A NOVEL METHOD FOR ARABIC CONSONANT/VOWEL SEGMENTATION USING WAVELET TRANSFORM

M. F. Tolba T. Nazmy A. A. Abdelhamid M. E. Gadallah

Faculty of computer and information sciences, Ain shams University, Abdelaziz.cs@gmail.com

Military Technical College.

***Abstract*:** *Automatic speech segmentation is a key step for building large vocabulary and continuous speech recognition systems. An alternative method is to use manual speech segmentation, which is tedious, time consuming, subjective and error prone. Many automatic speech segmentation methods have been proposed based on linguistic information such as phonetic transcription (phonetic string) but for real time systems this phonetic transcription is not always available. In this paper we propose a new algorithm for Arabic speech Consonant and Vowel (C/V) segmentation without linguistic information. This new method is based on wavelet transform and spectral analysis and focuses on searching the transient between Consonant and Vowel parts in certain levels from wavelet packet decomposition. To verify the proposed scheme, some experiments have been performed using a set of words (20 Arabic words), each word recorded six times. The accuracy rate is about 88.3% for Consonant/vowel segmentation. This rate remains fixed with low SNR value as well as high SNR.*

***Keywords:* Automatic speech recognition (ASR), Modern Standard Arabic (MSA), Wavelet Packet Transform (WPT), Spectral Analysis.**

1. INTRODUCTION

Speech segmentation is defined as the process of partitioning an entire speech into some isolated units with optimal boundaries [1]. Most automatic speech segmentation procedures use the associated linguistic knowledge such as orthographic and phonetic information [2].

In order to give speech recognition system the ability to recognize a large number of words pronounced continuously in a speech, the system should be trained to recognize the basic sub-word units of the underlying language. Before that, it should have a method to segment a speech utterance into its basic sub-word units. A vast majority of the currently available speech processing systems including medium to large vocabulary speech recognition [2] and [3], speaker recognition systems [1] and [4], and language identification systems [5], are based on acoustic sub-word units [2].

Intensive studies have been conducted on speech segmentation by employing different features and methods. Inspite of these research efforts, high accuracy speech segmentation is achieved for noise free speech utterances. T. El Arif [6] presented a theoretical framework and application of MSA speech segmentation using dynamic level building, achieving accuracy of 81%. However Amr.Gody [7] presented a speech segmentation method based on wavelet transform, achieving accuracy of 95% when using certain selected bands from the wavelet transform (6 bands are used). Meanwhile Wang, Stephen [8, 9] introduced some features for speech segmentation (i.e. counting zero crossing, energy profile, and pitch information), but the difficulty of using these features is in setting appropriate thresholds. Besides, it is found that these simple features based methods cannot offer satisfactory results particularly when low signal to noise ratios (SNR) are taken into consideration. Many other works have been conducted to improve the accuracy of speech segmentation for low SNR are found in [8, 9].

The goal of the work described in this paper is to develop a new algorithm for MSA speech segmentation based on wavelet transform and spectral analysis. The idea of the algorithm is to detect the start and end of each vowel in the speech utterance and the duration form vowel to vowel is considered as one or double consonants, and the duration from the start of the speech utterance to the first vowel is considered as the beginning consonant. That is based on the rules that control the existence of consonants and vowels in the Arabic speech such as, no transition from vowel to vowel, and there always one of six patterns CV, CV:, CVC, CVCC, CV:C, CV:CC [7], where C denotes a consonant, V denotes a vowel, and V: denotes a long vowel as shown in figure 1.

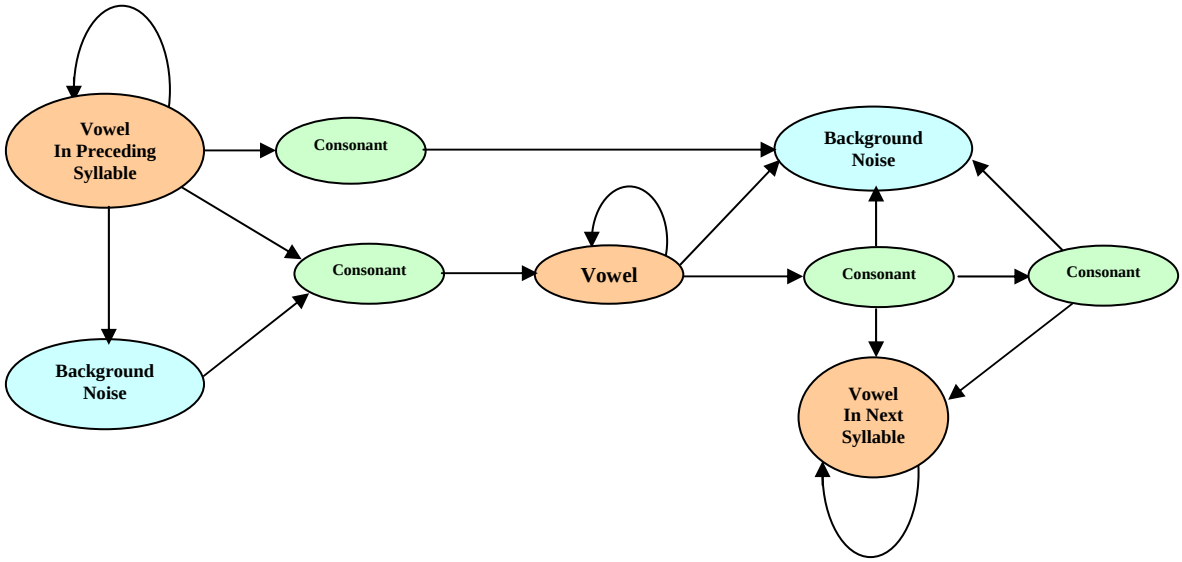


Fig 1: Illustration of all possible types of transient regions in the Arabic language of continuous speech.

There are essential difficulties exist in Arabic C/V segmentation. First of them, the existence of many dialects for the Arabic speech, these dialect differ according to the country as shown in table 1. Although these are different dialects, they intersect in the modern standard Arabic (MSA), for this reason we choose MSA for study. Another problem is the existence of two successive consonants pattern (CC). It is difficult to segment this pattern [7], and so we dealt with this pattern as single subword unit, especially that the number of occurrences for this pattern is limited in the Arabic language [7]. Besides, the low intensity of the recorded speech signal is considered as a problem since it may degrade the accuracy of the C/V segmentation algorithm, this problem is dealt with by normalization.

