Continuous Bangla Speech Segmentation, Classification and Feature Extraction

Md. Mijanur Rahman[1], Md. Farukuzzaman Khan[2] and Md. Al-Amin Bhuiyan[3]

[1]Dept. of Computer Science & Engineering, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymenshingh, Bangladesh.

[2]Dept. of Computer Science & Engineering, Islamic University, Kushtia, Bangladesh.

[3]Dept. of Computer Science & Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh.

Abstract
Continuous speech recognition is a multileveled pattern recognition task, which includes speech segmentation, classification, feature extraction and pattern recognition. In our work, a blind speech segmentation procedure was used to segment the continuously spoken Bangla sentences into words/sub-words like units using the end-point detection technique. These segmented words were classified according to the number of syllables and the sizes of the segmented words. MFCC signal analysis technique was used to extract the features of speech words, which including windowing. The developed system achieved the segmentation accuracy rate at about 98% and total 24 sub-classes of segmented words with MFCC features.
Keywords: *Segmentation, Feature Extraction, Speech Classification, MFCC and Windowing.*

1. Introduction

Speech can be used as a useful interface to interact with the machine. Speech processing is very much important in many fields and disciplines including acoustics, digital signal processing, pattern classification, linguistics, physiology, hearing, neuroscience, and computer science [1]. A vast majority of the currently available speech processing systems, including, medium to large vocabulary speech recognition systems [2, 3], speaker recognition systems [4, 5, 6], and language identification systems [7], are designed based on sub-word acoustic units. Speech recognition is very popular field of research and we have a continuous effort to develop a Bangla speech recognition system in which the robots will able to speak and understand Bangla speech to consolidate the relationship between robots and human beings. This work is a part of that effort.

2. Speech Segmentation

Sentence segmentation is very helpful in many applications, such as speech summarization [8], video summarization [9], speech document indexing and retrieval [10]. In general, there are two kinds of segmentation [11]. One is phonemic segmentation, which segments speech into phonemes and other is syllable-like unit segmentation, which segments speech into syllables, sub-words or words.

In our work, a blind speech segmentation procedure was used that allows a speech sample to be segmented into words/sub-word units without the knowledge of any linguistic information (such as, orthographic or phonetic transcription). This was done by using end-point detection technique [12] which detects the proper start and end points of the speech events. The start and end points are detected by tracing abrupt change of the data sequence, which is greater or less than a given threshold. Figure-1 shows the start and end points of four words within the sentence "কোন কাজই কঠিন নয়".

This method has some complexities. First is that the word boundaries are very unclear in continuous speech. Second is that the effects due to co-articulation [13] are much stronger in continuous speech. Third are stresses in articulation, particular words in a sentence and even some particular syllables in a word are often emphasized, while others are poorly articulated. To avoid these complexities, the articulation of continuous speech is such that there is sufficient pause between speech words as shown in Figure-1.

3. Speech Words Classification

Classification means collection of segmented words and sub-words into different classes based on some properties. In this research, an effort was made to categorize the segmented words and sub-words according to the number of syllables and the length of segmented units[12].

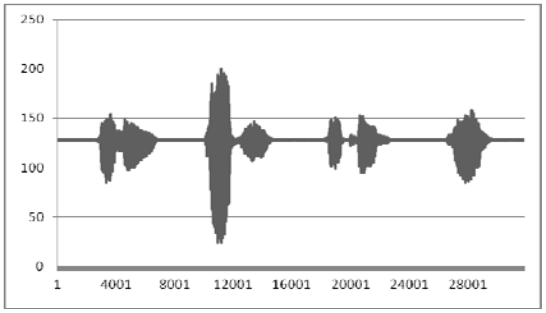


Figure-1. The start and end points of words in the continuous speech "কোন কাজই কঠিন নয় ".

3.1. Syllable-based classification

In the first-level of classification, named *syllable-based classification*, three different classes were formed according to the number of syllables of the segmented word. Table-1 shows this type of classification and Table-2 shows the example of classified words. The wave shapes of some classified speech words are shown in Figure-2. The algorithm for syllable-based classification is given below.

Syllable-based classification Algorithm:

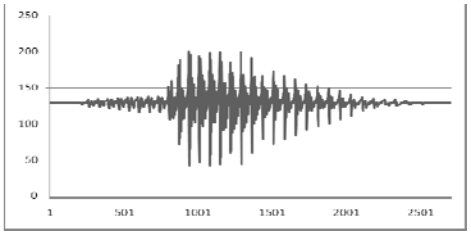
1. Select a segmented word, $W[i]$;
2. Calculate the number of gaps within $W[i]$ and set this number into N_g;
3. (i) If $N_g = 1$ then:
   Select a class c_1;
   Assign the segmented word $W[i]$ to class C_1;
   (ii) Else If $N_g = 2$ then:
   Select a class c_2;
   Assign the segmented word $W[i]$ to class C_2;
   (ii) Else:
   Select a class c_3;
   Assign the segmented word $W[i]$ to class C_3;
4. Repeat the above steps for all segmented words;

