$F1$	$F2$
Length of the pharyngeal oral tract	Inversely proportional	Inversely proportional
Oral constriction in the front half of the vocal tract	Inversely proportional	No effect
Pharyngeal constriction	Proportional	No effect
Back tongue constriction	No effect	Inversely proportional
Front tongue constriction	No effect	Proportional
Lip rounding	Inversely proportional	Inversely proportional

Table 1 give $F1$ and $F2$ give indication about the constrictions of the vocal tract in generating vowels [23].

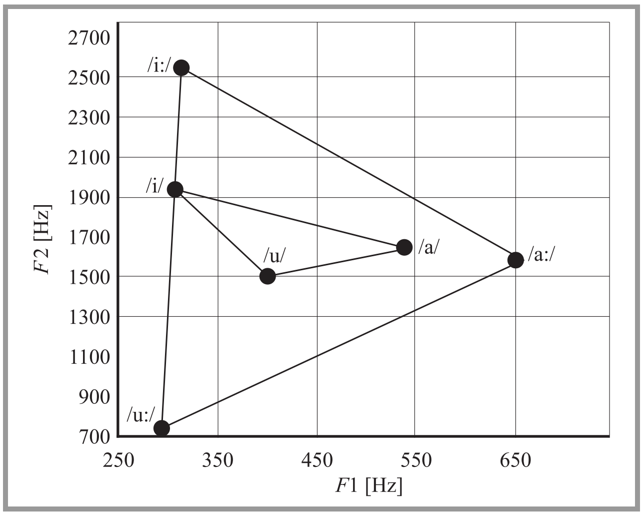

Fig. 2. The formant triangle of MSA vowel [15].

Based on the study [24], a formant-based analysis for the six vowels of MSA language was carried out and the values of the first three-formant frequencies were captured. Their results were compared to some previously published ones conducted on MSA and other Arabic dialects. The comparison was performed from geometric perspective using the Euclidean distance. The comparison results were found to be consistent with the visual inspection of the vowel triangles as shown in Fig. 2.

In Fig. 2, one can see that the vowels **/i:/** and **/i/** have low frequencies in $F1$ and high frequencies in $F2$. Moreover, the frequencies $F1$ and $F2$ are both low for the vowels **/u:/** and **/u/**. In the case of **/a:/** and **/a/**, both have $F1$ with high frequency and $F2$ with an average frequency. Therefore, when these vowels are plotted $F1$ to $F2$, they form two triangles.

The results of the analysis of the first three formants are summarized in Table 2 where speakers uttered perfectly Arabic phonemes without any influence by their local dialects [24].

Table 2
Results of $F1$, $F2$ and $F3$

Vowels	$F1$ [Hz]	$F2$ [Hz]	$F3$ [Hz]
/a:/	651.5	1588.1	3058.3
/i:/	314.1	2549.8	3278.9
/u:/	295.4	744.3	2560.2
/a/	535.0	1635.0	5890.6
/i/	307.5	1942.1	2702.7
/u/	407.9	1520.3	2777.7

4. Methodology

This section outlines in detail the settings of band frequencies, algorithm and computation conducted with references to the research presented in this paper. The proposed strategy based spectral analysis extracts phonemes from the raw speech waveforms. It requires no learning and it is language independent applied for Arabic speech.

4.1. Band Frequencies Definition

It is well known that an acoustic speech signal contains information beyond its linguistic content. This paralinguistic information includes clues to a speaker's accent and identity, which are exploited by automatic accent identification (AID) and speaker identification (SID) systems. The relationship between AID and SID is asymmetric, since accent information is relevant to SID but speaker information is a distraction in the context of AID.

For instance, the speaker identification study in [25], performed on the clean TIMIT corpus using mono Gaussian modeling, showed that the frequency regions below 600 Hz and above 3000 Hz provided better SID than the middle frequency regions. However, no similar study has been conducted for AID. In [26], the contrasting importance of different frequency bands for AID and SID are investigated, using contemporary GMM-based systems. These bands are defined in center frequency shown in Table 3 [26].

According to [26], it is useful to divide the spectrum into four areas: A (0 to 0.77 kHz), B (0.34 to 3.44 kHz), C (2.23 to 5.25 kHz) and D (3.40 to 11.02 kHz). The results suggest that speaker information dominates in areas A and D. The first area A, corresponding to primary vocal tract resonance information, and the second area D, corresponding to high-frequency sounds. These results are consistent with [25]. In contrast, area B is most useful for AID, indicating that the vocal tract resonance information in this region is linguistic biased, rather than speaker information. Area C contains both types of information, although speaker information appears dominant.

