Blocking Black Area Method for Speech Segmentation

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

Fatema Khatun
Dept. of Electrical & Electronic Engineering
Hamdard University Bangladesh
Sonargoan, Narayanganj, Bangladesh

Dr. Md. Al-Amin Bhuiyan
Dept. of Computer Engineering
King Faisal University
Al Ahssa 31982, Saudi Arabia

***Abstract*—Speech segmentation is an important sub problem of automatic speech recognition. This research is concerned with the development of a continuous speech segmentation system using Bangla Language. This paper presents a dynamic thresholding algorithm to segment the continuous Bngla speech sentences into words/sub-words. The research uses Otsu's method for dynamic thresholding and introduces a new approach, named blocking black area method to identify the voiced regions of the continuous speech in speech segmentation. The developed system has been justified with continuously spoken several Bangla sentences. To test the performance of the system, 100 Bangla sentences have been recorded from 5 (five) male speakers of different ages and 656 words have been presented in the 100 Bangla sentences. So, the speech database contains 500 Bangla sentences with 3280 words. All the algorithms and methods used in this research are implemented in MATLAB and the proposed system has been achieved the average segmentation accuracy of 90.58%.**

Keywords—Blocking Black Area; Boundary Detection; Dynamic Thresholding; Otsu's Algorithm; Speech Segmentation

I. INTRODUCTION

Automated Speech Recognition (ASR) is a popular and challenging area of research in developing human computer interactions. The main challenge of speech recognition lies in modeling the variations of the uttered speech, such as different geographical boundaries, social background, age, gender, occupation etc. Automated segmentation of speech signals has been under research for over 30 years [1]. It is a necessity for phonetic analysis of speech [2, 3], audio content classification [4] and many applications in the field of automatic speech recognition (ASR), including word recognition [5, 6]. Speech Recognition system requires segmentation of Speech waveform into fundamental acoustic units [7]. Segmentation is the very basic step in any voiced activated systems like speech recognition system and speech synthesis system. The set of fundamental acoustic units into which the speech waveform can be segmented are words, phonemes or syllables. Word is the preferred and natural unit of speech, because word units have well defined acoustic representation. So, this research chooses word as the basic unit for segmentation. Speech segmentation was done using wavelet [8], fuzzy methods [9], artificial neural networks [10] and Hidden Markov Model [11].

This paper will present the proposed dynamic thresholding algorithms for segmenting continuous Bangla speech sentences into words/sub-words. For speech segmentation, this research introduces a new approach, named *blocking black area method* to properly detect word boundaries in continuous speech segmentation. The paper is organized as follows: Section I describes the introduction of speech processing and the organization of this paper. In Section II, we will discuss about speech segmentation and types of segmentation. Section III will describe thresholding. In Section IV, Otsu's thresholding method will be discussed. Section V will present the blocking black area method. The implementation of the proposed system will be described in Section VI. Sections VII and VIII will describe the experimental results and conclusion, respectively.

II. SPEECH SEGMENTATION

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The general idea of segmentation can be described as dividing something continuous into discrete, non-overlapping entities [12]. In speech segmentation, the basic idea of segmentation is to divide a continuous speech signal into smaller parts, where each of these segments has phonetical or acoustical properties that distinguishes it from neighboring segments. Segmentation can be performed, for example, at the *segment, phone, syllable, word, and sentence* or *dialog turn* level. In isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely: Reliable word recognition is critically dependent on accurate endpoint detection and the computation for processing the speech is less, when the endpoints are accurately located [13]. Automatic speech segmentation methods can be classified in many ways, but one very common classification is the division to blind [14] and aided segmentation algorithms [15]. 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.

III. DYNAMIC THRESHOLDING

In general, thresholding is the simplest method of image segmentation.