Table-1. Syllable-based classes

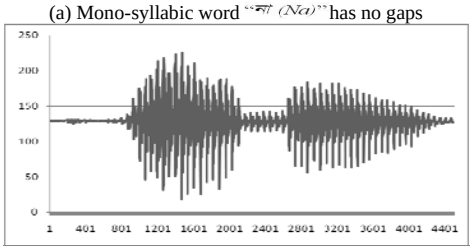
Name of Classes	Contents
Class-1 (C1)	Segmented words of mono-syllable
Class-2 (C2)	Segmented words of di-syllables
Class-3 (C3)	Segmented words of tri or more syllables

Table-2. Some classified Bangla words

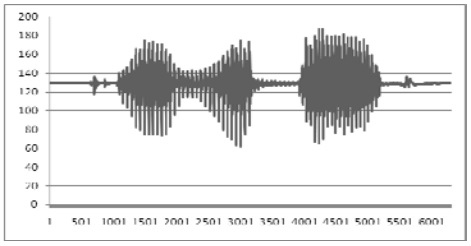
Words of Class-1
না, নেই, হয়, দেশ, ঐ, যা, যায়, বেশ, ভোট, যে, কেউ, কোন, শোক, জোট, আর, আয়, আজ, পর, প্রায়, বা, এ, এই, এক, ঐ, যা, যায়, রান, ত্রান, রাত, কাজ, তা, ও, সব, গাছ, হয়, ঘুষ, দিন, বিল, কি, তাই, তার, তাল
Words of Class-2
তিনি, পড়েন, জানা, তবে, হবে, কাটান, আছে, আটক, পারি, বিজয়, টাকা, আছে, যুদ্ধ, থাকে, থাকেন, থাকছে, দলে, কোটি, আমি, আমরা, আমার, পাতা, পুলিশ, বলা, উৎসব, যায়নি, যদি, কথা, করেনি, গ্রহণ, হবে, ঘরে, হলে, হতে, হবার, চলবে, তিনি, টাকা, তাদের
Words of Class-3
ধরনের, ঘটনায়, জানানোর, আমাদের, বর্তমান, প্রতিটি, ফলাফল, অবস্থার, বিষয়টি, স্বাধীনতার, অধিকার, সবাই, সরবরাহ, হাসপাতালে, সুবিধা, সেখানে, অনুষ্ঠানের, অবস্থার, আপনাদের, নারীদেরই, প্রদর্শিত, বর্তমানে, লাগানো, কর্তৃপক্ষের, সহযোগিতা, ধন্যবাদ, সৌন্দর্য, ধরনের, অগ্রণী, আমাদের, ধারনা, আকাশের, ইতিহাস, পদত্যাগ, পৃথিবীর, প্রদর্শনী, ভবিষ্যতের, শহরের, চিকিৎসাধীন, বিভিন্ন



(a) Mono-syllabic word "না (Na)" has no gaps



(b) Di-syllabic word "জানা (Jana)" has a single gap



(c). Tri-syllabic word "আমাদের (Amader)" has two gaps

Figure - 2. Examples of syllable-based classified words

3.2. Length-based classification

In the second-level of classification, named *length-based classification*, words of each three main classes are distributed among the eight different sub-classes according to the length/size of the segmented words. So, after classification 24 sub-classes are found. Table-3 shows this type of classification. The algorithm for length-based classification is given below.

Length-based classification algorithm:

1. Select a class, C_k;
2. Select a segmented word, *w[i]* from c_k;
3. Calculate the length (N) of *w[i]*;
4. (i) If *N<=2000* then:
   Select a sub-class sc_{k1};
   Assign the word *W[i]* to sub-class sc_{k1};
   (ii) Else If *N>2000 AND N<=3000* then:
   Select a sub-class sc_{k2};
   Assign the word *W[i]* to sub-class sc_{k2};
   (iii) Else If *N>3000 AND N<=4000* then:
   Select a sub-class sc_{k3};
   Assign the word *W[i]* to sub-class sc_{k3};
   (iv) Else If *N>4000 AND N<=5000* then:
   Select a sub-class sc_{k4};
   Assign the word *W[i]* to sub-class sc_{k4};
   (v) Else If *N>5000 AND N<=6000* then:
   Select a sub-class sc_{k5};
   Assign the word *W[i]* to sub-class sc_{k5};
   (vi) Else If *N>6000 AND N<=7000* then:
   Select a sub-class sc_{k6};
   Assign the word *W[i]* to sub-class sc_{k6};
   (vii) Else If *N>7000 AND N<=8000* then:
   Select a sub-class sc_{k7};
   Assign the word *W[i]* to sub-class sc_{k7};
   (viii) Else:
   Select a sub-class sc_{k8};
   Assign the word *W[i]* to sub-class sc_{k8};
5. Repeat the above steps for all segmented words;

Table-3. Length-based classes for a class C_k

Name of Sub-classes	Word Length (Samples)
Sub-class-1 (SC_{k1})	up to 2000
Sub-class-2 (SC_{k2})	2001 to 3000
Sub-class-3 (SC_{k3})	3001 to 4000
Sub-class-4 (SC_{k4})	4001 to 5000
Sub-class-5 (SC_{k5})	5001 to 6000
Sub-class-6 (SC_{k6})	6001 to 7000
Sub-class-7 (SC_{k7})	7001 to 8000
Sub-class-8 (SC_{k8})	Greater than 8000