Table 3
The center frequency for 31 Mel-spaced band-pass filters [26]

Filter number	Center frequency [Hz]	Filter number	Center frequency [Hz]
1	129	17	2239
2	258	18	2497
3	344	19	2799
4	473	20	3100
5	559	21	3445
6	645	22	3832
7	775	23	4263
8	861	24	4737
9	990	25	5254
10	1076	26	5857
11	1205	27	6503
12	1335	28	7235
13	1464	29	8253
14	1636	30	8957
15	1808	31	9948
16	2024		

Based on the assumption that the majority of phonemes used in Spanish language are used in Arabic language, we consider the study given in [27].

In [27], it is shown that for Portuguese language (or Spanish language) there are 48 different phonemes used for the SID grouped into 11 classes. These classes are: silence, voiced fricative and unvoiced fricative, voiced plosive and unvoiced plosive, nasal consonants, nasal vowels, front vowels, median vowels, back vowels and liquid consonants.

For each class, a given set of representative parameters is largely used for phoneme classification. The parameters used for each class are as follows [28].

Silence – only the total energy of the analysis window is used (threshold –35.8 dB). The boundary between the silence and other classes is set up at the frame where the total energy becomes greater than the threshold.

Vowels (median, front, back, nasal) – four parameters are used: total energy of the analysis window, first ($F1$) and second ($F2$) formant values and energy profile. The transition between vowels and the other classes is determined by using the total energy of the analysis windows (transition is set where the energy is below –28 dB). Energy profile, $F1$ and $F2$ values are used to separate vowels in diphthongs. $F1$ is used to separate median vowels from back and front vowels, and the boundary is set up at the frame where $F1$ is below 673 Hz. Energy profile and $F2$ value are used to separate front vowels from back vowels. The transition is determined at the frame where $F2$ is below 1845 Hz and the energy profile is below 2106 Hz. Energy profile represents the frequency band carrying a given percentage of the total energy and is calculated from the Discrete Fourier Transform (DFT) of the windowed speech signal.

Fricative (voiced and unvoiced) – two parameters are used: Zero Crossing Rate [29] (thresholds 0.35 for voiced fricatives and 0.62 for unvoiced fricatives) and gravity spectral center (threshold 2500 Hz). The gravity spectral center represents the frequencies where 50% of the total energy of the windowed signal is concentrated. The transition from fricatives to other classes is determined at the frame where the parameters values are below the thresholds.

Plosive (voiced and unvoiced) – three parameters are employed: energy in the frequency bands $[0-F3]$ and $[F3-f_s/2]$ [30] and the first order derivative of $F2$, where $F2$ and $F3$ represent the second and third formant frequencies and f_s is the sampling frequency. As the derivative of $F2$ exhibits a peak at the transition from plosive to other classes where the peak position represents the boundary. Energy is combined with the derivative permit to avoid spurious peaks. The energy in the frequency band $[0-F3]$ for voiced plosive is above of –5 dB and in the frequency band $[F3-f_s/2]$ is above of –2 dB. For unvoiced plosive the energy is above of 5 dB and 10 dB for the bands $[0-F3]$ and $[F3-f_s/2]$ respectively.

Nasal consonants – two parameters are used: $F1$ value (threshold 280 Hz) and the ratio between the spectral energy in the frequency bands $[0-353]$ Hz and $[358-5373]$ Hz (threshold 0.87). When the $F1$ value is greater than 280 Hz and the spectral energy ratio is below 0.87, a transition has occurred from nasal consonant to another class.

Liquids – two parameters are employed: spectral energy band $[0-2600]$ Hz (threshold above 6.5 dB) and its first order derivative. Transition from liquid to another class tends to exhibit a peak in the first derivative of the spectral energy. The peak determines the transition and at this frame, the spectral energy threshold must be below 6.5 dB.

Based the study in [26], [27] and the one of formant frequencies defined in Arabic speech developed in Section 3, one can see that if we divide all frequencies centers [26] indicated in Table 3 into six zones (Table 4), we get closer to the syntheses given above [28].

