DEVELOPMENT OF ISOLATED SPEECH RECOGNITION SYSTEM FOR BANGLA WORDS

Md. Mijanur Rahman[1] and Fatema Khatun[2]

[1]Dept. of Computer Science and Engineering
Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh, Bangladesh.
[2]Dept. of Electronics and Communication Engineering
Institute of Science, Trade and Technology (ISTT), Dhaka, Bangladesh.

Email: mijan_cse@yahoo.com, fatema_aece@yahoo.com

Abstract: *This research devoted to the development of Speech Recognition System in Bengali language that works with speaker independent, isolated and subword-unit-based approaches. In our work, the original Bangla speech words were recorded and stored as RIFF (.wav) file. Then these words were classified into three different groups according to the number of syllables of the speech words and these grouping speech signals were converted to digital form, in order to extract features. The features were extracted by the method of Mel Frequency Cepstrum Coefficient (MFCC) analysis. The recognition system includes direct Euclidean distance measurement technique. The test database contained 600 distinct Bangla speech words and each word was recorded from six different speakers. The development software is written in Turbo C and common feature of today's software have been included. The development system achieved recognition rate at about 96% for single speaker and 84.28% for multiple speakers.*

Keywords: *MFCC, Syllable-based grouping, Speaker independent, End-point detection and Euclidian distance.*

1. Introduction

Speech and music are the most basic means of adult human communication. 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 humankind. Such research has led to the development of speech and music synthesizers, speech transmission systems, and automatic speech recognition systems. In computer speech recognition, a person speaks over a microphone or telephone and the computer listens. Then the computer simply attempts to transcribe the speech into the text. 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] or small vocabulary continuous speech [5].

Most speech recognition systems can be classified according to the following categories [6]: (a) *Speaker Dependent vs. Speaker Independent*, a speaker-dependent speech-recognition system is one that is trained to recognize the speech of only one speaker, while a speaker-independent system is one that is trained such that anyone can use it; (b) *Isolated vs Continuous Speech Recognition*, in isolated speech, the speaker pauses momentarily between every word, while in continuous speech the speaker speaks in a continuous and possibly long stream, with little or no breaks in between; (c) *Keyword-based vs. Subword-unit-based, a* speech recognition system can be trained to recognized whole words, like "dog" or "cat" and another approach would be to train the recognition system recognize sub-word units like syllables or phonemes. In this paper, we have tried to represent a Bangla speech recognition system that works with speaker independent, isolated and subword-unit-based approaches.