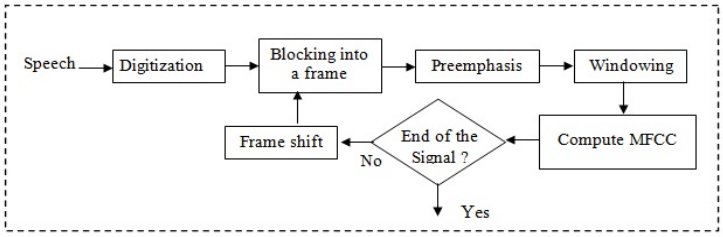


Figure – 3. MFCC feature extraction process

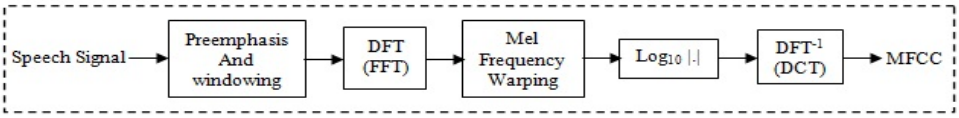


Figure – 4. Computation of MFCC feature

4. Feature Extraction

Feature extraction converts the speech waveform to some type of parametric representation (a collection of meaningful features). A good feature may produce a good result for any recognition system. MFCC is the stronger feature for Bangla speech recognition [14] and was used in this research. The process of MFCC feature extraction is sown in the Figure-3. Now, we shall explain the step-by-step computation of MFCC in this section.

5. Mel Frequency Cepstrum Coefficient (MFCC)

The signal is cut into short overlapping frames, and for each frame, a feature vector is is computed, which consists of Mel Frequency Cepstrum Coefficients. The cepstrum is the inverse Fourier transform of the log-spectrum. Thus, the computation of MFCC [15] includes a series of operations e.g. Fast Fourier Transform (FFT), Mel frequency warping, Logs of power and Discrete Cosine Transform (DCT), as shown in Figure-4.

5.1. Pre-emphasis

The speech signal *s(n)* is sent to a high-pass filter, $s_2(n) = s(n) - a*s(n-1)$, where $s_2(n)$ is the output signal and the value of a is usually between *0.9* and *1.0*. In our research we used $a = 0.98$. The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. The speech after pre-emphasis sounds became sharper with a smaller volume.

5.2. Frame blocking

The input speech signal is segmented into frames of 20~30 ms with optional overlap of 1/3~1/2 of the frame size. Usually the frame size (in terms of sample points) is equal power of two in order to facilitate the use of FFT. If this is not the case, we need to do zero padding to the nearest length of power of two.

5.3. Windowing

In windowing, each data frame has to be multiplied with a window function [16] in order to keep the continuity of the first and the last points in the frame. If the signal in a frame is denoted by *s(n), n = 0,…N-1*, then the signal after windowing is *s(n)*w(n)*, where *w(n)* is the window function. In our research, we used different types of window functions, such as Hamming, Hanning, Rectangular, Bohman, Trangle, Welch, Kaiser and Blackman windows. In general, the Hanning window is satisfactory in 95% of cases. It has good frequency resolution and reduced spectral leakage. If you do not know the nature of the signal but you want to apply a smoothing window, start with the Hanning window.

For $n = 0, 1, 2, \ldots, N-1$, where N is the length of the window, the following equations define various windowing functions.

The equation for the Hanning window is

$$w(n) = 0.5 - 0.5\cos\frac{2\pi n}{N} .$$

The equation for Hamming window is

$$w(n) = 0.54 - 0.46\cos\frac{2\pi n}{N} .$$

The equation for rectangular window is $w(n) = 1.0$.

The equation for Bohman window is

$$w(n) = \left(1 - \frac{|n - N/2|}{N/2}\right)\cos\left(\pi\frac{|n - N/2|}{N/2}\right) + \frac{1}{\pi}\sin\left(\pi\frac{|n - N/2|}{N/2}\right)$$

The equation for Parzen window is

$$w(n) = \begin{cases} 1 - 6\left(\frac{n - N/2}{N/2}\right)^2 + 6\left(\frac{|n - N/2|}{N/2}\right)^3 & 0 \le \left|n - \frac{N}{2}\right| \le \frac{N}{4} \\ 2\left(1 - \frac{|n - N/2|}{N/2}\right)^3 & \frac{N}{4} < \left|n - \frac{N}{2}\right| \le \frac{N}{2} \end{cases}$$

The equation for the triangle window is

$$w(n) = 1 - \left|\frac{2n - N}{N}\right| .$$

The equation for the Welch window is

$$w(n) = 1 - \left(\frac{n - N/2}{N/2}\right)^2 .$$

The equation for the Blackman window is

$$w(n) = \begin{cases} 0.42 - 0.5\cos(2\pi n/N) + 0.08\cos(4\pi n/N), & 0 \le n < N \\ 0, \quad elsewhere \end{cases}$$

We implemented the above windowing function in C++ for extracting features from speech words.