To investigate the effect of different frequencies areas, segmentation experiments were conducted using frequency band limited speech data comprising the outputs of adjacent filters regions. For example in **LF2** area, we considered $k=4$ overlapping sub-bands, where the N-th sub-band comprises the outputs of filters N to $+3$ $(N=1\dots4)$.

Table 4
Definition of the six region band frequencies

LF – low frequency			
Band LF1	Center frequency [Hz]	Band LF2	Center frequency [Hz]
1	129	1	559
2	258	2	645
3	344	3	775
4	473	4	861
5	559	5	960
		6	1071
MF – medium frequency			
Band MF1	Center frequency [Hz]	Band MF2	Center frequency [Hz]
1	1076	1	1808
2	1205	2	2024
3	1335	3	2239
4	1464	4	2457
5	1636	5	2799
6	1808	6	3100
HF – high frequency			
Band HF1	Center frequency [Hz]	Band HF2	Center frequency [Hz]
1	3100	1	5254
2	3445	2	5854
3	3832	3	6503
4	4263	4	7235
5	4737	5	8000
6	5254		

4.2. Energy Computation over a Frequency Band

For the human ear perceiving speech along a nonlinear scale in the frequency domain [31], one approach is to use a uniformly space-warped frequency scale, such as the Mel scale.

The relation between Mel-scale frequency and frequency (Hz) is described by the following equation:

$$Mel = 2595\log(1+f/700)\,, \tag{1}$$

where Mel is the Mel-frequency scale and f is in Hz. The filter bank is then designed according to the Mel scale. For example, we take 4 frequency bands in LF2 area (see Table 4) that are approximated by simulating 4 triangular band-pass filters, (i,k) $(1 \le i \le 4, 11 \le k \le 21)$. Over a frequency range of 559 – 1076 Hz, we consider that the speech signal is sampled at 16 kHz windowed over 10 ms (each window of 10 ms has 160 point), and the spacing as well as the bandwidth are determined by a constant Mel frequency interval by Eq. 1. Considering a given time-domain noisy speech signal, $x_{time}(m,n)$, representing the magnitude of the n-th point of the m-th frame, we first find the spectrum, $x_{freq}(m,k)$, of this signal by 160-point DFT:

$$x_{freq}(m,k) = \sum_{n=0}^{N-1} x_{time}(m,n) W_N^{kn}, \quad (2)$$

$$0 \leq k \leq N-1,\ 0 \leq m \leq M-1,$$

$$W_N = \mathrm{e}^{\frac{-j2\pi}{N}}, \quad (3)$$

where $x_{freq}(m,k)$ is the magnitude of the k-th point of the spectrum of the m-th frame, N is 160 and M is the number of frames of the speech signal for analysis. Then, we multiply the spectrum $x_{freq}(m,k)$ by the weighting factors $f(i,k)$ on the Mel-scale frequency bank and sum the products for all k to get the energy $x(m,i)$ of each frequency band i of the m-th frame:

$$x(m,i) = \sum_{k=0}^{N-1} \left|x_{freq}(m,k)\right| f(i,k), \quad (4)$$

$$0 \leq m \leq M,\ 1 \leq i \leq 20,$$

where i is the filter band index, k is the spectrum index, m is the frame number, and M is the number of frames for analysis.

In order to remove some undesired impulse noise in Eq. 4, we further smooth it by using a three-point median filter to get $x_s(m,i)$:

$$x_s(m,i) = \frac{x(m-1,i) + x(m,i) + x(m+1,i)}{3}. \quad (5)$$

Finally, the smoothed energy, $x_s(m,i)$, is normalized by removing the frequency energy of background noise, $Noise_{freq}$, to get the energy of almost pure speech signal, $X(m,i)$. The smoothed and normalized frequency energies of a clean speech signal, $X(m,i)$ is described by Eq. 6. The energy of background noise is estimated by averaging the frequency energy of the first five frames of the recording:

$$X(m,i) = x_s(m,i) - Noise_{freq} = x_s(m,i) - \frac{\sum_{n=0}^{4} x_s(n,i)}{5}. \quad (6)$$

With the smoothed and normalized energy of the i-th band of the m-th frame, $X(m,i)$, we can calculate the total energy of the almost pure speech signal at the i-th band as $E(i)$:

$$E(i) = \sum_{m=0}^{M-1} \left|X(m,i)\right|. \quad (7)$$

The goal is to select some useful bands area having the maximum word signal information. It is obvious that $E(i)$ in Eq. 7 is a good indicator since the band with higher $E(i)$ contains more pure speech information.