This research proposes thresholding techniques on speech segmentation. From a grayscale image, thresholding can be used to convert binary image [16]. In order to convert the image into a binary representation, the technique first converts the image into a grayscale representation and performs a particular threshold analysis process in order to determine which pixels are turned into black or which are white. This research proposes *dynamic thresholding* to convert 256 gray-levels images into monochrome ones. Two important thresholding techniques are fixed or static thresholding and dynamic thresholding. In fixed or static thresholding, the systems usually uses 127 (say) as default threshold value, but you could change this value and obtain darker or lighter images. In dynamic thresholding, the system uses a different threshold value for each pixel of the image. This value is selected automatically, analyzing the sub-image area around each pixel and finding the local contrast. If the contrast of this area is low, the pixel is binarized using a global pre-calculated threshold value, otherwise, when the contrast is high, the local threshold value is calculated and used. In thresholding technique, the output image replaces all pixels in the input image with luminance greater than a threshold with the value of 1 (white) or 0 (black). The problem is how to choose the desired threshold value. Different dynamic thresholding techniques have been used to compute the threshold value [17]. Hence, the research proposes Otsu's thresholding algorithm to compute the desired threshold.

IV. OTSU'S ALGORITHM

Otsu's method is a simple and effective automatic thresholding method, used in image segmentation [18], invented by Nobuyuki Otsu in 1979 [19], also known as binarization algorithm. It is used to automatically perform histogram shape-based image thresholding (i.e. the reduction of a grayscale image into a binary image). The algorithm assumes that the image is composed of two basic classes; such as foreground and background [19]. It then computes an optimal threshold value that minimizes the weighted within class variance; also maximizes the between class variance of these two classes. The algorithmic steps for calculating the threshold is given in Figure-1.

The mathematical formulation of the algorithm for computing the optimum threshold will explain in this section. Let *P(i)* represents the image histogram of speech spectrogram. The two class probabilities $w_1(t)$ and $w_2(t)$ at level *t* are computed by:

$$w_1(t) = \sum_{i=1}^{t} P(i) \tag{1}$$

and
$$w_2(t) = \sum_{i=t+1}^{I} P(i) \tag{2}$$

The class means, $\mu_1(t)$ and $\mu_2(t)$ are:

$$\mu_1(t) = \sum_{i=1}^{t} \frac{iP(i)}{w_1(t)} \tag{3}$$

and
$$\mu_2(t) = \sum_{i=t+1}^{I} \frac{iP(i)}{w_2(t)} \tag{4}$$

Individual class variances:

$$\sigma_1^2(t) = \sum_{i=1}^{t} [i - \mu_1(t)]^2 \frac{P(i)}{w_1(t)} \tag{5}$$

and
$$\sigma_2^2(t) = \sum_{i=t+1}^{I} [i - \mu_2(t)]^2 \frac{P(i)}{w_2(t)} \tag{6}$$

The within class variance (σ_w) is defined as a weighted sum of variances of the two classes and given by:

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{7}$$

Now we will calculate the *between class* variance. The between class variance (σ_b) is defined as a difference of total variance and within class variance and given by:

$$\begin{aligned} \sigma_b^2(t) &= \sigma^2(t) - \sigma_w^2(t) \\ &= w_1(t)[\mu_1(t) - \mu]^2 + w_2(t)[\mu_2(t) - \mu]^2 \\ &= w_1(t)w_2(t)[\mu_1(t) - \mu_2(t)]^2 \end{aligned} \tag{8}$$

$$\mathit{where}\ \mu = w_1(t)\mu_1(t) + w_2(t)\mu_2(t) \tag{9}$$

These two variances $\boldsymbol{\sigma_w}$ and $\boldsymbol{\sigma_b}$ are calculated for all possible thresholds, $t = 0 \ldots I$ (max. intensity). Otsu finds the best threshold that *minimizes the weighted* ***within class variance*** **(σ_w)**, *also maximizes the weighted* ***between class variance*** **(σ_b)**. Finally, the pixel luminance less than or equal to threshold is replaced by 0 (black) and greater than threshold is replaced by 1 (white) to obtain the binary or B/W image.

