Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches

Md. Mijanur Rahman
Dept. of Computer Science & Engineering,
Jatiya Kabi Kazi Nazrul Islam University
Trishal, Mymensingh, Bangladesh.

Md. Al-Amin Bhuiyan
Dept. of Computer Science & Engineering,
Jahangir nagar University
Savar, Dhaka, Bangladesh.

***Abstract*—This paper presents simple and novel feature extraction approaches for segmenting continuous Bangla speech sentences into words/sub-words. These methods are based on two simple speech features, namely the time-domain features and the frequency-domain features. The time-domain features, such as short-time signal energy, short-time average zero crossing rate and the frequency-domain features, such as spectral centroid and spectral flux features are extracted in this research work. After the feature sequences are extracted, a simple dynamic thresholding criterion is applied in order to detect the word boundaries and label the entire speech sentence into a sequence of words/sub-words. All the algorithms used in this research are implemented in Matlab and the implemented automatic speech segmentation system achieved segmentation accuracy of 96%.**

Keywords-Speech Segmentation; Features Extraction; Short-time Energy; Spectral Centroid; Dynamic Thresholding.

I. INTRODUCTION

Automated segmentation of speech signals has been under research for over 30 years [1]. Speech Recognization system requires segmentation of Speech waveform into fundamental acoustic units [2]. Segmentation is a process of decomposing the speech signal into smaller units. Segmentation is the very basic step in any voiced activated systems like speech recognition system and speech synthesis system. Speech segmentation was done using wavelet [3], fuzzy methods [4], artificial neural networks [5] and Hidden Markov Model [6]. But it was found that results still do not meet expectations. In order to have results more accurate, groups of several features were used [7, 8, 9 and 10]. This paper is continuation of feature extraction for speech segmentation research. The method implemented here is a very simple example of how the detection of speech segments can be achieved.

This paper is organized as follows: Section 2 describes techniques for segmentation of the speech signal. In Section 3, we will describe different short-term speech features. In section 4, the methodological steps of the proposed system will be discussed. Section 5 and 6 will describe the experimental results and conclusion, respectively.

II. SPEECH SEGMENTATION

Automatic speech segmentation is a necessary step that used in Speech Recognition and Synthesis systems. Speech segmentation is breaking continuous streams of sound into some basic units like words, phonemes or syllables that can be recognized. The general idea of segmentation can be described as dividing something continuous into discrete, non-overlapping entities [11]. Segmentation can be also used to distinguish different types of audio signals from large amounts of audio data, often referred to as *audio classification* [12].

Automatic speech segmentation methods can be classified in many ways, but one very common classification is the division to *blind* and *aided segmentation algorithms*. A central difference between aided and blind methods is in how much the segmentation algorithm uses previously obtained data or external knowledge to process the expected speech. We will discuss about these two approaches in the following sub-sections.

A. Blind segmentation

The term *blind* segmentation refers to methods where there is no pre-existing or external knowledge regarding linguistic properties, such as orthography or the full phonetic annotation, of the signal to be segmented. Blind segmentation is applied in different applications, such as speaker verification systems, speech recognition systems, language identification systems, and speech corpus segmentation and labeling [13].

Due to the lack of external or top-down information, the first phase of blind segmentation relies entirely on the acoustical features present in the signal. The second phase or bottom-up processing is usually built on a front-end parametrization of the speech signal, often using MFCC, LP-coefficients, or pure FFT spectrum [14].

B. Aided segmentation

Aided segmentation algorithms use some sort of external linguistic knowledge of the speech stream to segment it into corresponding segments of the desired type. An orthographic or phonetic transcription is used as a parallel input with the speech, or training the algorithm with such data [15]. One of the most common methods in ASR for utilizing phonetic annotations is with HMM-based systems [16]. HMM-based algorithms have dominated most speech recognition applications since the 1980's due to their so far superior performance in recognition and relatively small computational complexity in the field of speech recognition [17].

III. FEATURE EXTRACTION FOR SPEECH SEGMENTATION

The segmentation method described here is a purely bottom-up blind speech segmentation algorithm. The general principle of the algorithm is to track the amplitude or spectral changes in the signal by using short-time energy or spectral

features and to detect the segment boundaries at the locations where amplitude or spectral changes exceed a minimum threshold level. Two types of features are used for segmenting speech signal: *time-domain signal features* and *frequency-domain signal features*.

