Arabic speech segmentation: Automatic verses manual method and zero crossing measurements

Mohammed A. Al-Manie, Mohammed I. Alkanhal and Mansour M. Al-Ghamdi
Computer Research Institute, KACST, P.O. Box 6080, Riyadh 11442, Saudi Arabia
malmanie@gmail.com; manie@kacst.edu.sa; alkanhal@kacst.edu.sa; mghamdi@kacst.edu.sa

Abstract

In this paper, a comparison between the automatic and manual approach of speech segmentation for the Arabic speech is conducted. In this approach, the automatic method, using the energy level measurement is compared to the manual segmentation procedure. The traditional zero crossing method commonly used for speech processing is also included in this work. The energy measurement method is based on dividing the uttered tokens into different levels. For instance, the Arabic language phonemes are divided into two energy regions: unvoiced phonemes which can be categorized as low energy include the sounds /س / (/s/) and / هـ / (/h/). On the other hand, vowels and semi-vowels such as / ـَ / (فتحه) (/ ‘a /) and /و / (/w/) are labeled as high energy. Voiced fricatives, for instance, the sounds / ز / (/z/) and /ع/ (/aa/) are classified as high energy phonemes.

Keywords: Manual speech segmentation, automatic speech segmentation, voiced phonemes, voiceless phonemes.

Introduction

Speech processing involves a number of sub-problems, including but not limited to, speech recognition, speech synthesis, and building large prosaically labeled speech corpuses. Due to the useful applications involved, these areas have received a great deal of attention by many researchers around the world and were implemented on a number of different languages (Hemert 1991; Alghamdi, 2003). However, before proceeding to solve the aforementioned problems, one must take into account the preprocessing stages that must be performed, such as normalization and noise reduction. Another important preprocessing stage is the partitioning of the speech signal under study into manageable and will defined segments in order to facilitate the upcoming tasks. This process is known as speech segmentation, which can be defined as the detection of boundaries between words, syllables, or phonemes in a certain natural language.

Hence, the segmentation of a speech utterance may be divided into two different parts. The first is known as lexical where the sentence is broken up into smaller parts such as words. The second is referred to as phonetic, in which the speech signal is broken up into phones. Grammar, semantics and context of the spoken language under study are very important factors that must be taken into account during speech segmentation research. This field recently received a great deal of interest by a number of researchers due to its useful applications in the field of natural language processing (Lee *et al.,* 2003).

Consequently, one needs an efficient, accurate and a simple technique to perform segmentation of speech signals. The manual procedure performed by specialized phoneticians has been the traditional way to tackle this problem. However, this approach is based on listening and visual judgment in order to detect the required boundaries, which makes it unpractical for huge data bases. This calls for an alternative approach known as automatic speech segmentation, where the task of detecting the boundaries of phonemes in a speech signal is done by a carefully chosen procedure or algorithms (Tolba *et al.,* 2005).

As a result, some researchers developed different techniques for this objective such as special algorithms based on energy level of the uttered words or sentences in an attempt to achieve higher accuracy. Speech segmentation can be defined as the process of finding the boundaries in a certain natural language between words, syllables or phonemes. The main objectives of this method are to use the result for other speech analysis problems such as speech synthesis, data training for speech recognizers or to build and label prosodic data basis (Alghamdi, 2003). Hence, it can be considered an important sub-problem for a number of fields in speech analysis and research. However, in order to perform speech segmentation, one must consider grammar, semantics and context of spoken language under study. On the other hand a tradition approach used to facilitate the segmentation process is the zero crossing measurements. The number of times the signal crosses the zero for a certain interval is defined as the zero crossing rate. This kind of approach is usually used in natural language processing to classify the speech utterance into voiced and voiceless phonemes. In general, the zero crossing rate (ZCR) for unvoiced is much higher than that for the voiced speech signal. The corresponding energy for each tested phoneme is also calculated and tabulated. The tested Arabic phonemes in this paper include voiced and voiceless fricatives and stops. In addition, the ZCR plus energy rate calculations for nasals, laterals, and trills sounds were computed for three different speakers involving all the tested categories. The second Section of this paper is about the theoretical background and procedures used to perform the segmentation of the Arabic speech. The zero crossing rate calculation method is presented in the third section including tabulated results. This part also outlines the proposed segmentation algorithm implementations for both low and high energy regions detection. In the fourth Section, discussion and analysis of the final results obtained using the applied techniques are provided. The fifth and final Section of the paper presents a summary of the findings and conclusions.