5.4. Fast Fourier Transform (FFT)

A Fast Fourier transform (FFT) is an efficient algorithm to compute the Discrete Fourier Transform (DFT) and its inverse. An FFT computes the DFT and produces exactly the same result as evaluating the DFT definition directly; the only difference is that an FFT is much faster.

Let $x_0, \ldots, x_{N-1}$ be complex numbers. The DFT is defined by the formula

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi k\frac{n}{N}} \qquad k = 0, \ldots, N-1.$$

Evaluating this definition directly requires $O(N^2)$ operations: there are N outputs X_k, and each output

requires a sum of N terms. An FFT is any method to compute the same results in O(N log N) operations. In our research, a well-known split-radix FFT (RS-FFT) algorithm [17] was used. This is a divide and conquer algorithm that recursively breaks down a DFT of any composite size $N = N_1N_2$ into many smaller DFTs of sizes N_1 and N_2, along with O(N) multiplications by complex roots of unity traditionally called twiddle factors [18].

When we perform FFT on a frame, we assume that the signal within a frame is periodic, and continuous. If this is not the case, we can still perform FFT but the incontinuity at the frame's first and last points is likely to introduce undesirable effects in the frequency response which is known as spectral leakage [19]. To overcome this problem, we can use windowing (discussed above) before performing FFT.

6. Experimental Results

In this experiment, 120 Bangla spoken sentences, which contain 758 speech words, were originated from 6 male speakers. Therefore, the speech database contains 720 speech sentences with 4,548 words. The segmentation result for these 6 speakers is shown in Figure-5. The developed system achieved the average segmentation accuracy of 98.48%. After segmentation, the test database contains properly segmented 4224 words, 704 words for each speaker.

The segmented words were the input of the classification program. In the first-level of classification, the program received the segmented words as input and produced 3 different classes of words as output, based on the number of syllables. In the second-level of classification, words from 3 main classes were distributed among 8 different sub-classes, based on the word lengths. After classification, 24 different sub-classes were obtained. The detailed classification results are shown in Figure-6 and Figure-7.

Finally, the feature extraction program received the segmented words of each class as input and produced MFCC features from these segmented words as output. The MFCC feature graphs of some segmented speech words are shown in Table-4.

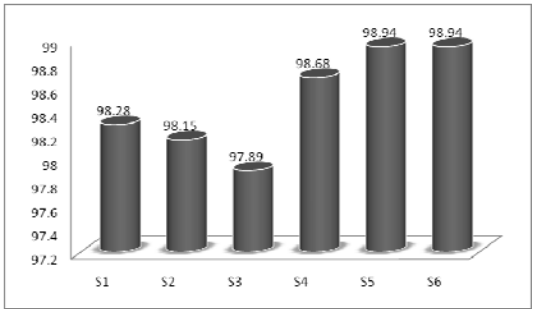
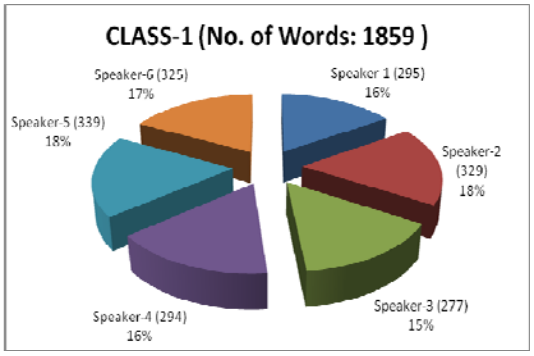
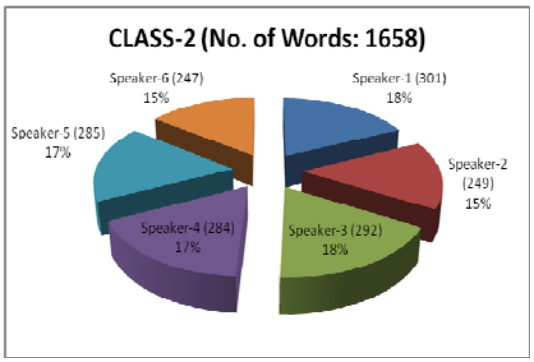