A. Time-Domain Signal Features

Time-domain features are widely used for speech segment extraction. These features are useful when it is needed to have algorithm with simple implementation and efficient calculation. The most used features are short-time energy and short-term average zero-crossing rate.

1) Short-Time Signal Energy

Short-term energy is the principal and most natural feature that has been used. Physically, energy is a measure of how much signal there is at any one time. Energy is used to discover voiced sounds, which have higher energy than silence/un-voiced, in a continuous speech, as shown in Figure-1.

The energy of a signal is typically calculated on a short-time basis, by windowing the signal at a particular time, squaring the samples and taking the average [18]. The square root of this result is the engineering quantity, known as the root-mean square (RMS) value, also used. The short-time energy function of a speech frame with length *N* is defined as

$$E_n = \frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2 \quad \ldots\ldots\ldots\ldots\ldots (1)$$

The short-term root mean squared (RMS) energy of this frame is given by:

$$E_{n(RMS)} = \sqrt{\frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2} \quad \ldots\ldots\ldots (2)$$

Where $x(m)$ is the discrete-time audio signal and $w(m)$ is rectangle window function, given by the following equation:

$$w(m) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

2) Short-Time Average Zero-Crossing Rate

The average zero-crossing rate refers to the number of times speech samples change algebraic sign in a given frame. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. It is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero [19]. Unvoiced speech components normally have much higher ZCR values than voiced ones, as shown in Figure 2. The short-time average zero-crossing rate is defined as

$$Z_n = \frac{1}{2}\sum_{m=1}^{N}|\,\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]\,|\,w(n-m) \quad \cdots (4)$$

$$\text{Where, } \mathrm{sgn}[x(m)] = \begin{cases} 1 & x(m) \ge 0 \\ -1 & x(m) < 0 \end{cases} \quad \ldots\ldots (5)$$

and*w(n)* is a rectangle window of length N, given in equation (3).

B. Frequency-Domain Signal Features

The most information of speech is concentrated in 250Hz – 6800Hz frequency range [20]. In order to extract frequency-domain features, discrete Fourier transform (that provides information about how much of each frequency is present in a signal) can be used. The Fourier representation of a signal shows the spectral composition of the signal.Widely used frequency-domain features are *spectral centroid* and *spectral flux* feature sequences that used discrete Fourier transform.

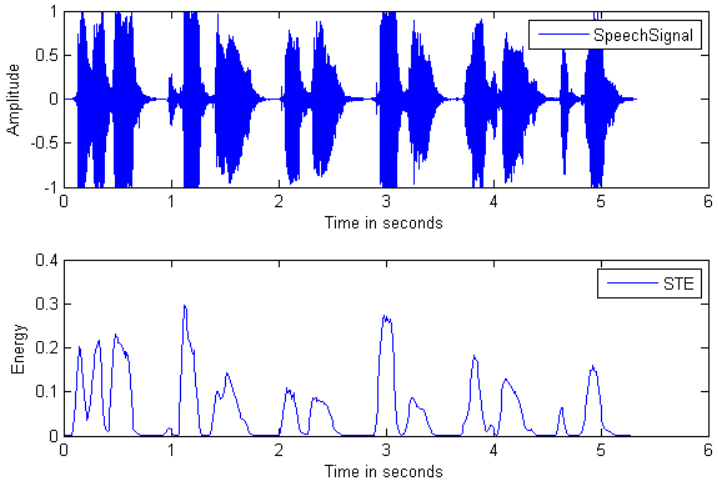

Figure 1. Original signal and short-time energy curves of the speech sentence, "আমাদেরজাতীয়কবিকাজীনজরুলইসলাম".

