SPEECH RECOGNITION FRONT-END FOR SEGMENTING AND CLUSTERING CONTINUOUS BANGLA SPEECH

Md. Mijanur Rahman[1], Md. Farukuzzaman Khan[2] and Mohammad Ali Moni[3]
[1]Dept. of CSE, Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh.
[2]Dept. of CSE, Islamic University, Kushtia, Bangladesh.
[3]Dept. of CSE, Pabna University of Science and Technology, Pabna, Bangladesh.
E-mail: mijan_cse@yahoo.com, mfkhan_bd@yahoo.com, moni_cse1@yahoo.com

Abstract: *This research is concerned with the development of speech recognition front-end for segmenting and clustering continuous Bangla speech sentence to some predefined clusters. From the study of different previous research works it was observed that the front-end is an important part of any speech recognition system. In our work, the original speech sentences were recorded and stored as RIFF (.wav) file format. Then a segmentation approach was used to segment the continuous speech into uniquely identifiable and meaningful units. Among the different techniques, the word/sub-word segmentation is simple and produces very good results. This is why this technique was selected for speech segmentation to obtain improved performance. After segmentation, the segmented words were clustered into different clusters according to the number of syllables and the sizes of the segmented words. The test database contained 758 words/sub-words segmented from 120 sentences. Each sentence was recorded from six different speakers and saved as a different wave file. The developed system achieved the segmentation accuracy rate at about 95%.*

Keywords: *Front-end, Phonemic and Word segmentation, Clustering, End Point Detection.*

1 Introduction

Speech is the most important manner of communication for humans, to exchange information. As technology advances and increasingly sophisticated tools become available to use with speech and music signals, scientists can study these sound more effectively and invent new ways of applying them for the benefit of humankind. Such research has led to the development of speech and music synthesizers, speech transmission systems, continuous speech segmentation systems, and automatic speech recognition systems. In the past few decades research in this fascinating field has produced remarkable results.

Bangla is an important language with a rich heritage and is spoken by approximately 8% of the world population [1]. Early researchers have developed Bangla speech recognition system for only phonemes [2] letters [1], words [3, 4, 5], small [6] or medium vocabulary speech system [7] with limited success. We have a continuous effort to develop a large vocabulary Bangla speech recognition system in the speech and image processing laboratory of Islamic University, Bangladesh. This work is a part of that effort.

2 Speech Segmentation

Speech segmentation is the signal processing front-end that segments continuous speech into uniquely identifiable or meaningful units as phonemes, syllables, words or sub-words and processes them to generate distinguishable features. Segmentation plays an important role in speech recognition to reduce memory size and minimize the computation complexity for large vocabulary systems. In general, there are two kinds of segmentation. One is phonemic segmentation [8], which segments speech into phonemes and other is syllable-like unit segmentation [9], which segments speech into syllables, sub-words or words. For both phonemic and syllabic unit segmentation, most of the approaches are based on the thresholds of the parameters used to segment the speech data. The thresholds of the segmentation features are usually set based on various studies. As it is often difficult to distinguish proper phonemes from continuous speech, syllable-like unit segmentation by end point detection technique [10] was selected for our research. In this method, Segmentation is done by detecting the proper start and end points of speech events. Fig.1 shows the start and end

points of three words within the sentence "পৃথিবী এগিয়ে চলেছে".

The start and end points are detected by tracing abrupt change of the data sequence,