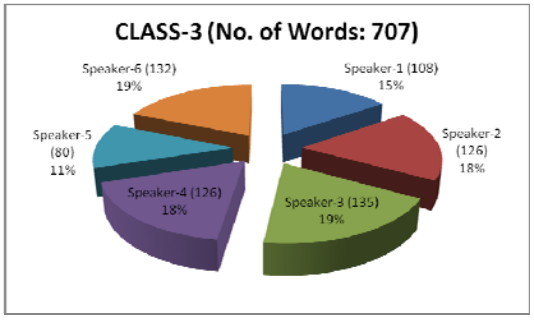
Figure-5. Segmentation Results



(a) Classification result for Class-1

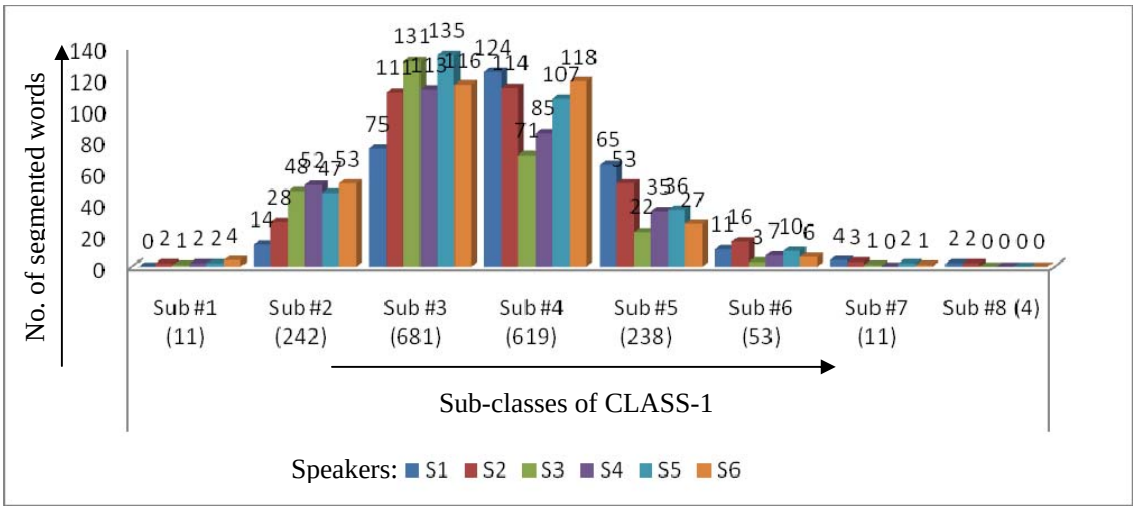


(b) Classification result for Class-2

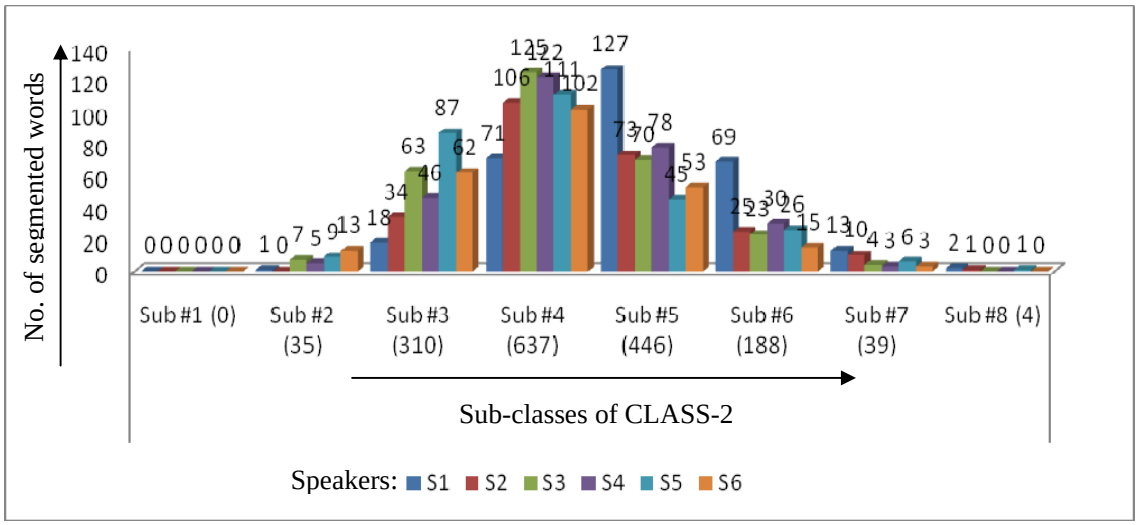


(c) Classification result for Class-3

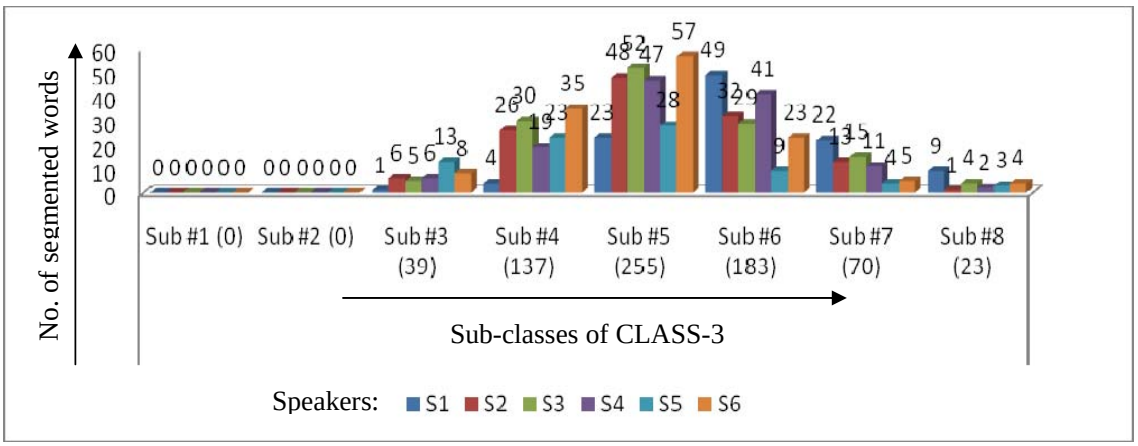
Figure-6: Syllable-based classification results of six different speakers's segmented words.