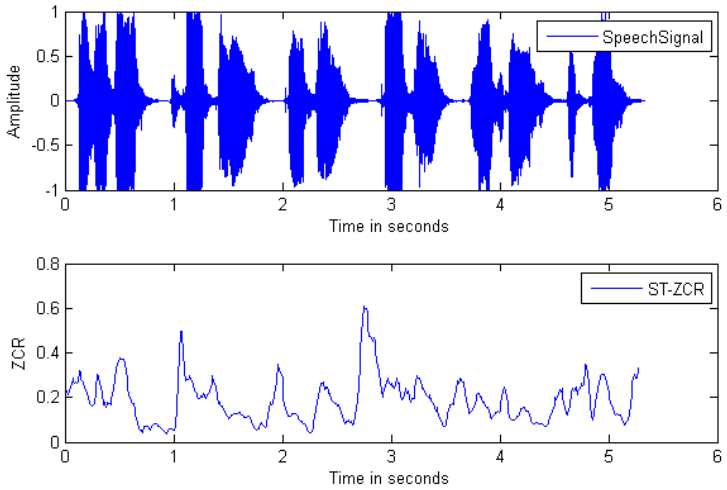

Figure 2. Original signal and short-term average zero-crossing rate curves of speech sentence, **"আমাদেরজাতীয়কবিকাজীনজরুলইসলাম".**

1) Spectral Centroid

The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of gravity" of the spectrum is. This feature is a measure of the spectral position, with high values corresponding to "brighter" sounds [21], as shown in Figure-3. The spectral centroid, SC_i, of the *i*-th frame is defined as the center of "gravity" of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m) X_i(m)}{\sum_{m=0}^{N-1} X_i(m)} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Here, f*(m)* represents the center frequency of *i*-th bin with length *N* and $X_i(m)$ is the amplitude corresponding to that bin in DFT spectrum. The DFT is given by the following equation and can be computed efficiently using a fast Fourier transform (FFT) algorithm [22].

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j 2\pi k \frac{n}{N}} \; ; k = 0, \ldots, N-1 \ldots\ldots\ldots (7)$$

2) Spectral flux

Spectral flux is a measure of how quickly the power spectrum of a signal is changing (as shown in Figure 4), calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame, also known as the Euclidean distance between the two normalized spectra. The spectral flux can be used to determine the timbre of an audio signal, or in onset detection [23], among other things. The equation of Spectral Flux, SF_i is given by:

$$SF_i = \sum_{k=1}^{N/2} (|X_i(k)| - |X_i(k-1)|)^2 \quad \ldots\ldots\ldots (8)$$

Here, $X_i(k)$ is the DFT coefficients of *i*-th short-term frame with length *N*, given in equation (7).

C. Speech Segments Detection

After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments. The following steps are included in this thresholding algorithm.

1. Get the feature sequences from the previous feature extraction module.
2. Apply median filtering in the feature sequences.
3. Compute the *Mean* or average values of smoothed feature sequences.
4. Compute the histogram of the smoothed feature sequences.
5. Find the local maxima of histogram.
6. If at least two maxima M_1 and M_2 have been found, Then:

$$\text{Threshold, } T = \frac{W * M_1 + M_2}{W + 1} \quad \ldots\ldots\ldots\ldots\ldots (9)$$

Otherwise,

$$\text{Threshold, } T = \frac{Mean}{2} \quad \ldots\ldots\ldots\ldots (10)$$

where W is a user-defined weight parameter [24]. Large values of *W* obviously lead to threshold values closer to M_1.Here, we used *W*=100.

The above process is applied for both feature sequences and finding two thresholds: T_1 based on the energy sequences and T_2 on the spectral centroid sequences. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds (for both feature sequences). Figure-5 shows both filtered feature sequences curves with threshold values.

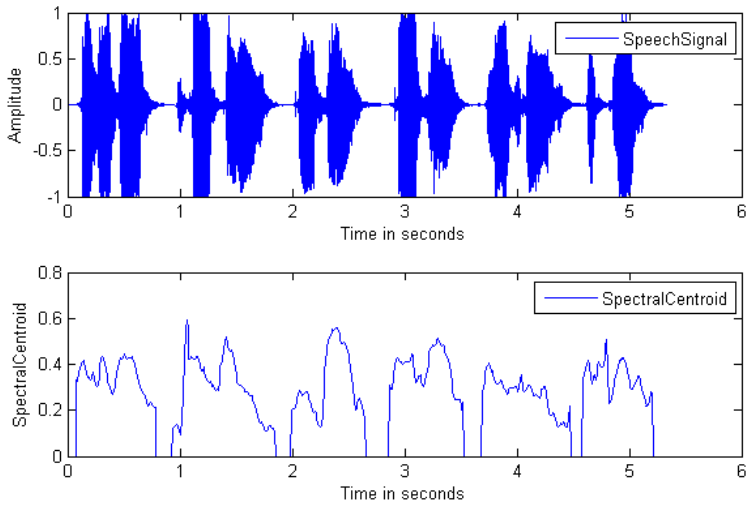

Figure 3. The graph of original signal and spectral centroid features of speech sentence, "আমাদেরজাতীয়কবিকাজীনজরুলইসলাম".