Segmentation procedures

This section provides a background on how the automatic Arabic speech segmentation techniques are implemented versus the other approach. The manual method is usually performed by a specialized phonetician with the help of time-frequency tools such as the spectrogram and also through repeated listening to the target phoneme. For the energy based technique, the Arabic phonemes are divided into different sections based on their acoustic manifestations. The onset, offset and the steady state of each sound are tagged.

In order to find the low energy region, the algorithm calculates the lowest point or the local minima in a given speech utterance which indicates the starts (onset) for this part. When the tested signal starts rising, this means the end of the low energy segment (offset). For instance, let the speech signal be $S(n)$, and $E(n)$ is the energy of the calculated for a certain window size, such as 15*ms* with an increment of 5*ms*. In this case, the energy boundaries denoted as (n_1, n_2) are calculated according to the following procedure (Essa, 2005):

$$n_1 = \arg\{\ \max(E'(n))\ \} \quad (1)$$

for $n_{\min} - \delta \le n \le n_{\min}$

$$n_2 = \arg\{\ \min(E'(n))\ \} \quad (2)$$

for $n_{\min} \le n \le n_{\min} + \delta$

Where $n_{\min} = \arg\{\ \min_n (E(n))\ \}$

That is the lowest energy point in the signal $S(n)$ and $E'(n)$ is the normalized delta energy defined as:

$$E'(n) = \sum_{i=0}^{6} \frac{E(n-i) - E(n+i)}{\min\{E(n-i), E(n+i)\}} \quad (3)$$

The algorithm proposed by Essa (2005) was modified to include high energy regions when performing automatic segmentation of the speech utterance. Some changes were also introduced to improve the performance of this method such as increasing the distance from the highest or lowest energy point so that it is made longer in one side than the other. Another important contribution of this paper is the testing of the procedure on a very large Arabic data base.

Experimental results

Zero crossing rate calculation

In this part of the paper, the zero crossing rates (ZCR); a procedure to find the number of times a certain signal passes through zero for some interval is performed. The signal in this case is the speech token which has a higher number of ZCR for unvoiced than that for the voiced. Therefore, this approach is used by some researchers in natural language processing to classify phonemes. The zero crossing can be found according to the following formula:

$$Z(k) = \frac{1}{N} \sum_{n=(k-1)\times N-1}^{k\times N-1} \frac{|\,sign(x(n+1) - sign(x(n))\,|}{2} \cdot w(m-n) \quad (4)$$

Where k is the frame number, N is the frame length, w is the window function, and $x(n)$ is the speech signal at a sample index n .

In this section, the Arabic phonemes are divided into voiced/unvoiced fricatives, stops, nasals, laterals, and trills. The zero crossings and zero crossings rates are calculated for the test signal using different speakers. Fig. 1 shows ZCR rates distribution for different Arabic phonemes while, Fig. 2 represents a comparison of the zero crossing rates distribution for the three speakers. The results are also tabulated for the different phonemes as shown in Table 1 and 2. Discussion of these findings is presented in the final section of the paper.

Low energy phonemes

In this section, the proposed algorithm for low energy detection is implemented. Voiceless phonemes in the Arabic language are considered low energy, for instance, the sounds / س / (/s/) and /ت / (/t/) belongs to this category. In this part of the paper, this particular class was chosen to test the automatic segmentation method. Consequently as an example of testing the proposed procedure for boundary limits, the low energy regions detection algorithm is used for the shortest voiceless phonemes in the Arabic language which has duration between 12*ms* and 24*ms* depending on the type of speaker. The phoneme /ر / (/r/) is selected for this purpose, using an analysis window length of 10*ms* and an overlap of 3*ms*. The resulting segmented signal with its corresponding energy is calculated by the algorithm and plotted in Fig. 3. A comparisons of the results obtained using the automatic and manual segmentations are shown in Table 3.