V. BLOCKING BLACK AREA METHOD

For speech segmentation, this research introduces a new approach, named *blocking black area method*. This method is used to block the voiced regions of the continuous speech, so that we can easily separate the voiced parts of the speech from silence or un-voiced parts in the continuous speech. The edges of the block are used as word boundaries in the continuous speech. The main task of speech segmentation is to detect the boundaries of speech units (i.e., start and end points detection). The algorithm is applied in the thresholded spectrogram image that produces rectangular black boxes in the voiced regions of the speech sentence, as shown in Figure-2. Each black box represents a speech unit (i.e., word or sub-word) of a speech sentence. The method works as follows:

- Summing the column-wise intensity values of thresholded spectrogram image.
- Find the image columns with fewer white pixels based on summing value and replace all pixels on this column with luminance 0 (black).
- Find the image columns with fewer black pixels based on summing value and replace all pixels on this column with luminance 1 (white).

- Detect the boundaries of voiced block and separate the voiced block as speech units.

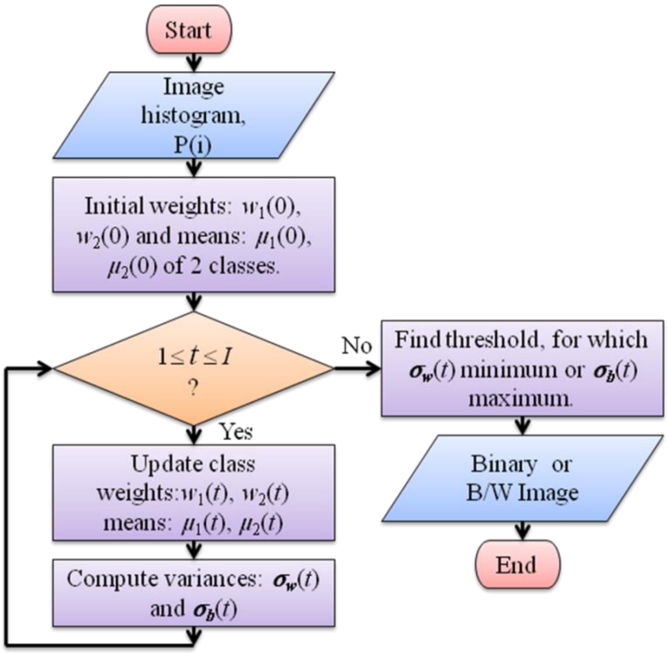

Start
Image histogram, P(i)
Initial weights: $w_1(0)$, $w_2(0)$ and means: $\mu_1(0)$, $\mu_2(0)$ of 2 classes.
$1 \leq t \leq I$?
No
Yes
Find threshold, for which $\sigma_w(t)$ minimum or $\sigma_b(t)$ maximum.
Binary or B/W Image
End
Update class weights: $w_1(t)$, $w_2(t)$ means: $\mu_1(t)$, $\mu_2(t)$
Compute variances: $\sigma_w(t)$ and $\sigma_b(t)$