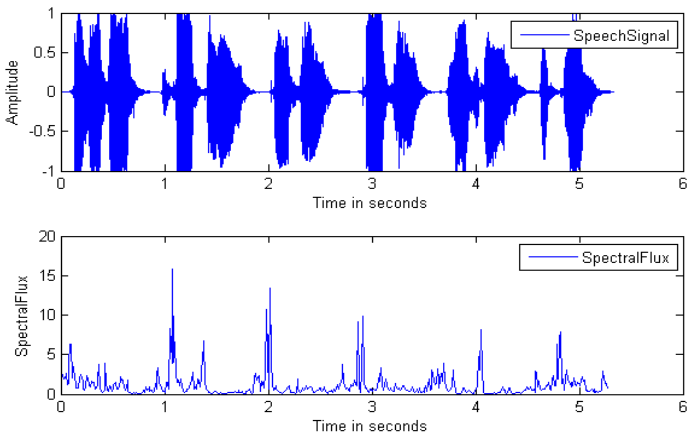

Figure 4. The curves of original signal and spectral flux features of speech sentence, "আমাদেরজাতীয়কবিকাজীনজরুলইসলাম".

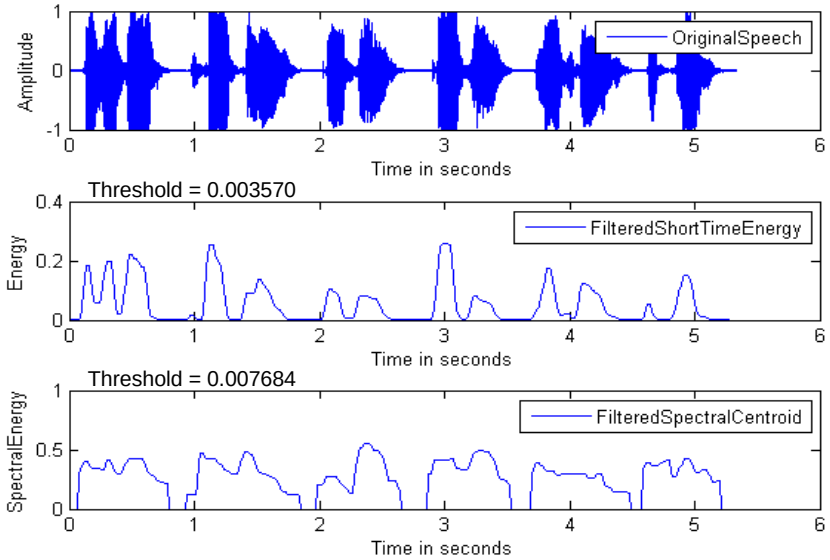

Figure 5. Original speech signal and median filtered feature sequences curves with threshold values of a speech sentence "আমাদেরজাতীয়কবিকাজীনজরুলইসলাম".