Table 1: Dialect regions of Arabic

Dialect region	Countries
Mahgreb Arabic	Morocco, Algeria, Tunisia, parts of Libyia.
Egyptian Arabic	Egypt, parts of Libyia.
Levantine Arabic	Syria, Lebanon, Israel, Palestine, Jordan.
Gulf Arabic	Kuwait, Qatar, Bahrain, UAE, Saudi Arabia, Oman

The rest of the paper is organized as follows; section 2 discusses the proposed segmentation algorithm. Section 3 introduces the implementation of the algorithm steps and experimental results. In section 4 the conclusion and future work are presented.

2. THE PROPOSED SEGMENTATION ALGORITHM

The steps of the algorithm are discussed in few details through the following sub-sections.

2.1 Pre-Emphasis.

The preemphasizer is used to spectrally flatten the speech signal. This is usually done by a high pass filter. The most commonly used filter for this step is the FIR filter [10] as described below:

$$H(z) = 1 - 0.95\ Z^{-1} \quad (1)$$

This filter in the time domain will be h(n) = {1,…,0.95}, and the filtering in the time domain will give the pre-emphasized signal S'(n):

$$S'(n) = \sum_{k=0}^{N-1} h(k).S(n-k) \quad (2)$$

where n is the sample index. N is the number of samples.

2.2 Wavelet Packet Transform (WPT)

Wavelets are developed in the mid 1980's by a group of researchers in France. They named them "ondeletts" which translated to English became "wavelets". The wavelet transform (WT) theory provides an alternative tool for short time analysis of quasi stationary signal, such as speech, as opposed to the traditional short-time Fourier transform (STFT). The WT has been applied widely in different speech analysis problems [10, 11, 12, 13, 14].

The WT is a non-parametric analysis tool which allows localization in both the time and frequency domains. The main difference between STFT and WT is that STFT is a constant-bandwidth analysis method; whereas WT is a constant-Q analysis method which resembles auditory filters [5]. Wavelet coefficients are obtained by computing the correlation between each wavelet and the signal. The wavelet is a small wave called "*Mother Wavelet*", from which many other waves are derived by translation and dilation. It can be defined as:

$$W_{jp}[m] = 2^{-j/2} w[2^{-j} m - p] \quad (3)$$

where j, p are scaling and translation factors.

The wavelet transform is expressed in terms of the mother wavelet W_{jp} as follows:

$$C_{jp} = \sum_m s[m].2^{-j/2}.w[2^{-j} m - p] = \sum_m s[m].w_{jp}[m] \quad (4)$$

Figure 2 introduces the block diagram of one level wavelet decomposition. This decomposition results in two sets of data namely the approximation coefficients A[n], and the detail coefficients D[n]. From the theory of wavelet, it is known that the approximation coefficients contain low frequency components while the detail coefficients contain high frequency components of the input speech utterance. We can reconstruct the high frequency components D[n] by setting the approximation coefficients to zero and perform wavelet reconstruction.

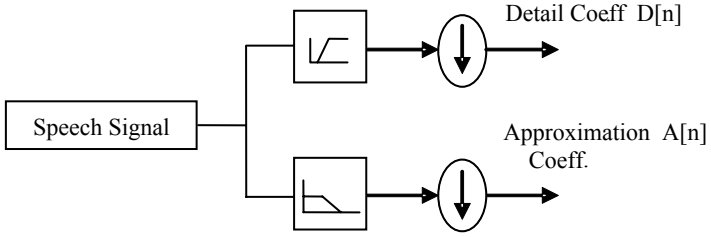


Fig 2. One Level wavelet decomposition.

Wavelet packet transform (WPT) is performed through applying the wavelet transform on both the approximation and detail coefficients, and repeating this process down to certain number of levels.

In the proposed method, we applied WPT down to 4 levels, then we select from the 4th level the nodes (4,1), (4,3), and (4,4) for the further processing. These nodes are selected based on the fact that most of the speech energy is concentrated below 1000 Hz. Since we are interested in detecting the start and end of each vowel in the speech utterance as mentioned earlier, and these vowels are characterized by their low frequency and high energy properties, we found that the Arabic vowels appear by some kind of clarity in the node (4,1). After many trials to improve the accuracy of detecting these vowels, we found also that the two other nodes (4,3) and (4,4) can help in this mission. The overall decomposition occurred on the signal is shown in figure 3.

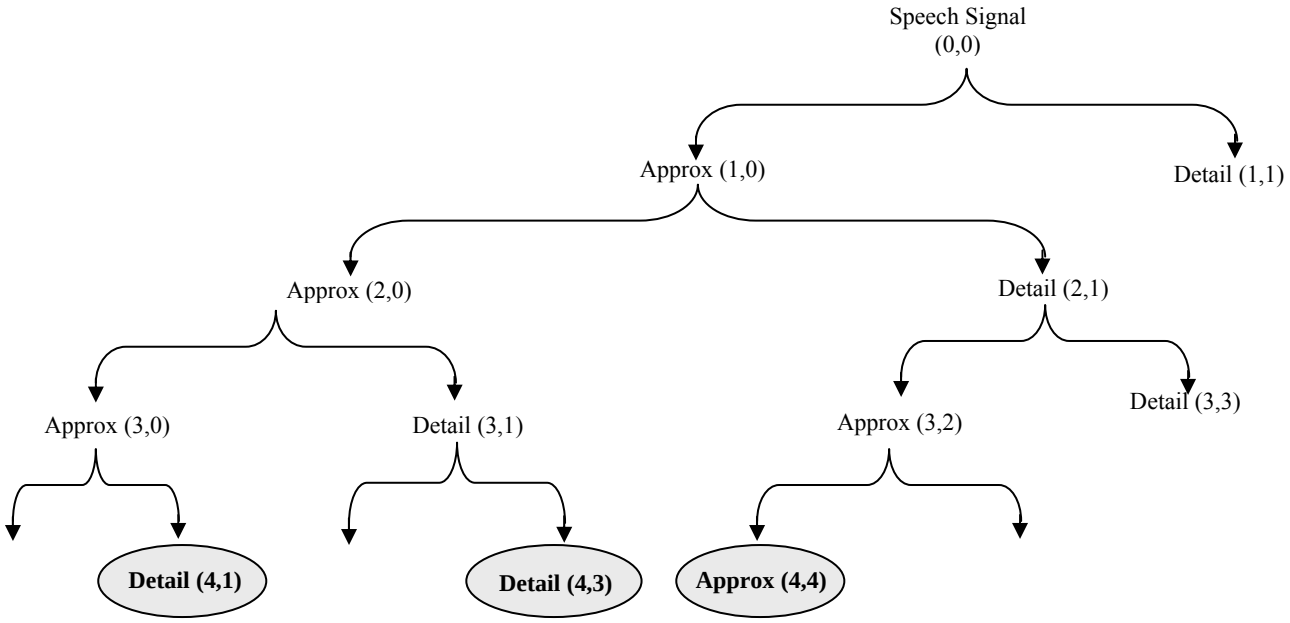


Fig 3. Wavelet packet decomposition and selected nodes are shadowed.