(a)



(b)



(c)

Figure-7. Length-based classification results (a) *for CLASS #1* (b) *for CLASS #2* and (c) *for CLASS #3*

Table-4. MFCC feature graphs of each segmented word of speech sentence "আজ আমরা সুন্দর একটি সমাজ গড়ে তুলতে পারি (Azz Aamra Sundar Akti Somaz Gore Tulte Pari)" including Hanning and Hamming window functions.

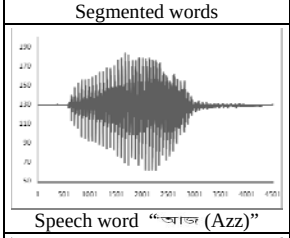
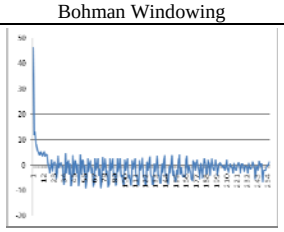
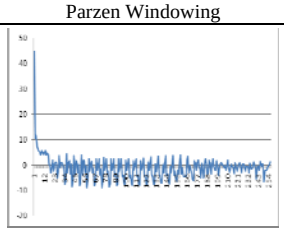
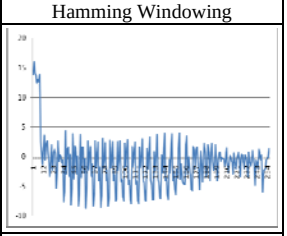
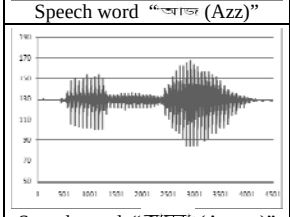
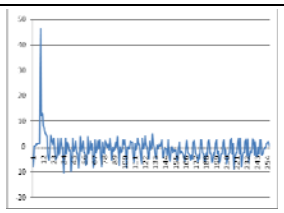
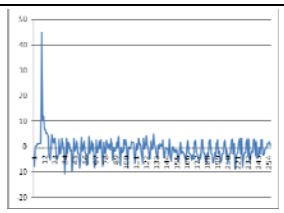
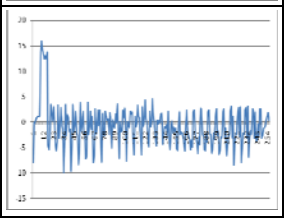
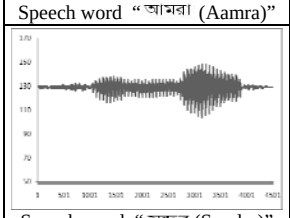
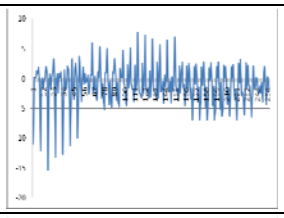
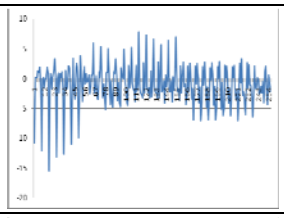
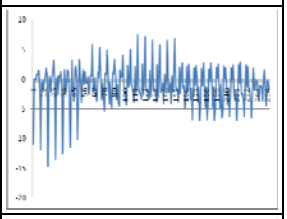
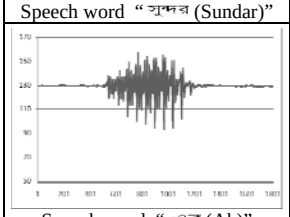
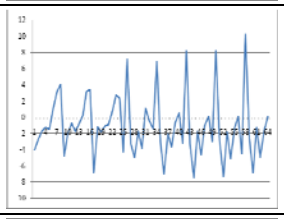
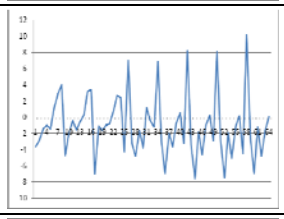
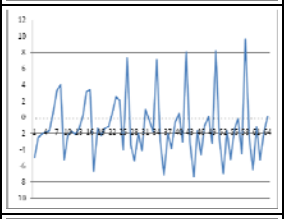
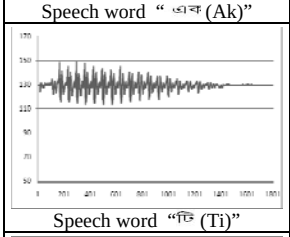
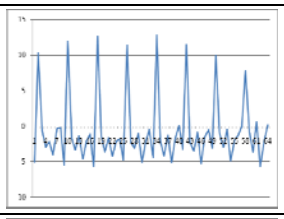
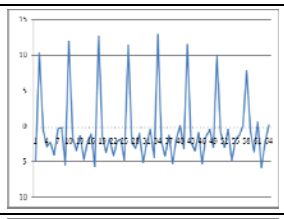