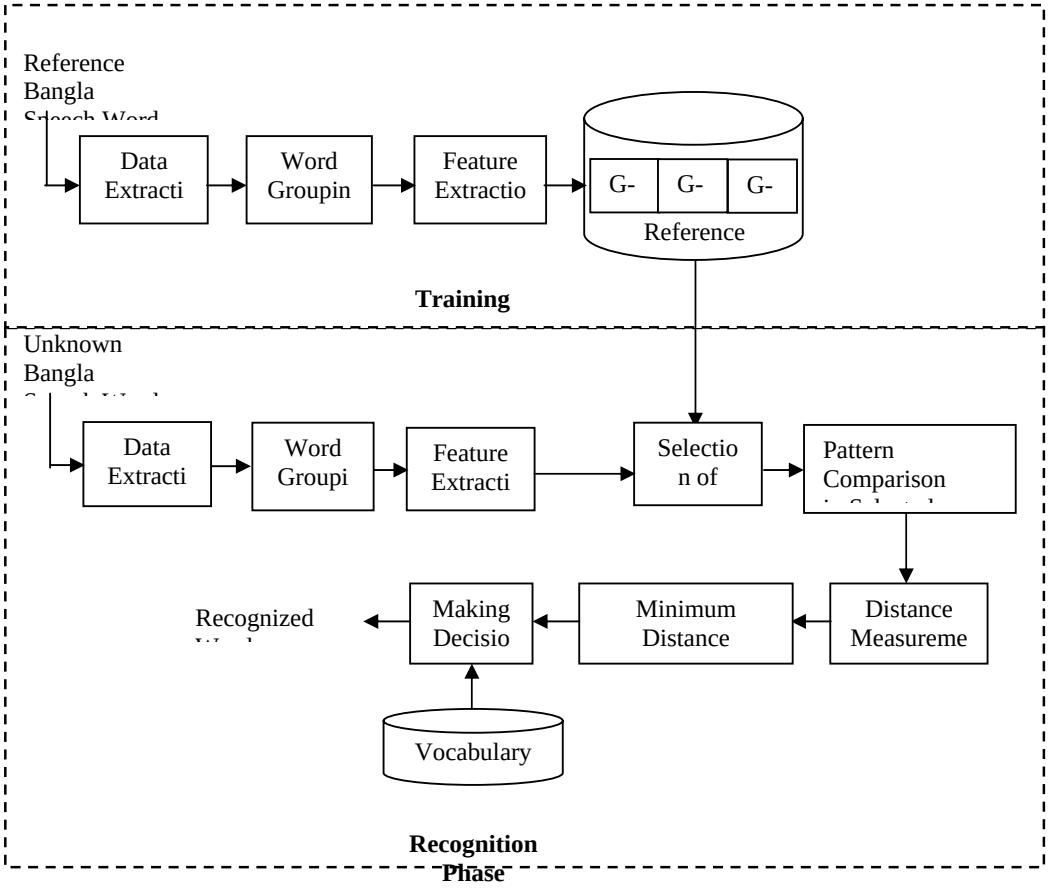
2. Methodologies

The complete recognition system for isolated Bangla speech words is shown in Figure-1. The individual steps are discussed in the following sub-sections.

2.1 Speech Acquisition

The recording of Bangla speech words was completed in a sound proof laboratory environment with the help of close-talking microphone, high quality sound card and sound recorder software. The 600 Bangla words originated from six speakers were recorded as *wav* file to make a sample database. Therefore, the reference database contained totally 3600 Bangla speech words. The utterances were recorded at a sampling rate of 8.00 KHz and coded in 8 bits PCM[7].

Date of submission : 01.06.2010 Date of acceptance : 10.10.2010



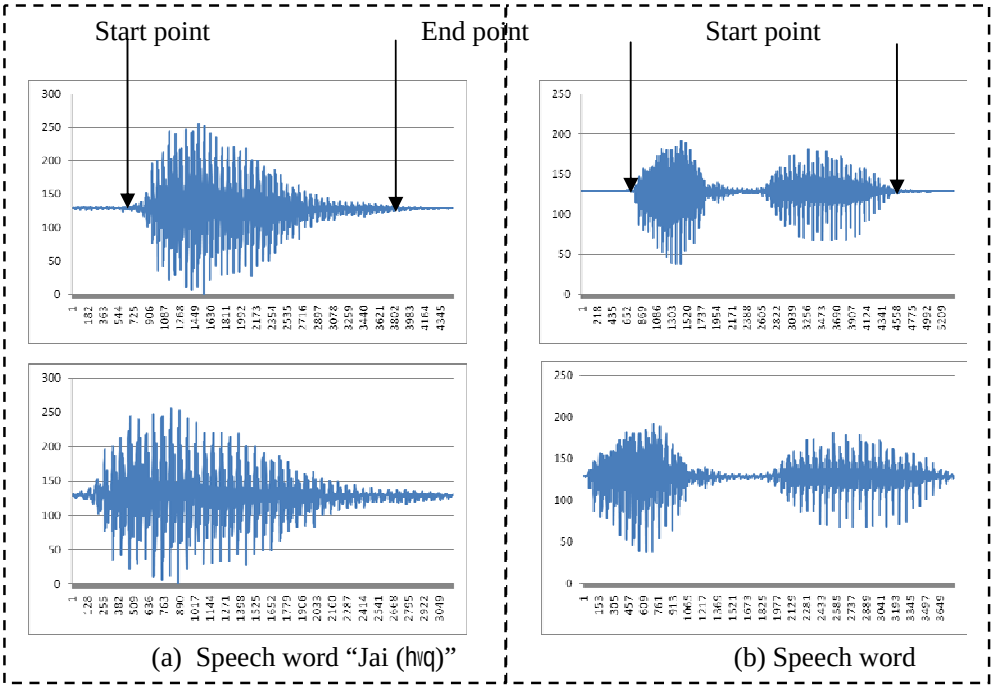


Fig. 1: Detection of start and end points of Bangla speech words.

2.2 Wave data Extraction

To extract wave data, we first discard 58 bytes (file header) from the beginning of the wave file and then read wave data as character [8]. The data extraction process extracts require voiced data from the input speech signal, which may contain silence, unvoice and voice. This data are stored in a text file as integer data.