Then each node is used separately to reconstruct the signal. So we used $X_1(n)$ to denote the reconstructed signal from node (4,1), $X_2(n)$ to denote the reconstructed signal from node (4,3), and $X_3(n)$ to denote the reconstructed signal from node (4,4).

2.3 Signal Framing.

Signal framing is the process of breaking down the signal into successive (overlapped or not overlapped) fixed length frames. This process is essential because the speech signal is inherently considered as a random signal [10], and to process this random signal we break in to smaller frames considering the signal within these frames is periodic or quasi periodic signal. In our work we applied the signal framing to X_1, X_2, and X_3 without overlap as shown in figure 4.

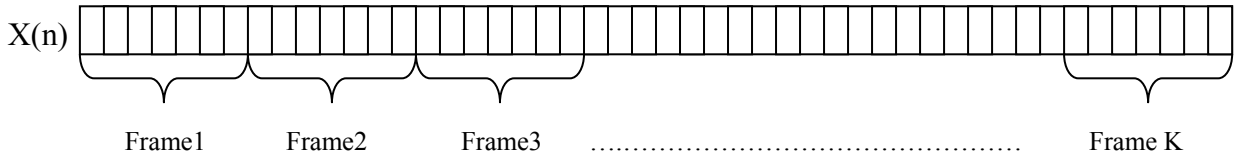


Fig 4: Framing of a sequence X(n).

2.4 Frame Windowing

The Next step is to apply a window to each frame in order to reduce signal discontinuity at either end of the frame. A commonly used window is the Hamming window [15]. It is defined as:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\Pi n}{N-1}\right) \quad (5)$$

where n is the sample index. N is the number of samples in the signal.

W (n) is applied to $X_1(n)$, $X_2(n)$ and $X_3(n)$ frames.

2.5 Fourier Transform.

Fourier transform is used to convert speech signal from time domain to frequency domain, for speech our auditory system seems to be very sensitive to the frequency characteristics. The perceptual attributes just like loudness, pitch, timbre and so on seem to have a strong correlation with the physical properties of intensity, fundamental frequency, and spectral shape. All of them mainly depend on the frequency characteristics of speech signal, although the connection between them is complex. So nowadays speech signal processing is mainly based on the frequency analysis.

$$X(k) = \sum_{j=1}^{N} x(j)\varpi_N^{(j-1)(k-1)} \quad (6)$$

where $\varpi_N = e^{(-2\Pi i)/N}$ (7)

2.6 Line Fitting.

Line fitting the spectra to a straight line is based on the least squares principle. More details about the algorithm can be found in [15]. This algorithm fits a set of N data points (x_i,y_i) to a straight line model of the form

$$y(x) = y(x; A, B) = Ax + B \quad (8)$$

The data fitting is accomplished by minimizing the chi-square merit function

$$E^2(A, B) = \sum_{i-1}^{N}\left(\frac{y_i - Ax_i - B}{\sigma_i}\right)^2 \quad (9)$$

where σ_i is the uncertainty association with each measurement y_i. It is important to point out that the spectral data are input as unweighted data (i.e. all $\sigma_i = 1$). The result will be slope (A) and constant (B), and we used the constant part (B) as will be shown in the following sections.
This process is applied to each frame in $X_1(n)$, $X_2(n)$ and $X_3(n)$.

2.7 Initial Segmentation

Based on the fact that the spectral tilt of the vowels is negative value and the spectral tilt of the consonants is positive value [10]. So we can use this property to determine the start and end of each vowel in the speech utterance by detecting the change in sign of the spectral tilt. But after many trials it was found that using the constant from line fitting gives more accurate detection for the start and end of the vowels.

In this step, we used the result from line fitting (constant part B) applied on $X_1(n)$, $X_2(n)$, and $X_3(n)$ separately to perform the initial segmentation. This segmentation is applied by using adaptive thresholds to search for the vowels.

2.7 Averaging

The results from initial segmentation are averaged to determine the actual start and end of each vowel in the speech utterance, and implicitly the duration between two successive vowels represents a consonant, and duration from the actual start of the utterance and first vowel represents the starting consonant as mentioned earlier. Figure 5 charts the steps of the proposed algorithm.