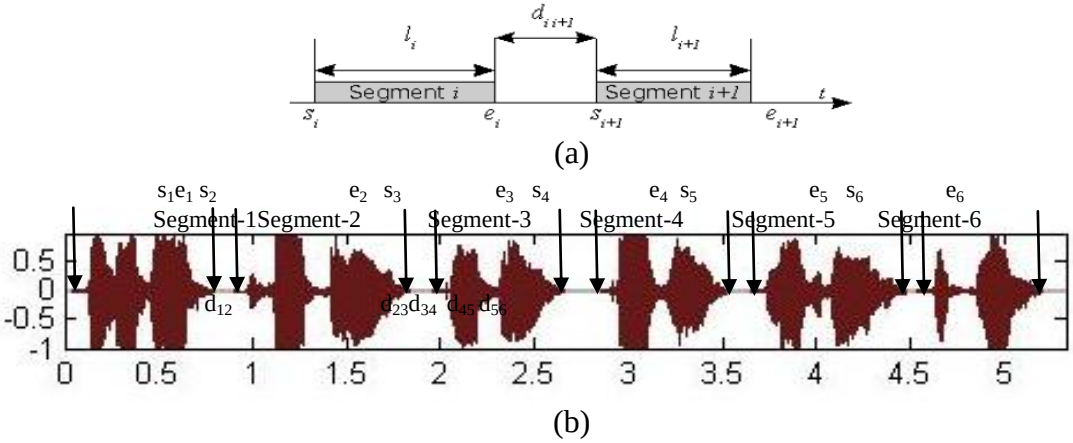

Figure 6. Speech Segments:(a) Segment, where l_i is the length of i segment, $d_{i\,i+1}$ is the distance between two segments and s_i is start time of i segment, e_i is the end time of i segment. (b) Detection of start and end point of each segment in a speech sentence.

D. Post Processing of Detected Segments

As shown in Figure 6, detected speech segments are analyzed in post-processing stage. Common segmentation errors are: short segments usually are noises/silences, and two segments with short space in between can be the same segment.

Post-processing with rule base can fix these and similar mistakes. Waheed [25] proposed to use 2 rules:

1. If $l_i < \min Length$ and $d_{ii+1} > \min Space$, then the segment i is discarded, similarly if $l_{i+1} < \min Length$ and $d_{ii+1} > \min Space$, then segment i+1 is discarded.
2. If $l_i \, or \, l_{i+1} > \min Length$ and $d_{ii+1} > \min Space$ and $l_i + l_{i+1} > FL$, then two segments are merged and anything between the two segments that was previously left, is made part of the speech.

IV. Implementation

The automatic speech segmentation system has been implemented in Windows environment and we have used MATLAB Tool Kit for developing this application. The proposed speech segmentation system has six major steps, as shown in Figure 7.

A. Speech Acquisition
B. Signal Preprocessing
C. Speech Features Extraction
D. Histogram Computation
E. Dynamic thresholding and
F. Post-Processing

A. Speech Acquisition

Speech acquisition is acquiring of continuous Bangla speech sentences through the microphone.

Speech capturing or speech recording is the first step of implementation. Recording has been done by native male speaker of Bangali. The sampling frequency is 16 KHz; sample size is 8 bits, and mono channels are used.

B. Signal Preprocessing

This step includes elimination of background noise, framing and windowing. Background noise is removed from the data so that only speech samples are the input to the further processing. Continuous speech signal has been separated into a number of segments called frames, also known as framing. After the pre-emphasis, filtered samples have been converted into frames, having frame size of 50 msec. Each frame overlaps by 10 msec. To reduce the edge effect of each frame segment windowing is done. The window, *w(n)*, determines the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. Rectangular window has been used.

C. Short-term Feature Extraction

After windowing, we have been computed the short-term energy features and spectral centroid features of each frame of the speech signal. These features have been discussed in details in Section 3. In this step, median filtering of these feature sequences also computed.

D. Histogram Computation

Histograms of both smoothed feature sequences are computed in order to find the local maxima of the histogram, from which the threshold values are calculated.

E. Dynamic Thresholding

Dynamic thresholding is applied for both feature sequences and finding two thresholds: T_1 and T_2, based on the energy sequences and the spectral centroid sequences respectively.

After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds.

F. Post-Processing

In order to segment words/sub-words, the detected speech segments are lengthened by 5 short term windows (each window of 50 msec), on both sides in the post-processing step. Two segments with short space in between have been merged to get final speech segments. These segmented speech words are saved as *.wav file format for further use.