Based on this computation for each band area cited in Table 4, the Fig. 3 shows the six energies computed of the six areas frequencies that specify each vocal segment of a speech Arabic signal for 2 s.

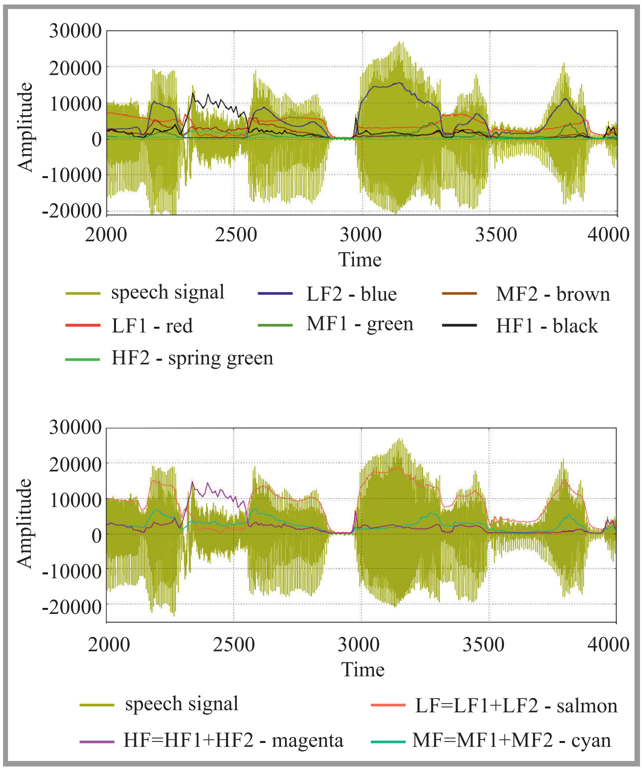

Fig. 3. Energies of six region bands in an Arabic speech signal frame. (See color pictures online at www.nit.eu/publications/journal-jtit)

4.3. Segmentation Algorithm

In each frame of an Arabic speech signal, the segmentation is based on three steps.

1. All closure and fricative phonemes for all point in segment where HF1 energy signal is greater than the sum of the energies signals LF1 and LF2 are selected (Fig. 4).
2. The vocalic segment for all point in segment where the sum of the energies signals LF1 and LF2 is greater than the mean of the energy signal HF2 is selected (Fig. 5).
3. The vowels and other voiced consonant in vocalic segment for all segment are selected (Fig. 6) where we are:
   - crossing between energy signal LF1 and energy signal LF2,
   - crossing between energy signal LF1 and the sum of the energies signals MF1 and MF2,
   - crossing between energy signal LF2 and the sum of the energies signals MF1 and MF2.

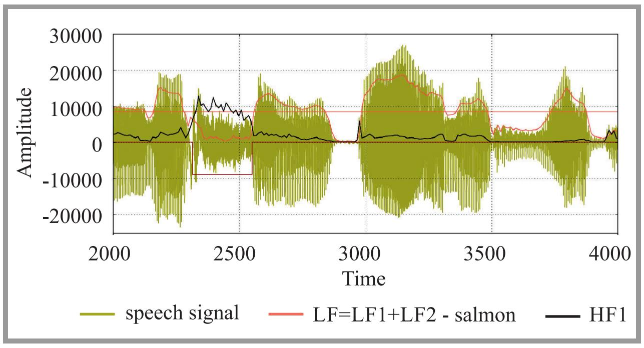

Fig. 4. Selection of a closure phoneme (first step).

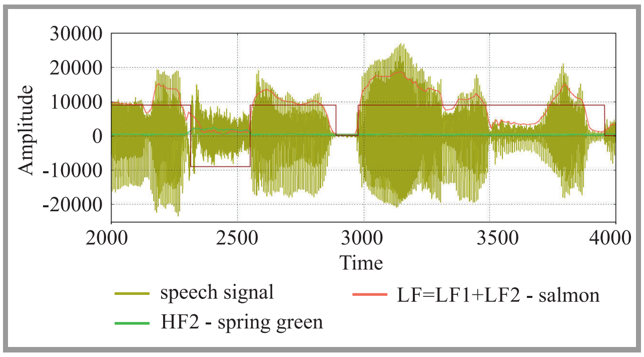

Fig. 5. Selection of vocalic segment (second step).