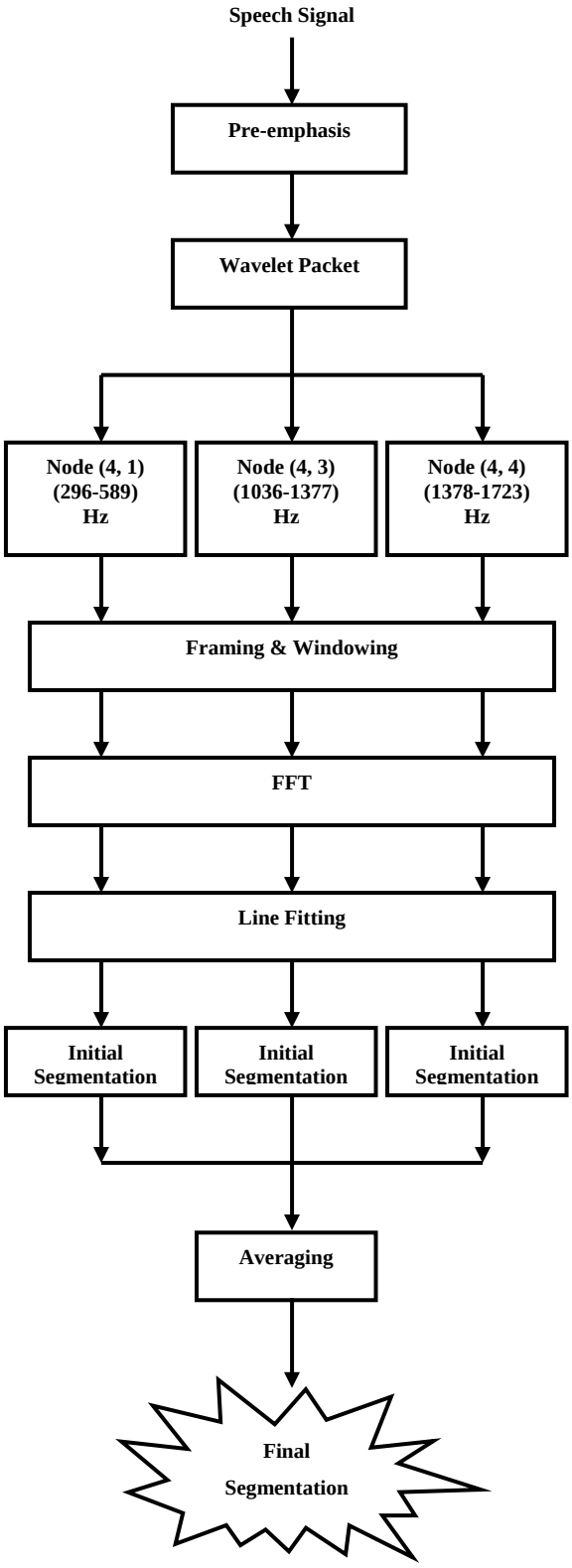


Fig5: The block diagram of the proposed algorithm steps.

3- EXPERIMENTAL RESULTS

The proposed algorithm was tested on a simple database containing 20 Arabic words articulated in modern standard Arabic. The following results show the application of the algorithm on a word from the database as shown in Figure 6-a, the word consists of the following C/V pattern CVCV: C. Wavelet packet transform is applied on the overall signal, and then the selected nodes (4, 1), (4, 3) and (4, 4) are used for segmentation.

Speech signal is reconstructed from each node, so we now have 3 versions from the speech signal each one contain a certain band of frequencies as shown in figures 6-b, 6-c, 6-d.

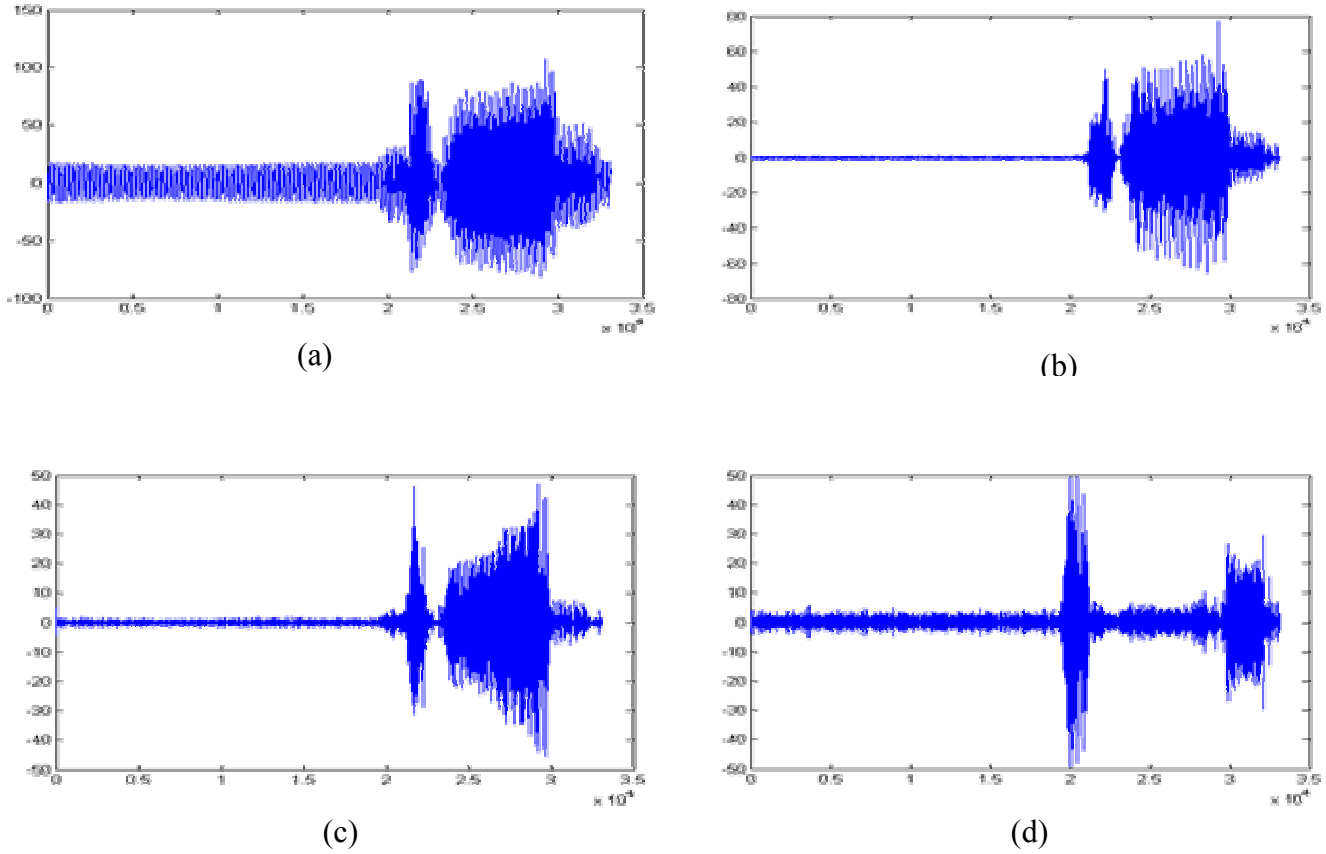


Figure 6 (a) Arabic word 'اشنون' recorded using 11025 sampling rate 16 bit/sample,
(b) reconstructed signal from wavelet node (4,1), this signal contains band (296Hz – 589Hz),
(c) reconstructed signal from wavelet node (4, 3), this signal contains band (1036Hz – 1377Hz),
(d) reconstructed signal from wavelet node (4, 4), this signal contains band (1378Hz – 1723Hz).

Each signal is preprocessed (framed with 256 samples/frame, and windowed), Fourier transform is applied on each frame, and line fitting is applied, then the constant part from line fitting is used from each frame. The result for each of the 3 version of the signal shown in figures 7-a, 7-b, 7-c respectively[1].

[1] For convenience, the results from line fitting, and from initial and final segmentations are stretched over the X-Axis.