Fig. 1. Otsu's Thresholding Algorithm

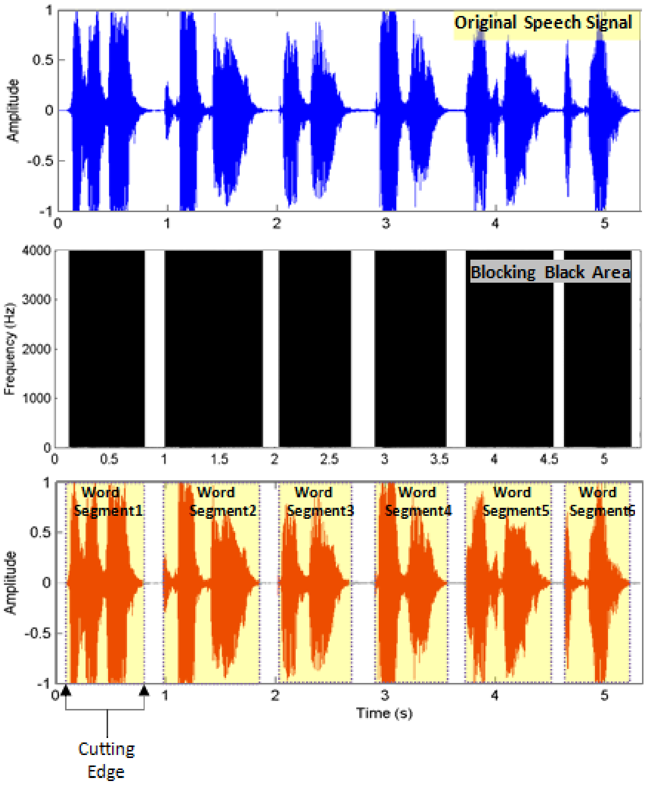


Fig. 2. Blocking the voiced regions by using blocking black area method

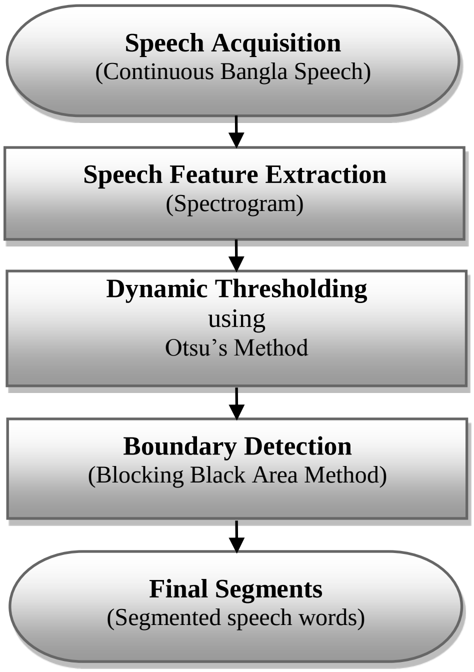


Fig. 3. Proposed Speech Segmentation Procedure

VI. Implementation

The proposed segmentation system, shown in Figure-3, has the following major steps and will discuss in the following sub-sections.

A. Speech Acquisition

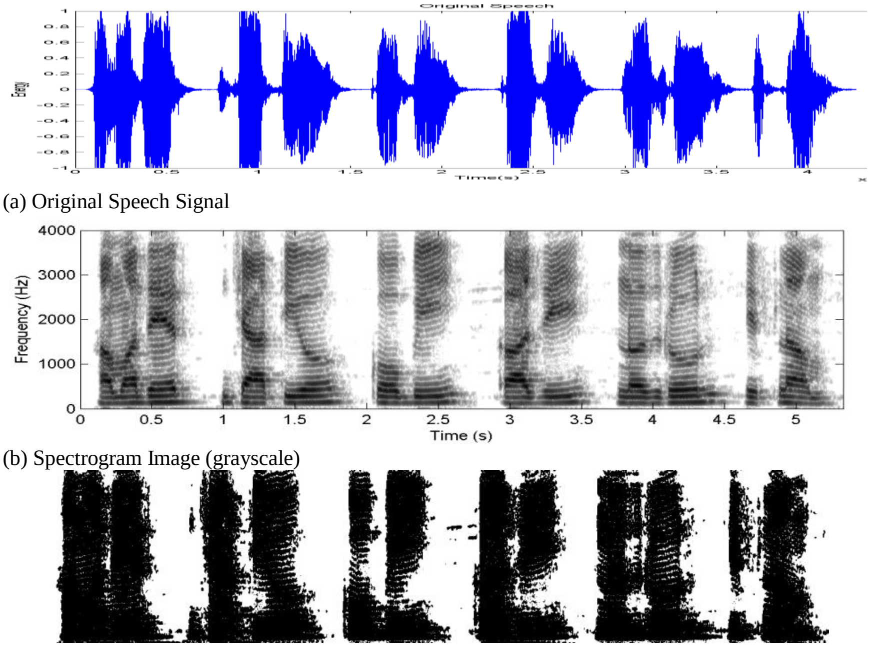
B. Feature Generation and Thresholding

C. Word Boundary Detection

D. Speech Segment Separation

A. Speech Acquisition

Speech acquisition is acquiring of continuous Bangla speech sentences through the microphone. Recording was done by 5 (five) native male speakers of Bengali. The sampling frequency is 16 KHz; sample size is 8 bits, and mono channels are used. The time-domain plot of a speech sentence ('আমাদের জাতীয় কবি কাজী নজরুল ইসলাম') is shown in Figure-4(a).