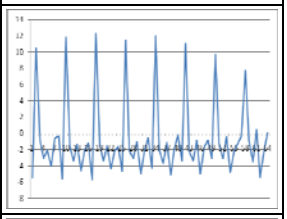
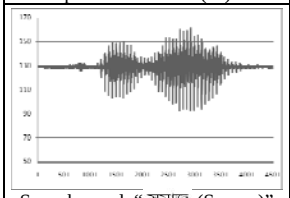
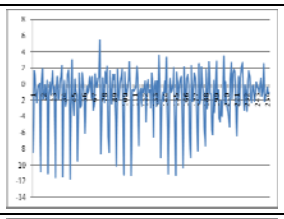
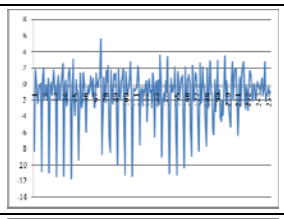
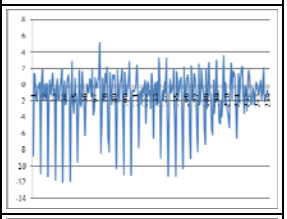
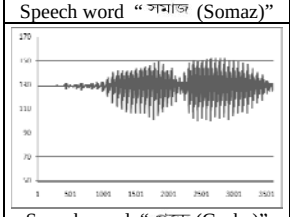
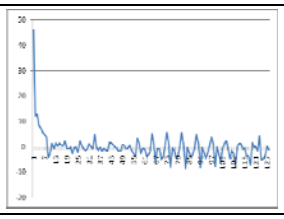
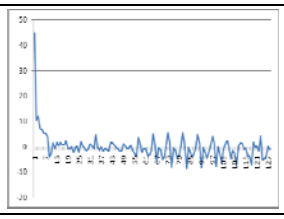
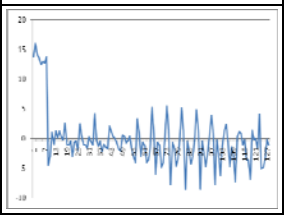
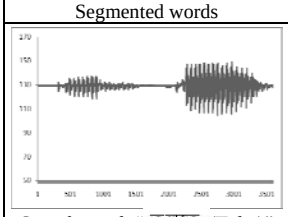
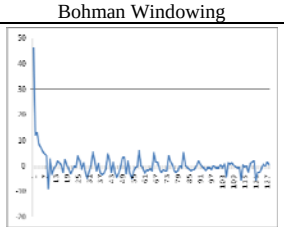
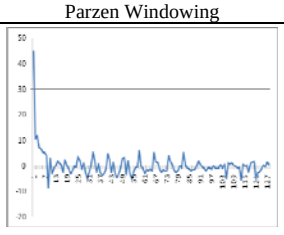
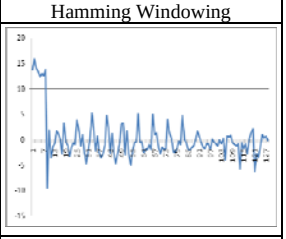
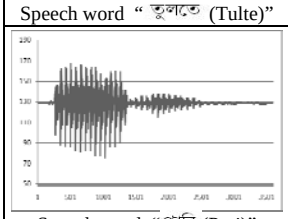
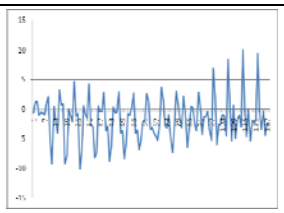
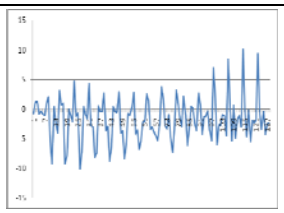
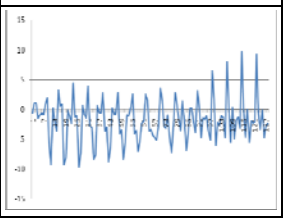
Original Graph of Segmented words	MFCC feature graph including Bohman Windowing	MFCC feature graph including Parzen Windowing	MFCC feature graph including Hamming Windowing
 Speech word "আজ (Azz)"			
 Speech word "আমরা (Aamra)"			
 Speech word "সুন্দর (Sundar)"			
 Speech word "একটি (Ak)"			
 Speech word "টি (Ti)"			
 Speech word "সমাজ (Somaz)"			
 Speech word "গড়ে (Gorhe)"			