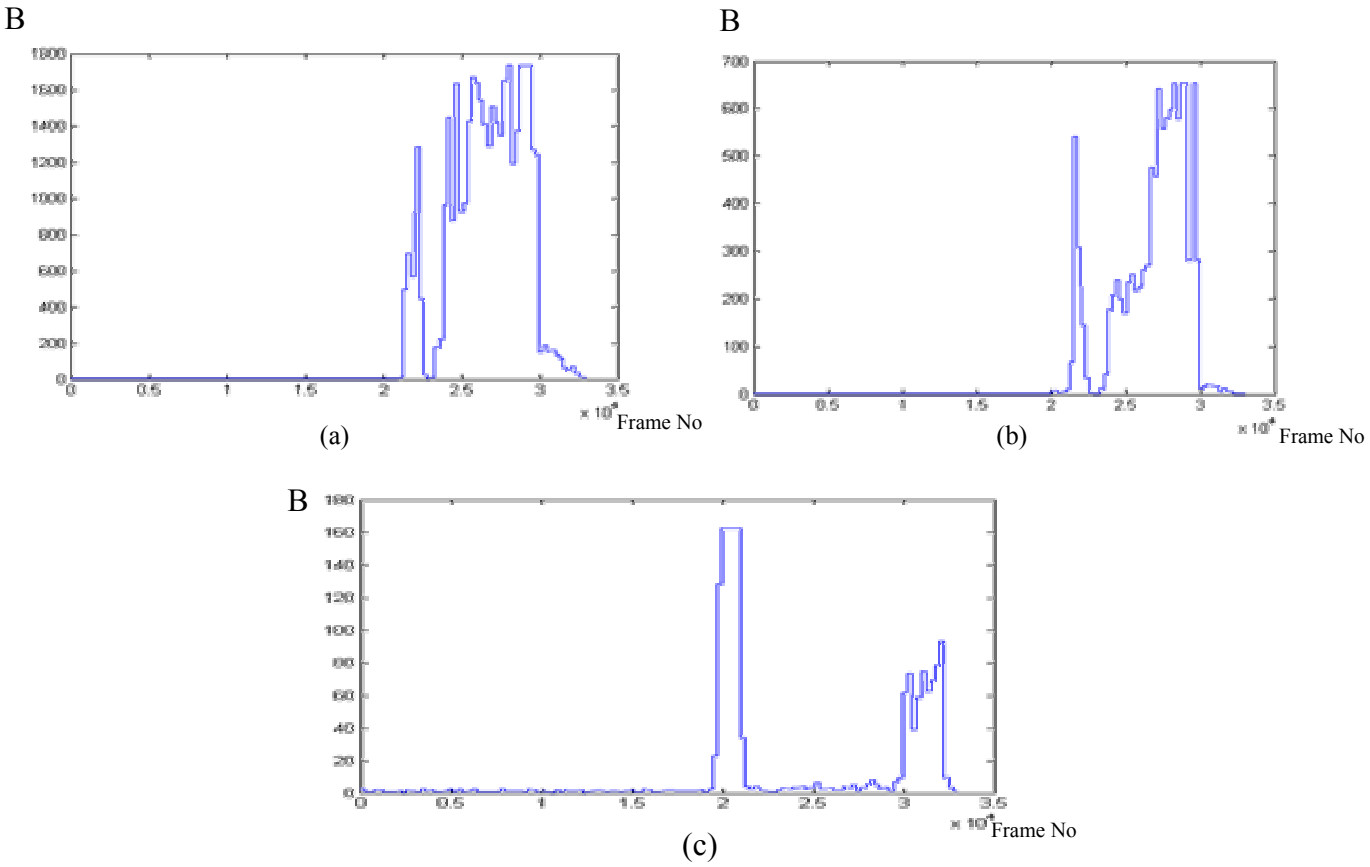


Figure 7 (a) the constant part from line fitting for the FFT on each frame in the signal reconstructed from wavelet node (4, 1).
(b) the constant part from line fitting for the FFT on each frame in the signal reconstructed from wavelet node (4, 3).
(c) the constant part from line fitting for the FFT on each frame in the signal reconstructed from wavelet node (4, 4).

The results are smoothed to eliminate unpredictable peaks, and then adaptive threshold used to determine the start and end of each segment. These adaptive thresholds[2] are as follows:

For the result in figure 7-a, it is estimated as:
Thr = 0.00011361*(mx_val^2) - 0.91749*mx_val + 1993.3 (10)
where mx_val is max value exist in the result of figure 7-a.

For the result in figure 7-b, it is estimated as:
Thr = 0.0001*(mx_val^2) - 0.2262*mx_val + 162.2080 (11)
where mx_val is max value exist in the result of figure 7-b.

For the result in figure 7-c, it is estimated as:
Thr = 0.1984*mx_val - 0.43918 (12)
where mx_val is max value exist in the result of figure 7-c.

Then after thresholding the results will be as shown in figures (8-a, 8-b, and 8-c). The final segmentation is obtained through averaging the previous results.

[2] These adaptive thresholds are determined through the mapping between the manual segmentation and the segmentation expected from the proposed algorithm.