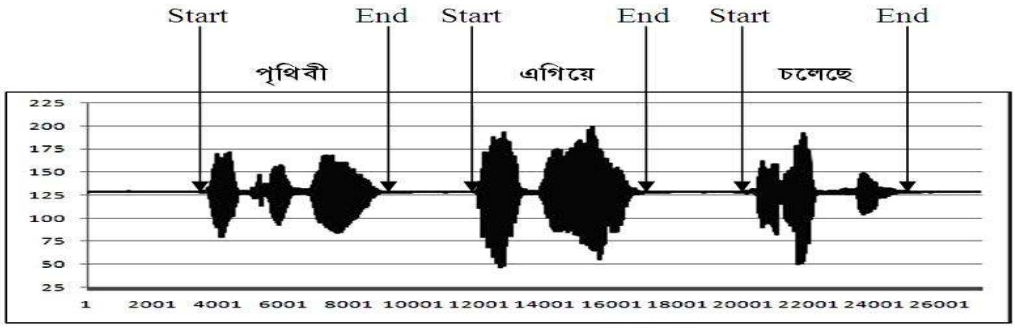


Fig. 1 The start and end points of words in the speech sentence "পৃথিবী এগিয়ে চলেছে".

which is greater or less than a given threshold. Though this method is very simple, but there are some complexities. First is that the word boundaries are very unclear in continuous speech. A very regular problem is that two successive words in a sentence are merged with an omission of phonemes and even the sentence becomes a long vocalic segment that makes difficult to detect word boundaries. Second is that the effects due to co-articulation [11] 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. Remembering these complexities, an acceptable system may be designed for segmentation using end point detection technique if the articulation of continuous speech is such that there is sufficient pause between speech units as shown in Fig.1.

3 Clustering

Clustering means collection of segmented words and sub-words into different clusters 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.

In the first-level of clustering (i.e., Syllable-based clustering), three different clusters were formed according to the number of syllables as shown in Table-1 using the following algorithm:

1. Select a segmented word, *w[i]*;
2. Count the number of gaps within *w[i]* and Choose a cluster, c_k;
3. Repeat steps 1 and 2 for all segmented words;

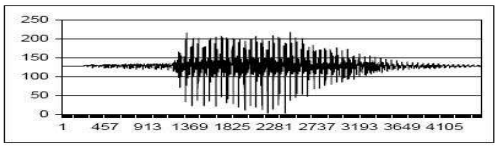
Fig.2 shows the examples of Mono, Di and Tri-Syllabic words.

In the second-level of clustering (i.e., Length-based clustering), words of each of the three main clusters are distributed among the eight different sub-clusters according to the length/size of the segmented words as shown in Table-2 using the following algorithm:

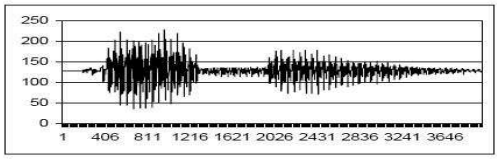
1. Select a segmented word, *w[i]* from a cluster, c_k;
2. Calculate the length of *w[i]* and Choose a sub-cluster, sc_k;
3. Repeat steps 1 and 2 for all segmented words of c_k;

Table1 Syllable-based clusters

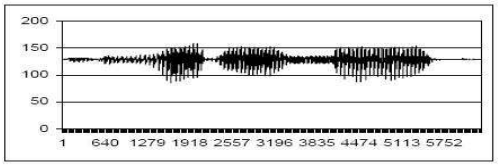
Name of Clusters	Contents
Cluster-1	Segmented words of mono-syllable
Cluster-2	Segmented words of di-syllables
Cluster-3	Segmented words of tri or more syllables



(a) Mono-syllabic word "নেই" has no gaps



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



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

Fig. 2 Syllable-based Clustering

Table 2 Length-based clusters

Name of Sub-clusters	Segment Size (in bytes)
Sub-cluster-1	up to 2000
Sub-cluster-2	2001 to 3000
Sub-cluster-3	3001 to 4000
Sub-cluster-4	4001 to 5000
Sub-cluster-5	5001 to 6000
Sub-cluster-6	6001 to 7000
Sub-cluster-7	7001 to 8000
Sub-cluster-8	Greater than 8000

4 Methodological Steps

The methodological steps of our system employ in this research are as follows:

Step-1: Speech Acquisition

The sample speech was recorded in laboratory with the help of a close-talking microphone, Creative Vibra-128 sound card and windows default sound recorder software. The continuously spoken 120 different sentences that contained 758 distinct words originated from six male speakers were recorded as *wav* file to make a sample database. The utterances were recorded at a sampling rate of 8.00 KHz and coded in 8 bits PCM.