High energy phonemes

In the previous section, the segmentation algorithm was implemented for low energy regions detection. In this part of the paper, the same procedure is modified to perform the segmentation for high energy regions using the same Arabic speech corpus. The high energy regions in the Arabic language include for instance the long and short vowels: the /a/ "*Fat'hah*", /i/ "*Kasrah*", and /u/ "*Dhammah*". In this paper, the long vowel /u:/*"long*

Fig. 1. Plot of the zero crossing rates (ZCR) distribution for different Arabic phonemes.

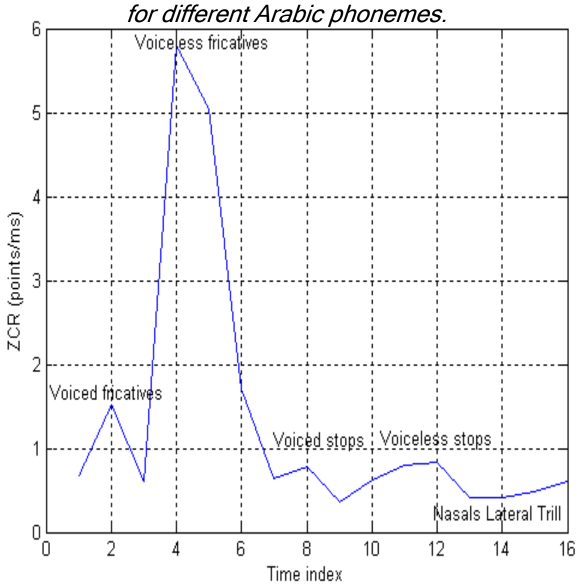

Fig. 2. Comparison of the zero crossing rates (ZCR) distribution for 5 speakers.

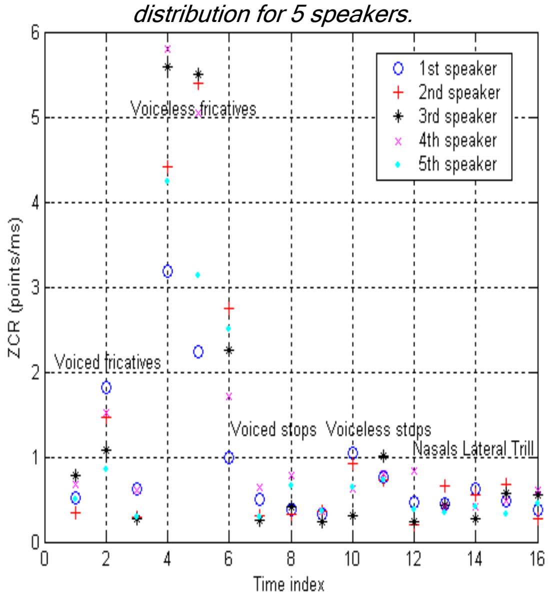

Fig. 3. Time series plot showing boundary limits for the phoneme / ر / (/r/) using low energy regions detection.

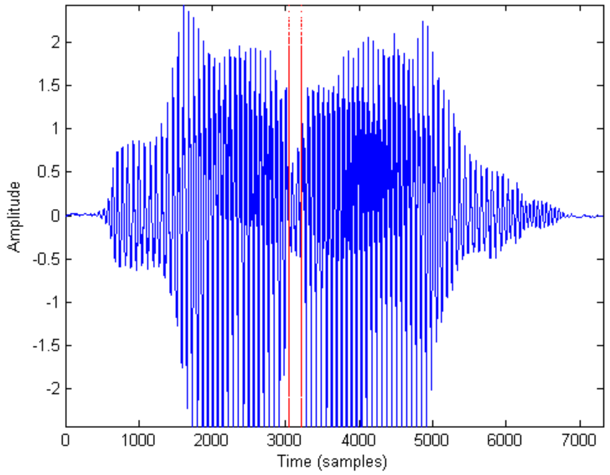

Fig. 4. Spectrogram plot for the phoneme / ر / (/r/) in fig. 3.