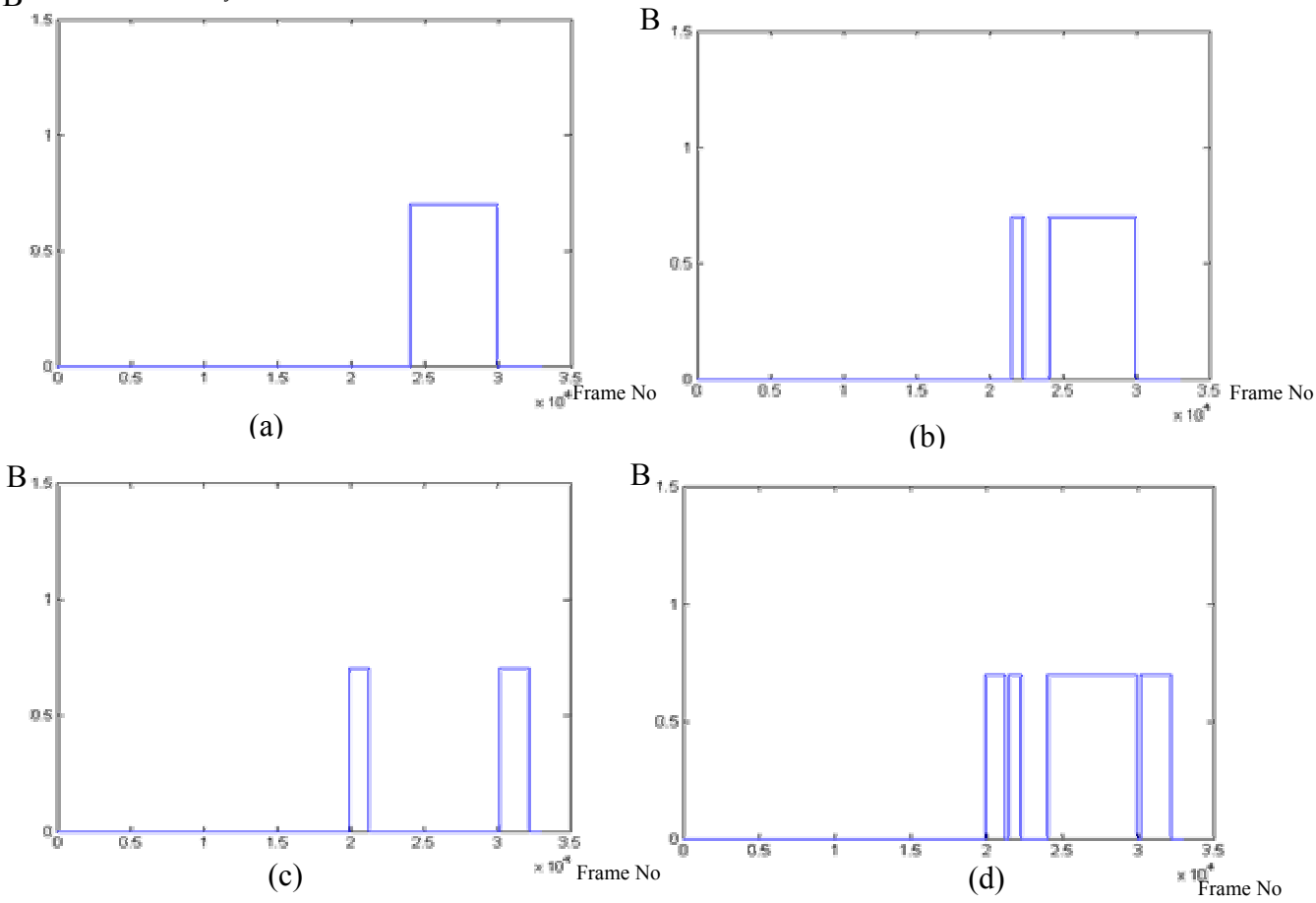


Figure 8 (a) the segment determined by thresholding result in figure 7-a, using corresponding adaptive threshold.
(b) the segment determined by thresholding result in figure 7-b, using corresponding adaptive threshold.
(c) the segment determined by thresholding result in figure 7-c, using corresponding adaptive threshold.
(d) the final segments obtained by averaging the results in figures (8-a, 8-b, and 8-c).

As shown in figure 8-d, the first segment represent phoneme 'ش', second segment represent phoneme 'و' which is Arabic vowel, third segment represent phoneme 'ـا' which is Arabic long vowel, and the final segment represent Arabic phoneme 'ن'.

From the given results, it has been noted that this method besides its capability to find boundaries of Arabic vowels, it also can find boundaries of some other consonant phonemes and that is considered as point of power in the proposed method.

The proposed algorithm was tested with high SNR as and low SNR values since it applies wavelet packet that acts as bank of filters that filters the recorded signal and focuses on certain bands which are not affected by noise[3]. The database used here contains all Arabic language's phonemes. Database consists of 20 Arabic words with 6 repetitions. The total duration is 163.08 Sec.
In order to verify the accuracy of the proposed method, the computed border locations of vowels are compared with those obtained from manual segmentation. The result is considered correct if the border location difference between the proposed method and manual is less than 13 milliseconds [16]. The measured segmentation accuracy is 88.3%. This method was compared with other methods for segmenting and was found that the proposed algorithm gives an acceptable accuracy for MSA C/V segmentation.

[3] This means that these bands are limited with certain frequencies that not increased by the existence of noise.

The complexity of the proposed algorithm depends only on the complexity of the wavelet decomposition and the complexity of FFT. These complexities are small enough for real time recognition systems.

For signals with low SNR, the performance of the proposed algorithm is still efficient. The following figures are the results for a signal containing the same Arabic word 'شئون' with SNR = 2. (Note: the results are still the same as for high SNR).