Step-2: Preprocessing

Some recorded speech sentences contain noise at the beginning or the end and long-time silence (Fig. 3). This may affect the segmentation performance. For this reason, these noises were removed from the speech sentences. These noises and long-time silences were cut off from speech by checking voice-data at those segments.

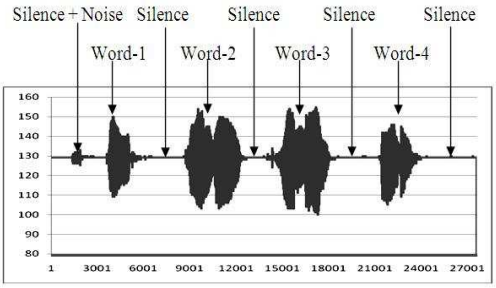


Fig. 3 Speech sentence that contains noise and silences

To extract speech data, the first 44 bytes header information was discarded from the stored *wav* file and then speech data were extracted and stored in a buffer as integers.

Step-3: Segmentation

After extraction from *wav* file, the speech data was segmented into word/sub-words using end-point detection technique [10] as stated earlier. Each segmented word/sub-word was the input of the next step for cluster identification.

Step-4: Clustering

As stated earlier, two-step clustering technique was used to categorize the segmented words into different 24 clusters, according to the number of syllables and the sizes of the segmented words/sub-words. The Dendrogram [12] for this clustering is shown in Fig. 4.

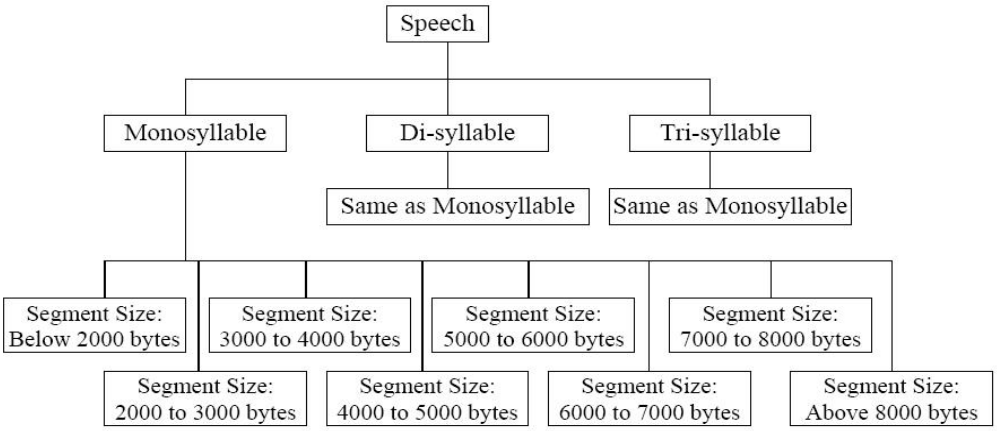


Fig. 4 The Dendrogram of clustering

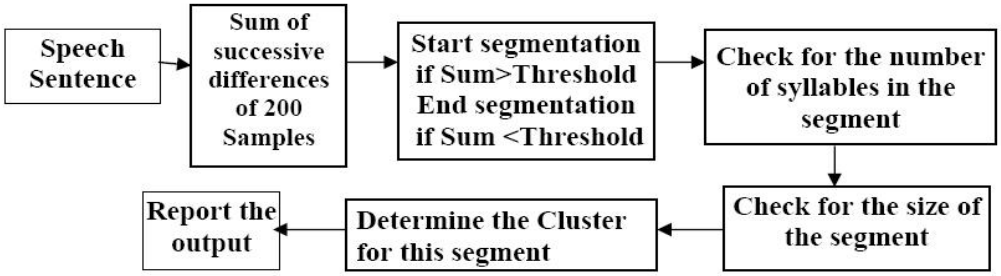


Fig. 5 The block diagram of the developed software system

5 System Development

The above methodology was implemented by software developed in C language. The modular design of the developed software system is shown in Fig.5.