(a) Original Speech Signal

(b) Spectrogram Image (grayscale)

(c) Thresholded Image

Fig. 4. Thresholded Spectrogram Images of the Speech Sentence 'আমাদের জাতীয় কবি কাজী নজরুল ইসলাম'

B. *Speech Feature Generation and Thresholding*

The feature extraction process generates spectrogram features from Bangla speech sentences. The grayscale spectrogram image of the speech sentence ('আমাদের জাতীয় কবি কাজী নজরুল ইসলাম') is shown in Figure-4(b). Spectrograms can be used to identify spoken words phonetically. For further processing of the spectrogram image, the labels of the image, such x-label, y-label and tile of the image, have been omitted, that's why label or title of the image is not shown in Figure-4(c). The thresholding algorithm is used to separate voiced regions from silence/un-voiced on continuous speech. The Matlab's '`graythresh`' function is used to implement the Algorithm-3. This algorithm returns a level (i.e., threshold) value for which the intra-class variance of the black and white pixels is minimum. The output image replaces all pixels in the input image with luminance greater than or equal to the threshold with the value of 1 (fully white) and less than threshold with 0 (fully black) to get fully black/white image (i.e., thresholded image). The thresholded image of the above speech sentence is shown in Figure-4(c).

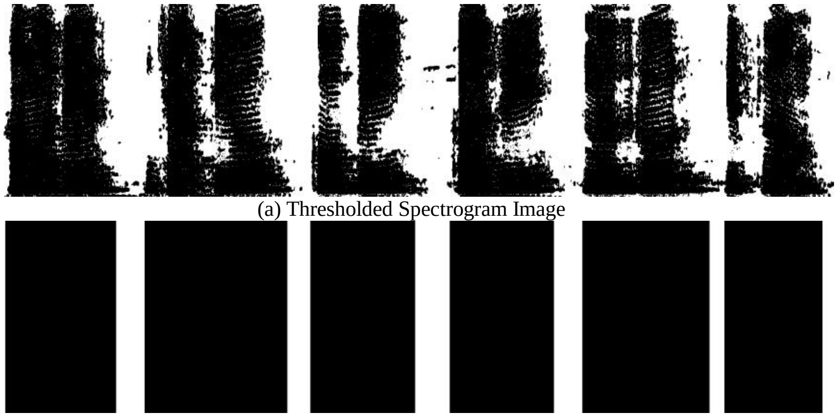
C. *Word Boundary Detection*

The newly introduced *blocking black area method* and *shape identification* techniques to properly detect word boundaries in continuous speech and label the entire speech sentence into a sequence of words/sub-words. The *block black area* method is applied in the thresholded spectrogram image that produces rectangular black boxes in the voiced regions of the speech sentence, as shown in Figure-5. Each rectangular black box represents a speech word or sub-word.

The method uses Matlab's '`regionprops`' function to identify each rectangular object in the binary image that represents speech words/sub-words. The function '`regionprops`' measures the properties of each connected object in the binary image. Different shape measurements properties, such as `'Area'`, `'BoundingBox'`, `'Centroid'` are used to identify each rectangular object in the binary image. The `'Extrema'` measurement, which is a vector of `[top-left top-right right-top right-bottom bottom-right bottom-left left-bottom left-top]`, is used to detect the start (`bottom-left`) and end (`bottom-right`) points of each rectangular object, as shown in Figure-6.

D. *Word Segment Separation*

Each rectangular black box represents a speech segment, such as a word or sub-word. After detecting the start and end points of each black box, the word boundaries in the original speech sentence are marked automatically by these two points and separated each speech segment from the speech sentence. Figure-7 shows that 6 (six) black boxes represent 6 (six) word segments in the speech sentence 'আমাদের জাতীয় কবি কাজী নজরুল ইসলাম'.