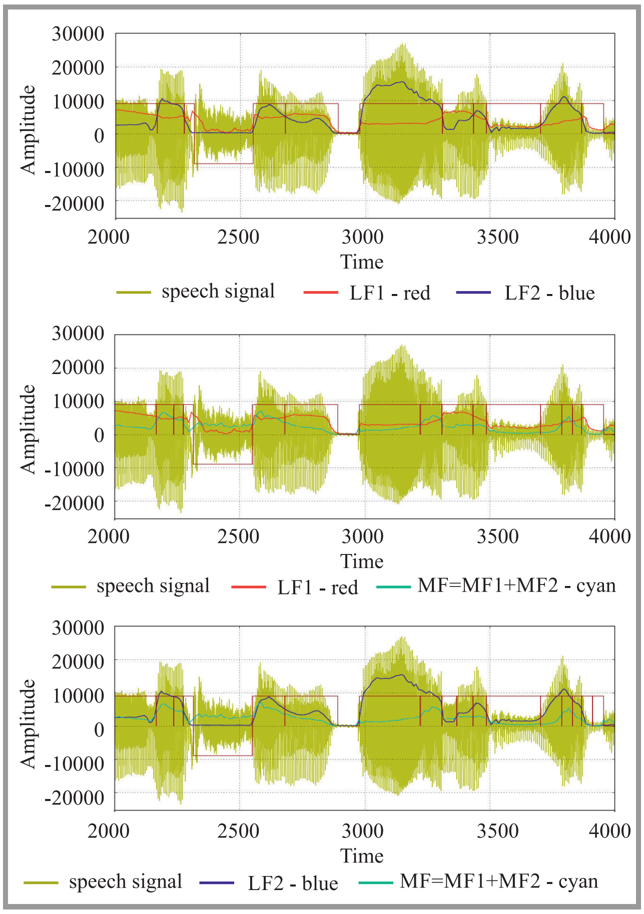

Fig. 6. Selection of vowels and voiced consonant (third step).

5. Experimentation and Evaluation

5.1. Data Set

The speech was recorded at a sampling rate of 44.1 kHz by means of a 16-bit mono analog-to-digital converter (ADC) per sample and they were down sampled to 16 kHz. Data are recorded with the help of a unidirectional microphone using Audacity recording tool in a normal room with minimum external noise. Ten subjects (10 male) in the 22–35 age range were participated in the recording process. All subjects were monodialectal speakers of MSA. They were free of any speech or hearing disorders by self-report based on a screening interview and as later judged by the investigator during the recording session. Each subject recorded twenty verses in Quranic recitation according to the *tajweed rules*. Then, all the files recorded of the data set are segmented into fixed size of 30 s. Additionally, a silence period is added to the beginning and end of each sample file. The input speech data are pre emphasized with coefficient of 0.97 using a first order digital filter and then window by a Hamming window. The resulting windowed frames of 20 ms are used for the phoneme boundary detection in our experiment. For comparison of boundaries detection does with proposed algorithm, the task of transcription of phonemes for our entire data set is done by an expert phonetician.

5.2. Performance Measure

In order to evaluate the proposed algorithm, the metrics required for speech segmentation performance evaluation are used whose definitions are as follow:

- H_R (hit rate): represents the rate of correctly detected boundaries ($\frac{N_H}{N_R}$). It utilizes the number of correctly detected boundaries (N_H) and the total number of boundaries (N_R);
- F_A (false alarm rate): represents the rate of erroneously detected boundaries $\frac{(N_T - N_H)}{N_T}$, which utilizes the total number of detected boundaries N_T and the number of correctly detected boundaries N_H;
- OS (over segmentation rate): shows how much more (or less) is total number of algorithm detections, compared to the total number of reference boundaries taken from the manual transcription $\frac{(N_T - N_R)}{N_R}$;
- PCR (precision rate) $=1 - F_A$: describes the likelihood of how often algorithm identifies a correct boundary whenever a boundary is detected.