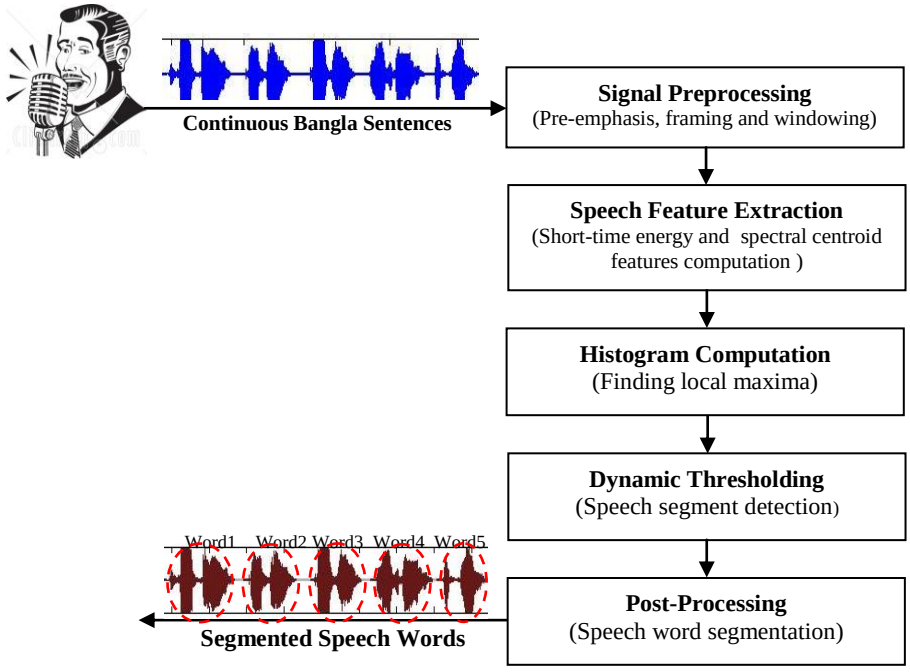

Figure 7. Block diagram of the proposed Automatic Speech Segmentation system.

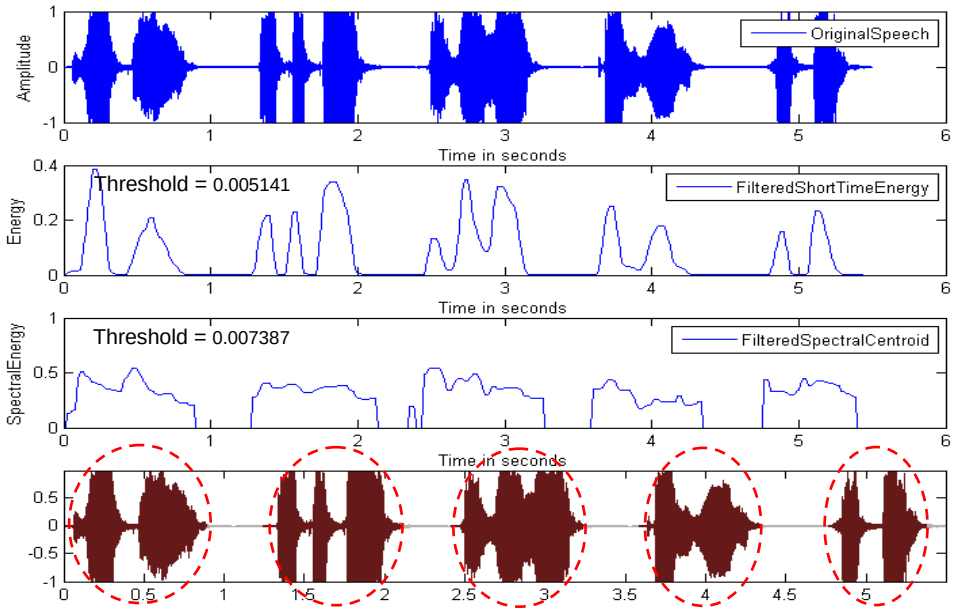

Figure-8. The segmentation results for a speech sentences "জাতীয়পতাকারডিজাইনারকামরুলহাসান" which contains 5 (five) speech words. The first subfigure shows the original signal. The second subfigure shows the sequences of the signal's energy. In the third subfigure the spectral centroid sequence is presented. In both cases, the respective thresholds are also shown. The final subfigure presents the segmented words in dashed circles.