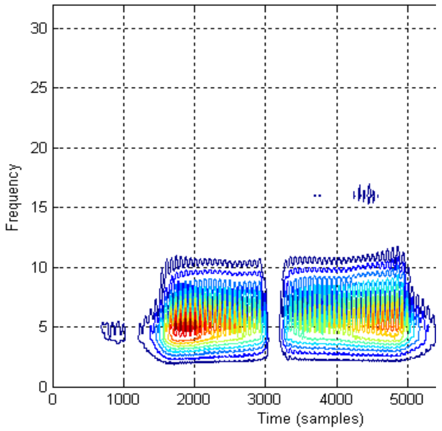

Fig. 5. Time series plot for the phoneme /u:/ "long dhammah" showing the boundary limits using high energy region detection.

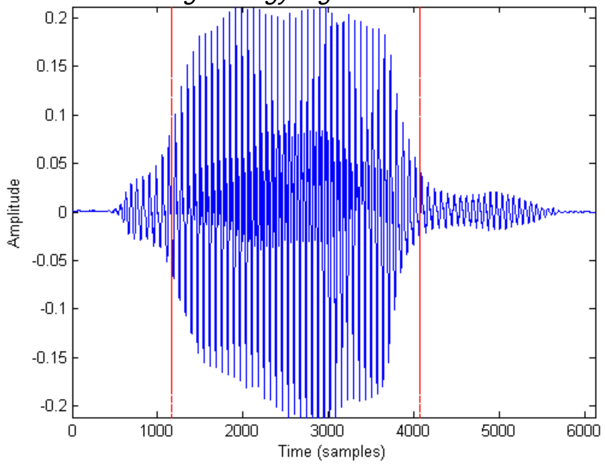

Fig. 6. Spectrogram plot for the phoneme /u:/ "long dhammah" in fig. 5.

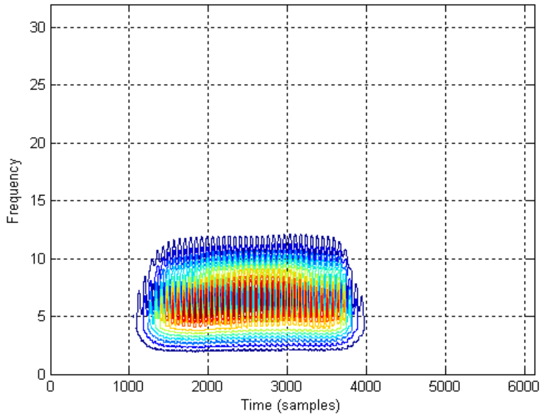

Table 1. The zero crossing (ZCR) and ZCR rate for voiced/voiceless fricatives & stops Arabic phonemes.

Phoneme type	Token	ZCR (points)	ZCR rate (ms)
Voiced fricatives	ززز زعز زذز	23 138 42	0.509978 1.81102 0.630631
Voiceless fricatives	زسز زشز زحز	295 226 78	3.17546 2.23762 0.984848
Voiced stops	زبز زضز زدز	35 22 22	0.497866 0.369128 0.319767
Voiceless stops	زتز زكز زقز	70 71 40	1.05263 0.771739 0.464037

dhammah”, is taken as an example for this particular category. The chosen window length is 12*ms* with an overlap of 2 *ms*. The Automatic segmentation results are presented in Fig. 4, Fig. 5 and Table 4 with Fig. 6 showing the spectrogram plot of the time series signal for results comparisons.

Table 2. The zero crossing (ZCR) & ZCR rate for nasals, lateral & trill Arabic phonemes.

Phoneme type	Token	ZCR (points)	ZCR rate (ms)
Nasals	زنز زمز	20 26	0.44150 0.62954
Lateral	زلز	9	0.47872
Trill	زرز	8	0.379147

Results and discussion

The first part of the paper presented zero crossing rate (ZCR) procedure which was estimated for the major speech signals in the Arabic language. The phonemes were divided into four groups; voiced fricative, voiced stops, unvoiced fricatives and unvoiced stops. It was found that the number of ZCR for the unvoiced speech was much higher than that for the voiced one as expected, especially for the case of unvoiced fricatives which produced a relatively much higher ZCR rates than the others. Fig. 1 depicts the results of the zero crossing rates (ZCR) distribution for different Arabic phonemes. As can be seen from the plot, the voiceless fricatives have the highest ZCR rate while nasals, laterals, and trills are the lowest. These results are also presented in Table 1. In addition, the same procedure was repeated for five different speakers using all Arabic phonemes categories. The zero crossing rates for this group were very close to each other. The findings are depicted in Fig. 2 which provides a comparison of ZCR rates distribution for different speakers. As can be seen from the figure, the ZCR values for the five speakers are very similar in terms of the distribution for the chosen phonemes.