The overall quality of proposed algorithm is described by computing F_{mesure} from precision rate and hit rate whose expression is $F_{mesure} = \frac{(2 \times PCR \times H_R)}{(PCR + H_R)}$. Another global measure, referred to as the R_{mesure}, decreases as the distance to the target grows, i.e. similarly as the F_{mesure} does, but is critical towards over-segmentation [32]. It calculated by $R_{mesure} = 1 - \frac{(|r1| + |r2|)}{2}$ with $r1 = \sqrt{(1-H_R)^2 + OS^2}$ and $r2 = \frac{(H_R - OS - 1)}{\sqrt{2}}$.

5.3. Performance Evaluation

By observing the Figs. 7 and 8, the segmentation appears in concordance with the spectrum. Compared to the manual transcription showed in the two figures, segmentation gives more information characteristic definition of the speaker and the phonemes have better boundaries.

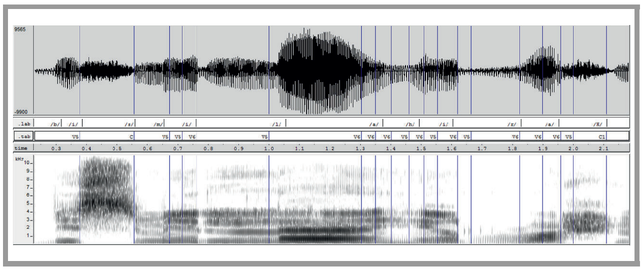

Fig. 7. Segmentation of Basmala (Surat Fatiha – Holy Coran).

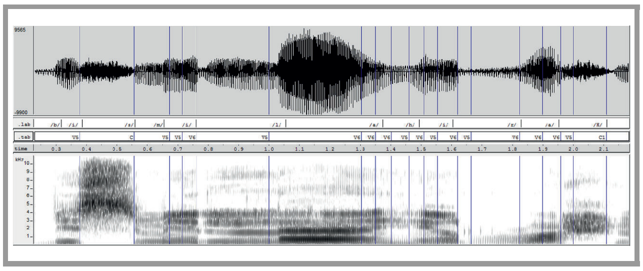

Fig. 8. Segmentation Verset 02 of Fatiha (Holy Coran).

The above calculations (Subsection 4.3) were performed for the analysis of the results obtained through the application of aforementioned algorithm.

The methodology was repeated on 83 different files of Quranic Arabic speech obtained from trained speakers.

Table 5
Segmentation performance

Files	N_H	N_R	N_T	$F_{measure}$ [%]
Speaker 01	87	109	105	81.31
Speaker 02	78	118	108	69.03
Speaker 03	76	105	98	74.88
Speaker 04	84	113	98	79.62
Speaker 05	83	109	101	79.05
Speaker 06	103	123	112	87.66
Speaker 07	105	127	118	85.71
Speaker 08	96	117	103	87.29
Speaker 09	107	124	115	89.54
Speaker 10	85	111	102	79.81
Mean measure	90	115	106	81.39

Speech signal was divided into different frames. For each frame, the trends of the signal to find the number of consecutive boundaries specifying phonemes were checked. As a result, each vowel or consonant detected, starting boundary, ending boundary of each phoneme is transcript.

To illustrate this, the results generated algorithmically from 10 different files of various speakers are presented in Table 5. The table shows the total number of different limits and measuring performance during the application of the proposed methodology.

5.4. Comparison Test

The proposed method was compared with the three same kinds of segmentation methods using mean $F_{measure}$ shown in Table 6. The first method [33] uses average level crossing rate (ALCR) and root-mean-square (RMS) energy to detect the phonetic boundary between obstruent initial consonant and preceding/following vowel. The second method [34] uses frequency synchrony and average signal level as input to a two-layered support vector machine based (SVM) system to detect phoneme boundaries. The third method [35] uses unsupervised phoneme boundary detection based on band-energy tracing technique.

Table 6
Comparison of segmentation performance

Method	*PCR* [%]	H_R [%]	$F_{measure}$ [%]
First method [33]	79.82	78.83	79.32
Second method [34]	81.12	78.91	79.99
Third method [34]	82.33	75.07	78.53
Proposed method	85.11	78.01	81.39

6. Conclusion

This work proves that it is possible to extract the information of phonemes from the energy of the acoustic signal. Following the formant technique, a study is done on Modern standard Arabic vowels. It shows that it has six basic vowels included in the constricting of vocal tract that has permit to the segmentation to be deployed in proposed system. The system shows that the formants are very effective for detecting phonemes correctly.