TABLE 1. SEGMENTATION RESULTS

Sentence ID	No. of word segments expected	No. of words properly segmented by system	Segmentation rate (%)	
			Success rate	Failure rate
S1	6	6	100	0
S2	10	10	100	0
S3	9	9	100	0
S4	5	4	80	20
S5	10	10	100	0
S6	7	7	100	0
S7	8	8	100	0
S8	11	10	90.9	9.1
S9	9	8	88.89	11.11
S10	5	5	100	0
Total	**80**	**77**	**96.25**	**3.75**

V. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed system different experiments were carried out. All the techniques and algorithms discussed in this paper have been implemented in Matlab 7.12.0 version. In this experiment, various speech sentences in Bangla language have been recorded, analyzed and segmented by using time-domain and frequency-domain features with dynamic thresholding technique. Figure 8 shows the filtered short-time energy and spectral centroid features of the Bangla speech sentence "জাতীয়পতাকারডিজাইনারকামরুলহাসান", where the boundaries of words are marked automatically by the system. Table-1 shows the details segmentation results for ten speech sentences and reveals that the average segmentation accuracy rate is 96.25%, and it is quite satisfactory.

VI. CONCLUSION AND FURTHER RESEARCH

We have presented a simple speech features extraction approach for segmenting continuous speech into word/sub-words in a simple and efficient way. The short-term speech features have been selected for several reasons. *First*, it provides a basis for distinguishing voiced speech components from unvoiced speech components, i.e., if the level of background noise is not very high, the energy of the voiced segments is larger than the energy of the silent or unvoiced segments. *Second*, if unvoiced segments simply contain environmental sounds, then the spectral centroid for the voiced segments is again larger. *Third*, its change pattern over the time may reveal the rhythm and periodicity nature of the underlying sound.

From the experiments, it was observed that some of the words were not segmented properly. This is due to different causes: (i) the utterance of words and sub-words differs depending on their position in the sentence, (ii) the pauses between the words or sub-words are not identical in all cases because of the variability of the speech signals and (ii) the non-uniform articulation of speech. Also, the speech signal is very much sensitive to the speaker's properties such as age, sex, and emotion. The proposed approach shows good results in speech segmentation that achieves about 96% of segmentation accuracy. This reduces the memory requirement and computational time in any speech recognition system.

The major goal of future research is to search for possible mechanisms that can be employed to enable top-down feedback and ultimately pattern discovery by learning. To design more reliable system, future systems should employ knowledge (syntactic or semantic) of linguistics and more powerful recognition approaches like Gaussion Mixture Models (GMMs), Time-Delay Neural Networks (TDNNs), Hidden Markov Model (HMM), Fuzzy logic, and so on.

REFERENCES

[1] OkkoRasanen, "*Speech Segmentation and Clustering Methods for a New Speech Recognition Architecture*", M.Sc Thesis, Department of Electrical and Communications Engineering, Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Espoo, November 2007.

[2] R. Thangarajan, A. M. Natarajan, M. Selvam , "*Syllable modeling in continuous speech recognition for Tamil language*", International Journal of Speech Technology , vol. 12, no. 1, pp. 47-57, 2009.

[3] Hioka Y and Namada N, "*Voice activity detection with array signal processing in the wavelet domain*", IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, 86(11):2802-2811, 2003.

[4] Beritelli F and Casale S, "*Robust voiced/unvoiced classification using fuzzy rules*", In 1997 IEEE workshop on speech coding for telecommunications proceeding, pages5-6, 1997.

[5] Qi Y and Hunt B, "*Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier*", IEEE Transactions on Speech and Audio Processing, I(2):250-255, 1993.

[6] Basu S, "*A linked-HMM model for robust voicing and speech detection*", In IEEE international conference on acoustics, speech and signal processing (ICAASSP'03), 2003.

[7] Atal B and Rabiner L, "*A pattern recognition approach to voice-unvoiced-silence classification with applications to speech recognition*", IEEEASSP, ASSP-24(3):201-212, June 1976.

[8] Siegel L and Bessey A, "*Voiced/unvoiced/mixed excitation classification of speech*", IEEE Transactions on Acoustics, Speech and Signal Processing, 30(3):451-460, 1982.

[9] Shin W, Lee B, Lee Y and Lee J, "*Speech/Non-speech classification using multiple features for robust endpoint detection*", In 2000 IEEE International Conference on Acoustics, Speech and Signal Processing,

ICASSP'00 Proceedings, Vol.3, 2000.

[10] Kida Y and Kawahara T, "*Voice activity detection based on optimally weighted combination of multiple features*", In 9th European Conference on Speech Communication and Technology, ISCA, 2005.

[11] Kvale K, "*Segmentation and Labeling of Speech",* PhD Dissertation, The Norwegian Institute of Technology, 1993.