(a) Thresholded Spectrogram Image

(b) Rectangle black boxes of thresholded image after applying BBA method

Fig. 5. Effect of applying Blocking Black Area (BBA) Method – Producing rectangle black boxes in voiced regions. (a) Before applying Blocking Black Area Method and (b) After applying Blocking Black Area Method – Each black box represents a word/sub-word of the continuous speech

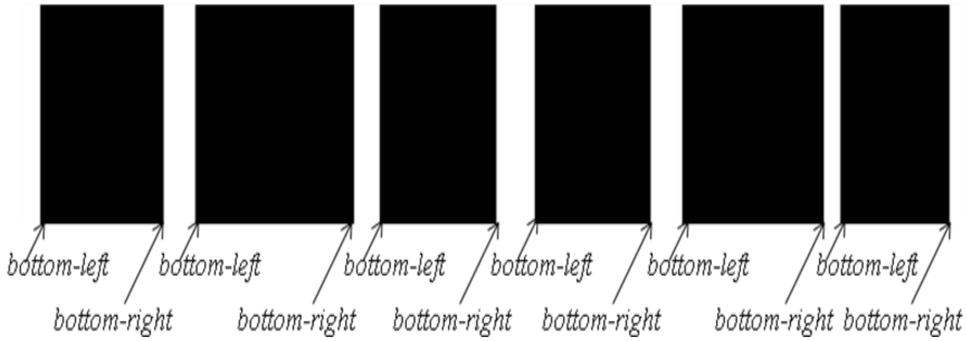


Fig. 6. Star and End point Detection of rectangular object

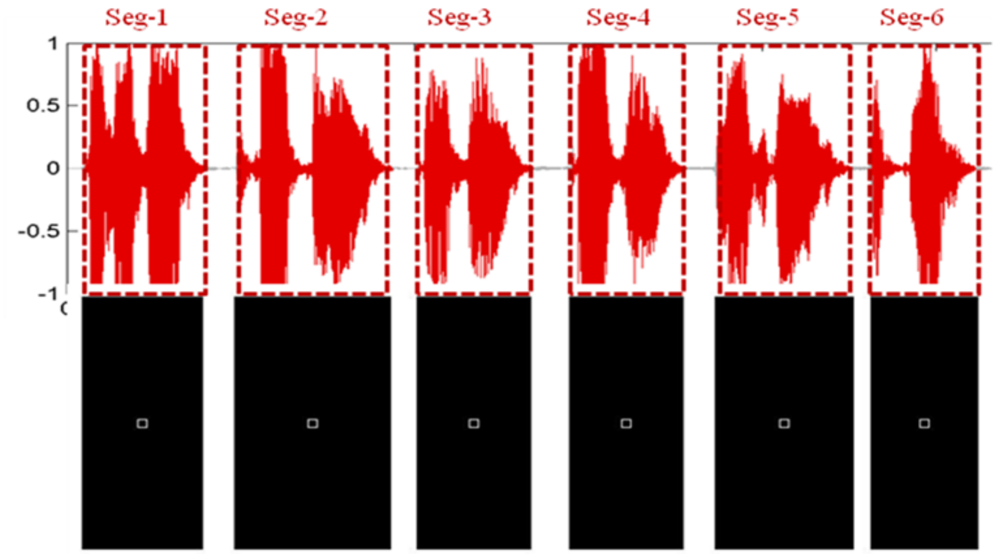


Fig. 7. Word Segments - 6 word segments in speech sentence ‘আমাদের জাতীয় কবি কাজী নজরুল ইসলাম’

VII. Experiments and Results

For speech segmentation, this research proposes the dynamic thresholding algorithm with newly introduced *blocking black area method* to segment the continuously spoken Bangla sentence into words or sub-words. All the programs related to the speech segmentation approaches have been implemented in Matlab. The ‘*myspectrogram.m*’ program computes spectrogram image from the original speech signal.