The experimentation shows that with this method, we can detect a mean of 81% of all boundaries manually transcribed of a speech raw file, and give better result than other methods developed in the literature.

References

[1] K. Vicsi and D. Sztahó, "Recognition of emotions on the basis of different levels of speech segments", *J. of Adv. Comput. Intell. and Intelligent Inform.*, vol. 16, no. 2, pp. 335–340, 2012.

[2] K. Vicsi, D. Sztahó, and G. Kiss, "Examination of the sensitivity of acoustic-phonetic parameters of speech to depression", in *Proc. 3rd IEEE Int. Conf. on Cognitive Infocommun. CogInfoCom 2012*, Kosice, Slovakia, 2012, pp. 511–515 (doi: 10.1109/CogInfoCom.2012.6422035).

[3] K. Vicsi, V. Imre, and G. Kiss, "Improving the classification of healthy and pathological continuous speech", in *Proc. 15th Int. Conf. Text, Speech and Dialogue TSD 2012*, Brno, Czech Republic, 2012, pp. 581–588.

[4] J. P. Goldman, "EasyAlign: An automatic phonetic alignment tool under Praat", in *Proc. 12th Ann. Conf. of the Int. Speech Commun. Assoc. Interspeech 2011*, Florence, Italy, 2011.

[5] B. Bigi and D. Hirst, "Speech phonetization alignment and syllabication (SPPAS): A tool for the automatic analysis of speech prosody", in *Proc. 6th Int. Conf. Speech Prosody*, Shanghai, China, 2012.

[6] S. Brognaux and T. Drugman, "HMM-based speech segmentation: Improvements of fully automatic approaches", *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 1, pp. 5–15, 2016.

[7] G. Gosztolya and L. Toth, "Detection of phoneme boundaries using spiking neurons", in *Proc. 9th Intell. Conf. on Artif. Intell. and Soft Comput. ICAISC 2008*, Zakopane, Poland, 2008, pp. 782–793.

[8] E. C. Zsiga, *The Sounds of Language: An Introduction to Phonetics and Phonology*. Chichester, UK: Wiley, 2012.

[9] M. Malcangi, "Soft computing approach to segmentation of speech in phonetic units", *Int. J. of Computers and Commun.*, vol. 3, no. 3, pp. 41–48, 2009.

[10] G. Kiss, D. Sztahó, and K. Vicsi, "Language independent automatic speech segmentation into phoneme-like units on the base of acoustic distinctive features", in *Proc. 4th IEEE Int. Conf. on Cognitive Infocommun. CogInfoCom 2013*, Budapest, Hungary, 2013, pp. 579–582.

[11] A. Stolcke *et al.*, "Highly accurate phonetic segmentation using boundary correction models and system fusion", in *Proc. of IEEE Int. Conf. on Acoust., Speech and Signal Process. ICASSP 2014*, Florence, Italy, 2014, pp. 5552–5556 (doi: 10.1109/ICASSP.2014.6854665).

[12] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries", *J. of the Acoust. Soc. of America*, vol. 127, no. 2, pp. 1084–1095, 2010 (doi: 10.1121/1.3277194).

[13] M. Sharma and R. Mammone, "Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge", in *Proc. of Int. Conf. on Spoken Lang. Process. ICSLP 96*, Philadelpia, USA, 1996, pp. 1237–1240.

[14] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries", in *Proc. 9th Int. Conf. on Spoken Lang. Process. INTERSPEECH 2006 – ICSLP*, Pittsburgh, PA, USA, 2006, pp. 645–648.

[15] Y. A. Alotaibi and S. A. Selouani, "Evaluating the MSA West Point Speech Corpus", *Int. J. of Comp. Process. of Lang.*, vol. 22, no. 4, pp. 285–304, 2009.

[16] O. A. A. Ali, M. M. Moselhy, and A. Bzeih, "A comparative study of Arabic speech recognition", in *Proc. 16th IEEE Mediterranean in Electrotech. Conf. MELECON 2012*, Hammamet, Tunisia, 2012.

[17] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling", in *Proc. of Worksh. on Computat. Approaches to Semitic Lang.*, Athens, Greece, pp. 53–61, 2009.

[18] N. Hajj and M. Awad, "Weighted entropy cortical algorithms for isolated Arabic speech recognition", in *Proc. Int. Joint Conf. on Neural Netw. IJCNN 2013*, Dallas, TX, USA, 2013 (doi: 10.1109/IJCNN.2013.6706753).