[12] Antal M, "*Speaker Independent Phoneme Classification in Continuous Speech*",Studia Univ. Babes-Bolyal, Informatica, Vol. 49, No. 2, 2004.

[13] Sharma M and Mammone R, "*Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge*", Spoken Language, 1996. ICSLP 96. Proceedings. Vol. 2, pp. 1237-1240, 1996.

[14] Sai Jayram A K V, Ramasubramanian V and Sreenivas T V, "*Robust parameters for automatic segmentation of speech*", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02), Vol. 1, pp. 513-516, 2002.

[15] Schiel F, "*Automatic Phonetic Transcription of Non-Prompted Speech*",Proceedings of the ICPhS 1999. San Francisco, August 1999. pp. 607-610, 1999.

[16] Knill K and and Young S, "*Hidden Markov Models in Speech and Language Processing*", Kluwer Academic Publishers, pp. 27-68, 1997.

[17] Juang B H and Rabiner L R, "*Automatic Speech Recognition – A Brief History of The Technology Development*", Elsevier Encyclopedia of Language and Linguistics, Second Edition, 2005.

[18] Tong Zhang and Jay C CKuo, "*Hierarchical classification of audio data for acrhiving and retrieving*", In International Conference on Acoustics, Speech and Signal Processing, volume VI, pages 3001–3004. IEEE, 1999.

[19] L R Rabiner and M R Sambur, "*An Algorithm for determining the endpoints of Isolated Utterances*", The Bell System Technical Journal, February 1975, pp 298-315.

[20] Niederjohn R and Grotelueschen J, "*The enhancement of speech intelligibility in high noise level by high-pass filtering followed by rapid amplitude compression*", IEEE Transactions on Acoustics, Speech and Signal Processing, 24(4), pp277-282, 1976.

[21] T Giannakopoulos, "*Study and application of acoustic information for the detection of harmful content and fusion with visual information*" Ph.D. dissertation, Dept. of Informatics and Telecommunications, University of Athens, Greece, 2009.

[22] Cooley, James W. and Tukey, John W. "*An algorithm for the machine calculation of complex Fourier series*", Mathematics of Computation: Journal Review, 19: 297–301, 1965.

[23] Bello J P, Daudet L, Abdallah S, Duxbury C, Davies M, and Sandler MB, "*A Tutorial on Onset Detection in Music Signals*", IEEE Transactions on Speech and Audio Processing 13(5), pp 1035–1047, 2005.

[24] T Giannakopoulos, A Pikrakis and S. Theodoridis "*A Novel Efficient Approach for Audio Segmentation*", Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008), December 8-11 2008, Tampa, Florida, USA.

[25] Waheed K, Weaver K and Salam F, "*A robust algorithm for detecting speech segments using an entropic contrast*", In Proc. of the IEEE Midwest Symposium on Circuits and Systems, Vol.45, Lida Ray Technologies Inc., 2002.

AUTHORS PROFILE

Md. MijanurRahman

Mr. Md. Mijanur Rahman is working as Assistant Professor of the Dept. of Computer Science and Engineering at Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymen singh, Bangladesh. Mr. Rahman obtained his B. Sc. (Hons) and M. Sc degrees, both with first class first in Computer Science and Engineering from Islamic University, Kushtia, Bangladesh. Now he is a PhD researcher of the department of Computer Science and Engineering at Jahangir nagar University, Savar, Dhaka, Bangladesh.

His teaching and research interest lies in the areas such as Digital Speech Processing, Pattern Recognition, Database management System, Artificial Intelligence, etc. He has got many research articles published in both national and international journals.

Dr. Md. Al-Amin Bhuiyan

Dr. Md. Al-Amin Bhuiyan is serving as a professor of the Dept. of Computer Science and Engineering. Dr. Bhuiyan obtained his B. Sc. (Hons) and M. Sc degrees both with first class in Applied Physics & Electronics from University of Dhaka, Bangladesh. He completed his PhD study in Information & Communication Engineering from Osaka City University, Japan.

His teaching and research interest lies in the areas such as Image Processing, Computer Vision, Computer Graphics, Pattern Recognition, Soft Computing, Artificial Intelligence, Robotics, etc. He has got many articles published in both national and international journals.