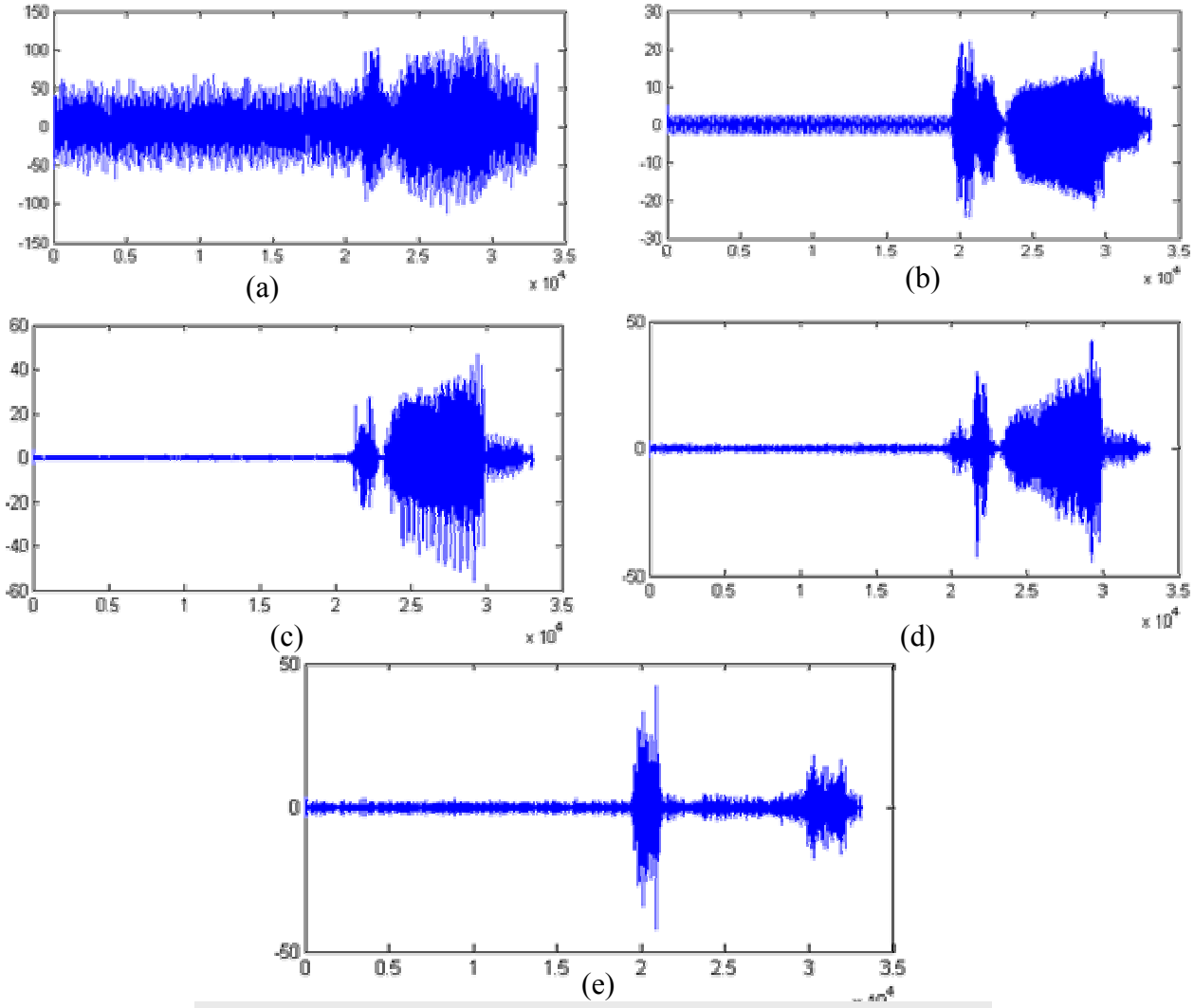
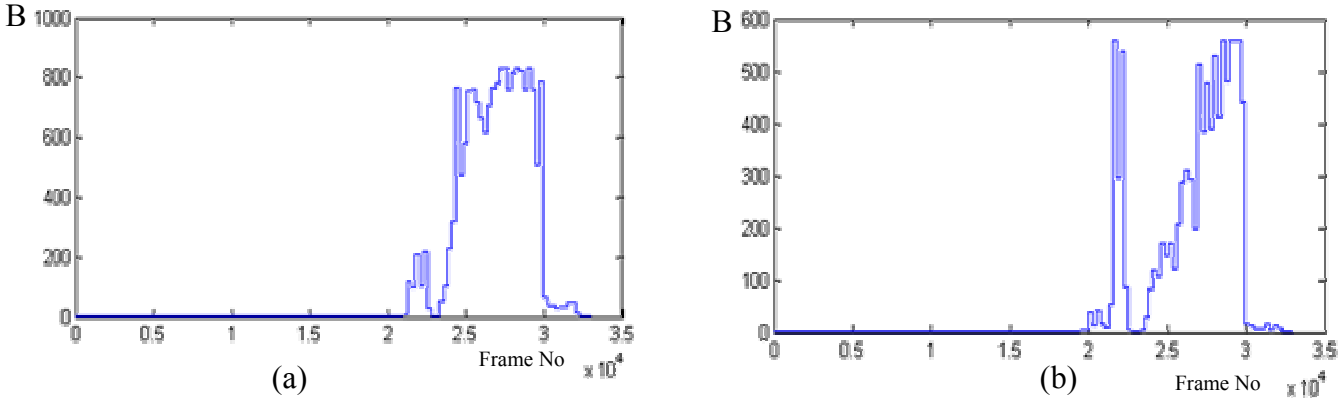
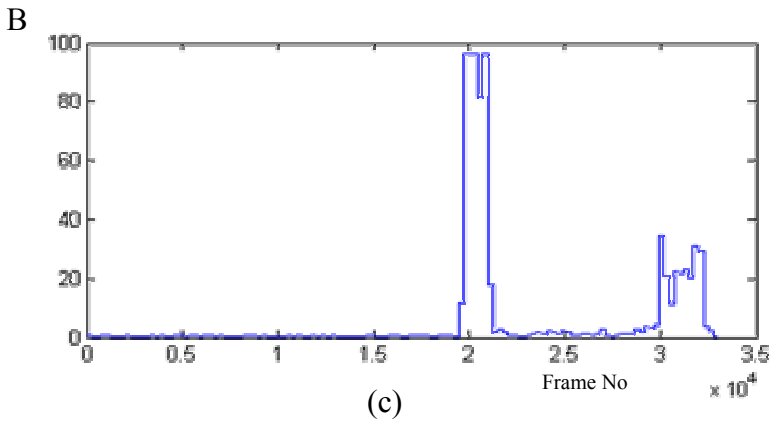


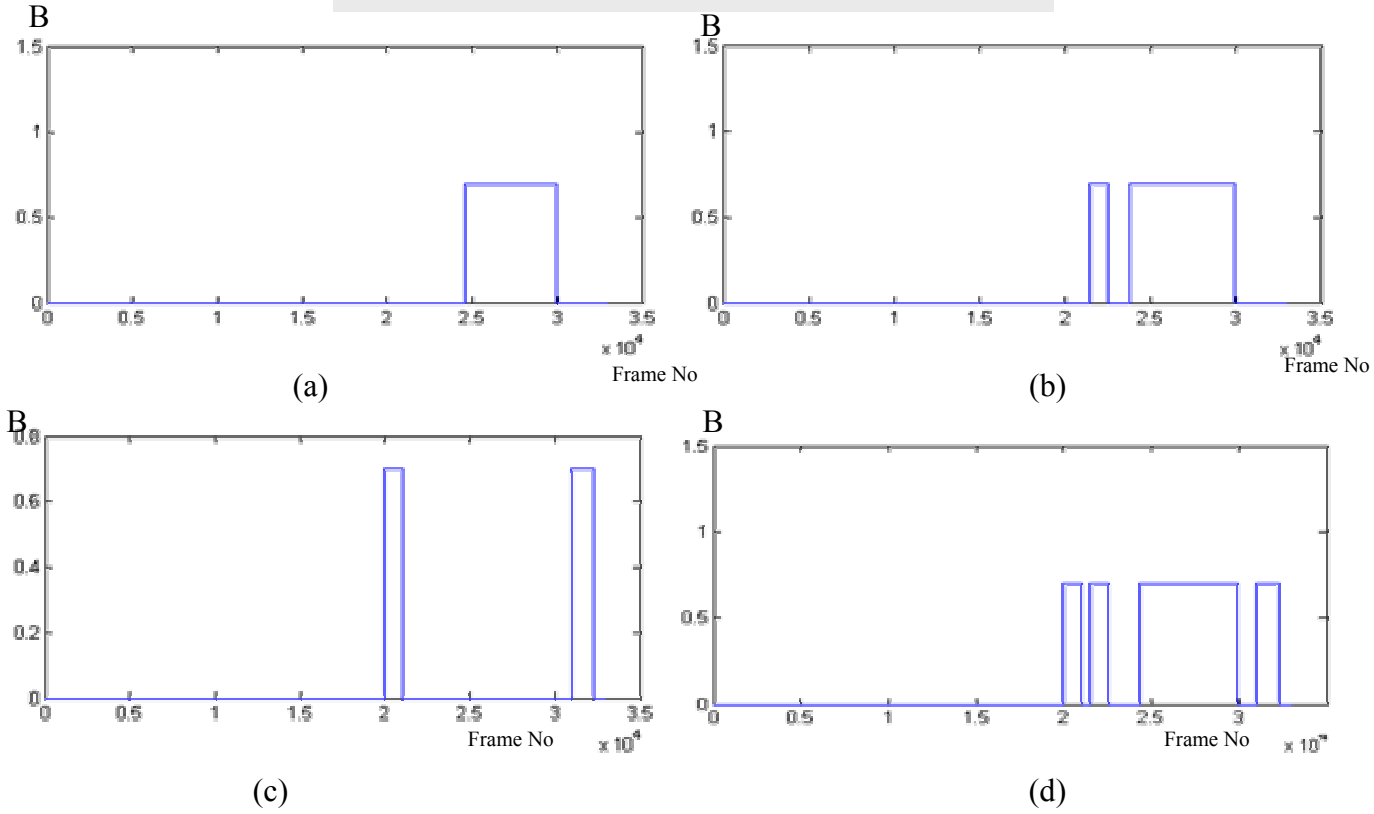
Figure 9 (a) Signal with SNR = 2.
(b) Signal after Preemphasis.
(c) Signal of node (4, 1) from wavelet packet decomposition.
(d) Signal of node (4, 3) from wavelet packet decomposition.
(e) Signal of node (4, 3) from wavelet packet decomposition.