[19] J. F. Bonnot, "Experimentale de Certains aspects de la germination et de l'emphase en Arabe", *Travaux de l'Institut Phonétique de Strasbourg*, vol. 11, pp. 109–118, 1979 (in French).

[20] M. Alkhouli, "Alaswaat Alaghawaiyah", Daar Alfalah, Jordan, 1990 (in Arabic).

[21] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Admissible wavelet packet sub-band-based harmonic energy features for Hindi phoneme recognition", *J. of IET Sig. Process.*, vol. 9, no. 6, pp. 511–519, 2015.

[22] A. Nabil and M. Hesham, "Formant distortion after codecs for Arabic", in *Proceeding of the 4th Int. Symp. on Commun. Control and Sig. Process. ISCCSP 2010*, Limassol, Cyprus, 2010, pp. 1–5.

[23] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*. Wiley, 2000.

[24] Y. M. Seddiq and Y. A. Alotaibi, "Formant based analysis of vowels in Modern Standard Arabic – Preliminary results", in *Proc. 11th Int. Conf. on Inform. Sci., Sig. Process. and their Appl. ISSPA 2012*, Montreal, QC, Canada, 2012, pp. 689–694.

[25] L. Besacier, J. Bonastre, and C. Fredouille, "Localization and selection of speaker-specific information with statistical modeling", *Speech Commun.*, vol. 31, pp. 89–106, 2000.

[26] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey, "Contrasting the effects of different frequency bands on speaker and accent identification", in *Proc. of IEEE Sig. Process. Lett.*, vol. 19, no. 12, pp. 829–832, 2012.

[27] A. M. Selmini and F. Violaro, "Acoustic-phonetic features for refining the explicit speech segmentation", in *Proc. 8th Ann. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2007*, Antwerp, Belgium, 2007, pp. 1314–1317.

[28] A. M. Selmini and F. Violaro, "Improving the Explicit automatic speech segmentation provided by HMMs", in *Proc. of the Int. Worksh. on Telecommun. IWT 2007*, Santa Rita do Sapucaí, Brazil, 2007, pp. 220–226.

[29] M. A. Ben Messaoud, A. Bouzid, and N. Ellouze, "Automatic segmentation of the clean speech signal", *Int. J. of Elec., Comp., Energe., Electron. & Commun. Engin.*, vol. 9, no. 1, pp. 114–117, 2015.

[30] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", PhD Thesis, University of Maryland, College Park, USA, 2004.

[31] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *Proc. IEEE Trans. Acoust. Speech & Sig. Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[32] O. J. Rasanen, U. K. Laine, and T. Altosaar, "An improved speech segmentation quality measure: the R-value", in *Proc. 10th Ann. Conf. of the Int. Speech Commun. Assoc. INTERSPEECH 2009*, Brighton, UK, 2009, pp. 1851–1854.

[33] S. Potisuk, "A novel method for blind segmentation of Thai continuous speech", in *Proc. of IEEE Sig. Process. & Signal Process. Edu. Worksh. SP/SPE 2015*, Snowbird, UT, USA, 2015, pp. 415–420.

[34] S. King and M. Hasegawa-Johnson, "Accurate speech segmentation by mimicking human auditory processing", in *Proc. of IEEE Int. Conf. on Acoust., Speech & Sig. Process. ICASSP 2013*, Vancouver, BC, Canada, 2013, pp. 8096–8100.

[35] D.-T. Hoang and H.-C. Wang, "A phone segmentation method and its evaluation on mandarin speech corpus", in *Proc. of 8th Int. Symp. on Chinese Spoken Lang. Process. ISCSLP 2012*, Hong Kong, China, 2012, pp. 373–377.

Noureddine Lachachi received his Ph.D. in Computer Science at the University of Oran 1 Ahmed ben Bella. He is a Professor at the Department of computer science, University of Oran 1 Ahmed ben Bella. His research interests include issues related to speech processing specialized in Arabic and Maghreb dialects identification systems. He is an author of some research studies published at national and international journals, conference proceedings.

E-mail: Lach_nour@yahoo.fr
Department of Computer Science
Faculty of Exact and Applied Sciences
University of Oran 1 Ahmed Ben Bella
Oran, Algeria