In the segmentation program, the summation of the absolute differences of successive samples over a frame of 200 samples was compared with threshold. When the sum was exceeded the threshold the samples were started to store in a buffer and the process end when the sum was less than the threshold. In the clustering program, the syllables were traced in a similar fashion except that the number of samples for summation was only 25. In the second level of clustering, the number of samples was counted and according to this count a number (1 to 24) that identify a cluster was reported as output.

6 Experimental Results

According to the result of our study, a threshold of 50 was produced good result for segmentation. The percentage of accuracy rate and the failure rate of segmentation had been calculated using the following equation.

$$Segmentation\ accuracy = \frac{No.of\ correct\ segments\ \ obtained}{No.of\ correct\ segments\ \ \exp ected} \times 100\%$$

Table 3. Segmentation results

Speaker ID	No. of Correct Segments Expected	No. of Correct Segments Obtained	No. of failure	Accuracy (%)	Failure (%)
S1	758	745	13	98.28	1.72
S2	758	744	14	98.15	1.85
S3	758	742	16	97.89	2.11
S4	758	748	10	98.68	1.32
S5	758	750	8	98.94	1.06
S6	758	750	8	98.94	1.06
Average Rate (%)				**98.48**	**1.52**

In this experiment, 120 spoken sentences were allowed to segment by the system. The system output was words and sub-words. The detailed segmentation results are given in Table 3.

The segmented words were the input of the clustering program. The program received the segmented words as input and produced three different clusters of words as output, based on the number of syllables. The speech words, which have no gap between two successive syllables, were considered as single-syllable words. In the second level, words from three main clusters were input to the system and the system was distributed them among eight different sub-clusters, based on the word lengths. After clustering, 24 different clusters were obtained. The detailed clustering results are given in Table 4 and Table 5.

Table 4. Syllable-based clustering results

Speaker Id	Cluster #1 (No. of words)	Cluster #2 (No. of words)	Cluster #3 (No. of words)	Total no. of words
S1	295	301	108	704
S2	329	249	126	704
S3	277	292	135	704
S4	294	284	126	704
S5	339	285	80	704
S6	325	247	132	704
Total	**1859**	**1658**	**707**	**4224**

7 Discussion

In this research, the main goal was to develop a system that automatically segments continuous Bangla speech and clusters speech segments into some predefined clusters. It is seen in table-3, the average segmentation accuracy rate is 98.48% and it is quite satisfactory. The syllable-based and length-based clustering results are shown in

Table-4 and Table-5 for all segmented words. It was 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, and emotion, and environment.

Table 5 (a) Length-based clustering results for Cluster #1

Speaker Id	Cluster #1								Total no. of words
	Sub #1	Sub #2	Sub #3	Sub #4	Sub #5	Sub #6	Sub #7	Sub #8	
S1	0	14	75	124	65	11	4	2	**295**
S2	2	28	111	114	53	16	3	2	**329**
S3	1	48	131	71	22	3	1	0	**277**
S4	2	52	113	85	35	7	0	0	**294**
S5	2	47	135	107	36	10	2	0	**339**
S6	4	53	116	118	27	6	1	0	**325**
Total	**11**	**242**	**681**	**619**	**238**	**53**	**11**	**4**	**1859**

Table 5 (b) Length-based clustering results for Cluster #2

Speaker Id	Cluster #2								Total no. of words
	Sub #1	Sub #2	Sub #3	Sub #4	Sub #5	Sub #6	Sub #7	Sub #8	
S1	0	1	18	71	127	69	13	2	**301**
S2	0	0	34	106	73	25	10	1	**149**
S3	0	7	63	125	70	23	4	0	**292**
S4	0	5	46	122	78	30	3	0	**284**
S5	0	9	87	111	45	26	6	1	**285**
S6	0	13	62	102	53	15	3	0	**247**
Total	**0**	**35**	**310**	**637**	**446**	**188**	**39**	**4**	**1658**