This research uses MATLAB's '*graythresh*' function to implement modified Otsu's algorithm that returns the desired threshold. The output image replaces all pixels in the input image with luminance greater than or equal to the threshold with the value of 1 (fully white) and less than threshold with 0 (fully black). The '*Blocking Black Area*' method has been implemented in the program '*blockingBlackArea.m*' that produces rectangular black boxes in the thresholded spectrogram image. The research uses MATLAB function '*regionprops*' to identify each rectangular object and the function's '*Extrema*' is used to detect the start and end points of each black box. The word boundaries of the original speech sentence are marked automatically by these two points and cut the word segments from the speech sentence and finally, the speech segments are save as .wav file format.

The developed system has been justified with continuously spoken several Bangla sentences. To test the performance the system, 100 Bangla sentences have been recorded from 5 (five) male speakers of different ages and 656 words have been presented in the 100 Bangla sentences. So, the speech database contains 500 (100x5) Bangla sentences with 3280 (656x5) words. Each sentence has been recorded separately and saved as .wav file format to make the speech database. In segmentation this research expects only properly segmented words as segmentation output, but the program produced some sub-words. The developed system achieved the average segmentation accuracy of **90.58**%; the details result of segmentation is given in Table-1.

TABLE I. SPEECH SEGMENTATION RESULTS

Speaker ID	No. of Sentences	No. of Words Present	No. of Properly Segmented Words	Accuracy (%)
S1	100	656	517	78.81
S2	100	656	601	91.62
S3	100	656	612	93.29
S4	100	656	619	94.36
S5	100	656	622	94.82
Total	**500**	**3280**	**2971**	**90.58**

VIII. CONCLUSION

The main objective of this research is to develop an efficient system that can automatically segments words from the continuously spoken Bangla sentences. This research introduces some ideas to develop the system. This research proposes dynamic thresholding algorithm a new approach, named "Blocking Black Area" method to detect proper word/sub-word boundaries in speech segmentation. Some words are not properly segmented. No or very little gap between two successive words causes two or more words in a single segment. Also the gap within a word causes sub-word segmentation. This is due to some sources of variability is speech, such as, Phonetic identity (two samples might correspond to different phonetic segments), Pitch and Amplitude, Speaker (based age, sex, emotion, etc.), Microphone and Media, and Environment (including background noise, room acoustics, distance from microphone, etc).

For further improvements and expansions of the speech segmentation developed system, this research can be employed by using noise reduction algorithms in a noisy environment. Also a fuzzy logic based speech segmentation approach can be employed.

REFERENCES

[1] Okko Rasanen, "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] Mermelstein P, "Automatic segmentation of speech into syllabic units", Journal of Acoustical Society of America, Vol. 58, No. 4, pp. 880-883, Oct. 1975.

[3] S L Mattys, P W Jusczyk, "Phonotactic cues for segmentation of fluent speech by infants", Cognition 78, 91–121, 2001.

[4] Zhang T and Kuo C C J, "Hierarchical classification of audio data for archiving and retrieving", Proceedings of the Acoustics, Speech, and Signal Processing 1999 on 1999 IEEE International Conference, Vol. 6, pp. 3001-3004, 1999.

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

[6] D Dahan and M R Brent, "On the discovery of novel word like units from utterances: an artificial-language study with implications for native-language acquisition", J. Exp. Psychol. 128 (1999) 165–185.

[7] Thangarajan R and Natarajan A M, "Syllable Based Continuous Speech Recognition for Tamil", South Asian Language Review VOL.XVIII, No.1, 2008.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech Recognition", Prentice Hall, Englewood Cliffs, N.J., 1993.

[14] 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.

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

[16] Shapiro, Linda G. and Stockman, George C., "Computer Vision", Prentice Hall, ISBN 0-13-030796-3, 2002.

[17] Md. Mijanur Rahman and Md. Al-Amin Bhuiyan, "Dynamic Thresholding on Speech Segmentation", IJRET: International Journal of Research in Engineering and Technology, Volume: 02 Issue: 09, Sep-2013.

[18] Gonzalez, Rafael C. & Woods, Richard E, "Thresholding", In Digital Image Processing, pp. 595–611. Pearson Education, 2002.

[19] Nobuyuki Otsu, "A threshold selection method from gray-level histograms", IEEE Trans. Sys., Man., Cyber. 9 (1): 62–66, 1979.