The tested algorithm performed reasonably well when compared with the manual approach for the Arabic speech segmentation. In Table 3, the deviation between the automatic and manual is 8*ms* for the onset while this value is only 6*ms* for the offset part. This result was selected as an example since it represents the shortest phonemes in the Arabic language, which is only in the range of 18*ms* in length. Furthermore, the plots in Fig. 3 and Fig.4 of the segmented speech signal show an accurate detection of boundary limits for the target phoneme.

On the other hand, the proposed algorithm performance was not as robust for high energy region detection as it was for the low energy case. Some problems were encountered when the highest energy point was very close to the start or end of the tested signal. In this situation the calculation procedure needed some adjustments. Unlike the low energy method, the values of the window and overlap must be selected very precisely to avoid any calculation error which is another disadvantage. These major draw backs mean that the algorithm must be supervised in order to perform satisfactorily especially when used for high energy phonemes. Fig. 5 and Fig.6 represent an accurate detection of the boundary limits for the phoneme “*long damah”.*

Table 3. Comparison of automatic and manual results obtained for the target phoneme / ر / (/r/).

Token	Auto-segmentation		Manual-segmentation		Phoneme length
زَرَزْ zaraz	Onset (ms)	Offset (ms)	Onset (ms)	Offset (ms)	(Auto-Seg.) (ms)
	304	322	296	328	18

Table 4. Comparison of automatic & manual results for the target phoneme “long dhammah”.

Token	Auto-segmentation		Manual-segmentation		Phoneme length
/u:/ Long Dhammah	Onset (ms)	Offse t (ms)	Onset (ms)	Offset (ms)	(Auto-Seg.) (ms)
	116	407	122	378	291

Conclusions

In this paper, the zero crossing rates (ZCR) for Arabic speech were calculated using equation (4) where the sampled speech signal is represented as $x(n)$. The tested phonemes were divided into voiced/unvoiced fricatives, stops, nasals, laterals, and trills before the required result $Z(k)$ was found. The procedure was repeated for five different speakers. When the distributions of the obtained values were compared to each other, the results were very similar as shown in Table 1 and 2. Moreover, an automatic segmentation method based on dividing the Arabic speech into low and high energy regions corresponding to voiced and un-voiced phonemes was implemented. This approach needs some prior knowledge of the tested segment since an appropriate window and overlap must be selected for different utterances. In general, the algorithm was found to be more accurate and robust for the low energy regions detection than for the high energy case. In order to carry out the experiment, the Arabic Phonetic Database (KAPD) built by the Computer and Electronics Research Institute at KACST was utilized.

Acknowledgment

The authors would like to thank the computer and research institute (CERI) for providing the Arabic corpus and KACST for its financial support.

References

1. Alghmadi M (2003) KACST Arabic phonetic database. The 15th Int. Congress of Phonetics Sci. 3109-3112.
2. Cherif A, Bouafif L and Dabbabi T (2001) Pitch detection and formant analysis of Arabic speech processing. *Appl. Acoustics.* 1129-1140.
3. Demuynck K and Laureys T (2002) A comparison of different approaches to automatic speech segmentation. *Lecturer notes in Computer Sci.* 2448, 385-406.
4. Essa O (2005) Using prosody in automatic segmentation of speech. Computer Science Dept., University of South Carolina.
5. Hemert V (1991) Automatic segmentation of speech. IEEE Trans. on Signal Processing. 1008-1012.
6. Lee Y, Papineni K, Roukos S, Emam O and Hassan H (2003) Language model based arabic word segmentation. Proc. of the 41st Annual meeting of the Association for Computational Linguistics. 399-406.
7. Tolba M, Nazmy T, Abdelhamid A and Gadallah M (2005) A novel method for Arabic consonant/vowel segmentation using wavelet transform. Proc. of IJICIS. 353-364.