Table 5 (c). Length-based clustering results for Cluster #3

Speaker Id	Cluster #3								Total no. of words
	Sub #1	Sub #2	Sub #3	Sub #4	Sub #5	Sub #6	Sub #7	Sub #8	
S1	0	0	1	4	23	49	22	9	**108**
S2	0	0	6	26	48	32	13	1	**126**
S3	0	0	5	30	52	29	15	4	**135**
S4	0	0	6	19	47	41	11	2	**126**
S5	0	0	13	23	28	9	4	3	**80**
S6	0	0	8	35	57	23	5	4	**132**
Total	**0**	**0**	**39**	**137**	**255**	**183**	**70**	**23**	**707**

8 Conclusion

We have described a novel approach for segmenting and clustering continuous Bangla speech. Perfect segmentation method is an unavoidable prerequisite for the development of a continuous speech recognition system. From the result of our experiment it can be concluded that the proposed approach for segmenting and clustering is very efficient. Thus, the developed system may be used to develop large vocabulary continuous speech recognition system. 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 recognition tools.

References

[1] Abul Hasanat, Md. Rezaul Karim, Md. Shahidur Rahman and Md. Zafar Iqbal, "Recognition of Spoken letters in Bangla", 5th ICCIT 2002, East West University, Dhaka, Bangladesh, 27-28 December 2002.

[2] S. M. Jahangir Alam, an M.Sc. Thesis on "System Development for Bangla Phoneme Recognition", Dept. of Computer Science & Engineering, Islamic University, Kushtia-7003, July-2004.

[3] Kaushik Roy, Dipankar Das and M. Ganjer Ali, "Development of the Speech Recognition System using Artificial Neural Network", 5th ICCIT 2002, East West University, Dhaka, Bangladesh, 27-28 December 2002.

[4] Md. Farukuzzaman Khan, "Computer Recognition of Bangla Speech", M.Phill. Thesis, Dept. of Computer Science and Engineering, Islamic University, Kushtia, September, 2002.

[5] Md. Farukuzzaman Khan, Md. Mijanur Rahman, Md. Mostafizur Rahman, "Development of Bangla Voice Command Driven DOS Utility System", Journal of Applied Science & Technology, Islamic University, Kushtia-7003, Bangladesh, Vol. 03, No. 02, p93-98, December-2003.

[6] Md. Saidur Rahman, "Small Vocabulary Speech Recognition in Bangla Language", M.Sc. Thesis, Dept. of Computer Science & Engineering, Islamic University, Kushtia-7003, July-2004.

[7] Md. Rabiul Huq, "A medium vocabulary speech to text system", M. Sc. Thesis, Dept. of Computer Science & Engineering, Islamic University, Kushtia-7003, February-2005.

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

[9] R. G. Chen, "Autoatic *segmentation techniques for Mandarin speech recognition*," M. S. Thesis, Department of Electrical Engineering, National Taiwan University, 1978.

[10] Md. Farukuzzaman Khan and Dr. Ramesh Chandra Debnath, "*Bangla Sentence Recognition Using End-Point Detection Technique*", Rajshahi University Studies, Rajshahi University, 2000.

[11] 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.

[12] Earl Gose, R. J. Baugh and Steve Jost, "Pattern Recognition and Image Analysis", Prentice-Hall of India, 1997.

Md. Mijanur Rahman is working as a Lecturer of the department of Computer Science and Engineering in Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh. He served as an Instructor (Tech.) Computer in Govt. Polytechnic Institute under the directorate of Technical Education, 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 in the department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh.

Md. Farukuzzaman Khan is working as an Associate Professor of the department of Computer Science and Engineering in Islamic University, Kushtia, Bangladesh. He completed his B Sc (Hons) and M Sc degree from Rajshahi University, Rajshahi, Bangladesh. He is a PhD researcher in the department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh.

Mohammad Ali Moni is working as a Lecturer of the department of Computer Science and Engineering in Pabna University of Science and Technology, Pabna, Bangladesh. He served as a Lecturer 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.