(c)

Figure 10 (a) Constant part from line fitting for node (4, 1).
(b) Constant part from line fitting for node (4, 3).
(c) Constant part from line fitting for node (4, 4).



(a) (b)

(c) (d)

Figure 11-(a) Initial segmentation for Signal of node (4, 1).
(b) Initial segmentation for Signal of node (4, 3).
(c) Initial segmentation for Signal of node (4, 4).
(d) Final segmentation for the Arabic speech Signal.

4. CONCLUSIONS AND FUTURE WORK

In this paper a novel method for Arabic speech segmentation was presented, depending on wavelet transform and spectral analysis. This method doesn't depend on any features to be extracted from speech signal, but deals direct with the signal itself which speeds up time needed for segmentation. Besides, the proposed algorithm can be implemented using parallel processing; since the reconstructed signal from the 3 wavelet nodes, are processed independently and so can be implemented in parallel, which also will help in speeding up the processing time.

The algorithm works efficiently with low SNR as well as high SNR, since it depends on certain bands of frequencies in the speech utterance.

This work is totally implemented using MATLAB, and all it functions are integrated in a single module, for easily usage with other modules.

The future work will be integrating this segmentation module with a recognition module using HTK for building a complete Arabic speech recognizer based on sub word units.

REFERENCES

[1] M. Sharma and R. Mammone. "*Sub-word Based Text Dependent Speaker Verification System with User Selectable Passwords*", *In Proceedings of ICASSP*, pp. 93-96, 1996.

[2] C. Becchetti and L. R. Ricotti. *Speech Recognition: Theory and C++ Implementation*,Fondazione Ugo Bordoni, Rome, Italy, Jhon Wiley & Sons ltd, 1999.

[3] Mitchell K. F. Lee, " *Automatic Speech Recognition - The Development of the SPHINX system*.", Kluwer Academic, Boston, 1989.

[4] A. E. Rosenberg, C-H Lee and F.K. Soong. "*Sub-word Unit Talker Verification using Hidden Markov Models*", *In Proceedings of ICASSP*, pp. 269-272, 1990.

[5] Y. K. Muthusamy, E. Barnard and R. A. Cole. "*Reviewing Automatic Language Identification*"., *IEEE Signal Processing magazine*, 11(4): pp. 33-41, October 1994.

[6] T. I. El Arif, Z. T. Fayed, M. E. Gad Allah, and A.I.Eldesoky "*Automatic phonetic segmentation of Arabic speech without linguistic knowledge*" , ICICIS conference, pp. 265-271, 2005.

[7] Amr M.Gody "*Speech Processing Using Wavelet Based Algorithms*", PhD Thesis Cairo University 1999.

[8] J. F. Wang, C. H. Wu, S. H. Chang, and J. Y Lee, "*A Hierarchical Neural Network Model Based on a C/V Segmentation Algorithm for Isolated Mandarin Speech Recognition*," IEEE Trans. On Signal Processing, vol 39, No 9, pp.2131-2136, September 1991.

[9] Stephen W. K. Fu, C. H. Lee, O. L. Clubb, "A Robust C/V Segmentation Algorithm for Cantonese," IEEE TENCON, pp. 42-45, 1996.

[10] M. Nilsson and M. Ejnarsson. *"Speech Recognition using Hidden Markov Model: Performance evaluation in Noisy Environment*", Master thesis, Department of telecommunications and signal processing, Blekinge Institute of technology, Sweden, March 2002.

[11] A. Van Erp, C. Houben, B. Barry, M. Grice, L. J. Boë, G. Braun, P. Cosi, N. Dyhr, G. Perennou, N. Vigouroux and D. Auteserre. "*A Unified Approach to The Labeling of Speech: First Multi-lingual Results*", *Eurospeech- 89*, vol 2, pp. 88-91, 1989.

[12] Al-Ani, S. H. ,"*Arabic Phonology*", The Hague: Mouton, 1970.

[13] Deller, J. R. Jr., Proakis, J. G., and Hansen, J. H. L.,"*Discrete-time Processing of Speech Signal"s*, Macmillan Publishing company, New York 1993.

[14] A. E. Rosenberg, C-H Lee and F.K. Soong. "*Sub-word Unit Talker Verification using Hidden Markov Models*", *In Proceedings of ICASSP*, pp. 269-272, 1990.

[15] Pickover C A and Khorasani A L, *"Fractal characterization of speech waveform graphs*, *Computer and Graphic"s*, Vol. 10, no. 1, pp. 51-61, 1986.

[16] S.Ratsameewichai, N.Theera-Umpon "*Thai Phoneme Segmentation using Dual-Band Energy Contour*". 2001.