Table-4 to be continued…

Original Graph of Segmented words	MFCC feature graph including Bohman Windowing	MFCC feature graph including Parzen Windowing	MFCC feature graph including Hamming Windowing
 Speech word " তুলতে (Tulte)"			
 Speech word "পারি (Pari)"			

7. Discussion

In this research, the main goal was to develop a system that automatically segments continuous Bangla speech, categories segmented words and extracts features from segmented words. Among the different techniques, the end-point detection technique was used for word/sub-words segmentation produced very good results. It is seen in table-3, the average segmentation accuracy rate is 98.48%, and it is quite satisfactory. The classification and feature selection is the most important factors in designing a speech recognition system. From the study of different previous research works it was observed that among the different features the MFCC produces better results in recognition system. Also the selection of a window function is not a simple task. Each window function has its own characteristics and preferred application.

It was observed that some of the words were not segmented properly. It was also observed that same words were appeared in different clusters in some cases. This is due to some common causes frequently occurred in the continuous speech recognition system. The utterance of words/ sub-words differs depending on their position in the sentence. The pauses between the words/sub-words are not identical in all cases because of the variability nature of the speech signals. The other important cause is the non-uniform articulation of speech. Even for a single speaker it is difficult to maintain the uniformity in articulation for the same speech. The speech signal is very much sensitive to the speaker's properties such as age, sex, emotion, etc., and environment.

8. Conclusion and Future Research

We presented a speech recognition front-end that is used for segmenting continuous Bangla speech, classifying and extracting features from speech words. Proper segmentation, classification and feature extraction is a crucial task in the development of large vocabulary continuous speech recognition system. Experiments evaluate the proposed approach and each feature set. The result is satisfying. It achieves comparable accuracy as the method using speech recognition but with lower computational cost. The recognition task will be done to develop the complete interface between robots and human for further research. Also, to design more reliable speech recognition system, the future researchers should employ more speakers of different ages and genders, and consider noisy speech data using different speech processing and recognition tools like Neural Networks, Hidden Markov Model (HMM), and Fuzzy Logic.

References

[1] Jesse C. Hansen, "*Modulation Based Parameters for Automatic Speech Recognition*", MSc Thesis, Department of Electrical Engineering, University of Rhode Island, 2003.

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

[3] B-H Juang and L.R. Rabiner, "*Fundamentals of Speech Recognition*", Prentice-Hall, NJ, 1992.

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

[5] T. Matsui and S. Furui. Concatenated phoneme models for text-variable Speaker Recognition. In Proceedings of ICASSP, pp. 391-394, 1994.

[6] M. Sharma and R. Mammone, "*Subword-based text dependent speaker verification system with user-selectable passwords*", In Proceedings of ICASSP, pp. 93-96, 1996.

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

[8] K. Zechner, "*Summarization of Spoken Language – Challenges, Methods, and Prospects*" Technology ExperteZine, Issue 6, January 2002.

[9] Y.-F. Ma, L. Lu, H.-J. Zhang and M. J. Li. "*An Attention Model for Video Summarization*", 10th ACM International Conference on Multimedia 2002.

[10] K. Koumpis and S. Renals, "*The role of Prosody in a Voicemail Summarization System*" Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding on Prosody in Speech Recognition and Understanding, pp.139-146, Red Bank, NJ.

[11] C. T. Hsieh and J. T. Chien, "*Segmentation of continuous speech into phonemic units*", *Proceedings of International Conference on Information and Systems*, pp. 420-424, 1991.

[12] Md. Mijanur Rahman, Md. Farukuzzaman Khan and Mohammad Ali Moni, "*Speech Recognition Front-end for Segmenting and Clustering Continuous Bangla Speech*", DIU Journal of Science and Technology, Daffodil International University, Dhaka, Bangladesh, Vol 5(1), Jan 2010.

[13] Kai-Fu Lee and Fil Alleva, "*Continuous Speech Recognition*", An article of "Advances in Speech Signal Processing – Edited by S. Furui and M. M. Sondhi", Marcel Dekker, Inc., New York, USA, 1992.

[14] Md. Farukuzzaman Khan and Dr. Ramesh Chandra Debnath, "*Comparative Study of Feature Extraction Methods for Bangla Phoneme Recognition*", 5th ICCIT 2002, East West University, Dhaka, Bangladesh, PP 27-28, December 2002.

[15] Fang Zheng, Guoliang Zhang and Zhanjiang Song, "*Comparison of Different Implementations of MFCC*", J. Computer Science & Technology, 16(6): 582–589, 2001.

[16] Eric W. Weisstein. CRC Concise Encyclopedia of Mathematics. *CRC Press*, ISBN 1584883472, 2003.

[17] P. Duhamel and H. Hollman, "*Split-radix FFT algorithms. Electronics Letters*", 20, 14-16, Jan 1984.

[18] Cooley, James W., and John W. Tukey, "*An algorithm for the machine calculation of complex Fourier series*", Math. Comput. 19: 297–301, 1965.

[19] Ramirez, Robert W., "*The FFT, Fundamentals and Concepts*", Prentice-Hall, New Jersey, 1985.

Biographies of Authors

Md. Mijanur Rahman

Mr. Rahman is working as an assistant professor of the department of Computer Science and Engineering in Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh. He completed his B Sc (Hons) and M Sc in CSE degree from Islamic University, Kushtia, Bangladesh. At present he is continuing his PhD research work in the department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. He has got a number of research articles published in different local and international journals.

Prof. Md. Farukuzzaman Khan

Mr. Khan is working as professor of the department of Computer Science and Engineering in Islamic University, Kushtia, Bangladesh. He completed his B Sc (Hons), M Sc and M. Phil degree from Rajshahi University, Rajshahi, Bangladesh. He is a PhD researcher in the department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh. He has has got a number of research articles published in different local and international journals.

Prof. Dr. Md. Al Amin Bhuiyan

Dr. Bhuiyan is working as a faculty member of the department of Computer Science and Engineering in Jahangirnagar University, Savar, Dhaka, Bangladesh. He awarded his Ph D on Information & Communication Engineering from Osaka City University, Japan. He completed his B Sc and M Sc in Applied Physics and Electronics from Dhaka University, Bangladesh. He has many publications in local and international journals.