This is usually done by detecting the proper start and end points of the speech events (voicing and unvoicing) and then separated into different pieces containing the audio signals on the basis of the detected start and end points [9], as shown in Figure -2. Proper data extraction ensures better extraction of speech feature, which in turn results in better recognition accuracy.

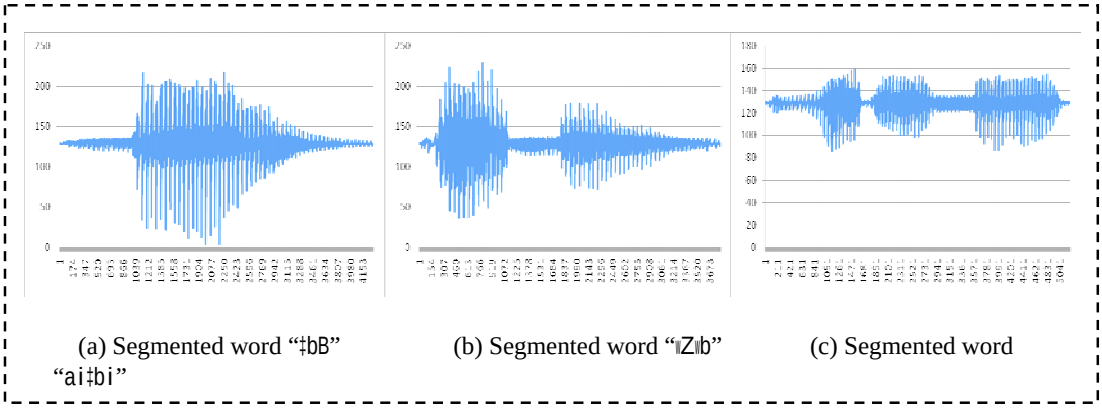
2.3 Grouping of Words

Grouping means collection of spoken words and sub-words into different groups based on some properties. It is very important for medium and large vocabulary speech recognition system. It increases recognition speed and accuracy. In this research, an effort was made to categorize the speech words according to the number of syllables of spoken words, which is known as syllable-based grouping [6]. According to our study three different groups were formed, as shown in Table 1 and Fig. 2 shows the examples of grouping words.

Grouping is a very difficult task for speech recognition, because the same words of speech may vary from speaker to speaker. This is caused by non-uniform articulation of speech [10]. Sometimes it is difficult to maintain the uniformity in articulation for the same speech of the same speaker. The size also varies depending on the properties of the speaker, such as age, sex and emotion. Because of the grouping complexities, all same words and sub-words may not fall in the same group for all speakers. So, we have performed a union operation among the same groups of all speakers and made final reference pattern for this group.

Table 1**:** *Syllable-based Grouping*

Group Name	Contents
Group1 (G-1)	Mono-syllabic words
Group2 (G-2)	Di-syllabic words
Group3 (G-3)	Tri or more syllabic words



(a) Segmented word “‡bB” “ai‡bi” (b) Segmented word “\Z\b” (c) Segmented word

Fig. 2: Example of grouping words

2.4 Feature Extraction

The greatest important part of all recognition systems is the feature extraction, which converts the speech signal to some digital form of meaningful features. Obviously, a good feature may produce a good result for any recognition system. Feature extraction is the combination of some signal processing steps including frame blocking, preemphasis, windowing and the computation of Mel Frequency Cepstrum Coefficient (MFCC), as shown in Figure -4.

At first, each speech word was segmented in a set of samples, called frame that representing typically 16 to 32 ms of speech. Preemphasis compensates for the negative spectral slope of the voiced portions of the speech signal. A typical signal preemphasis is given by $y(n) = s(n) - C \times s(n-1)$, where C is the preemphasis constant generally falls between 0.9 and 1.0 [11].

Windowing of speech signal involves multiplying a speech signal by a finite-duration window. One of the most popular windows used in speech recognition is the Hamming window defined by the following equation:

$$h(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad .....(0 \le n \le N-1)$$

$$= 0, \quad otherwise$$

where N is the window length [11].

Now the preprocessed speech signal is passed through some computational steps to extract a set of features that represents Mel Frequency Cepstrum Coefficients (MFCC) of the signal. The computation steps of MFCC including Discrete Fourier Transform (DFT), computation of first two formant frequencies, Mel frequency warping, Discrete Cosine Transform (DCT) and finally the computation of Mel Frequency Cepstrum Coefficient (MFCC), as shown in Figure – 5 [12][13].

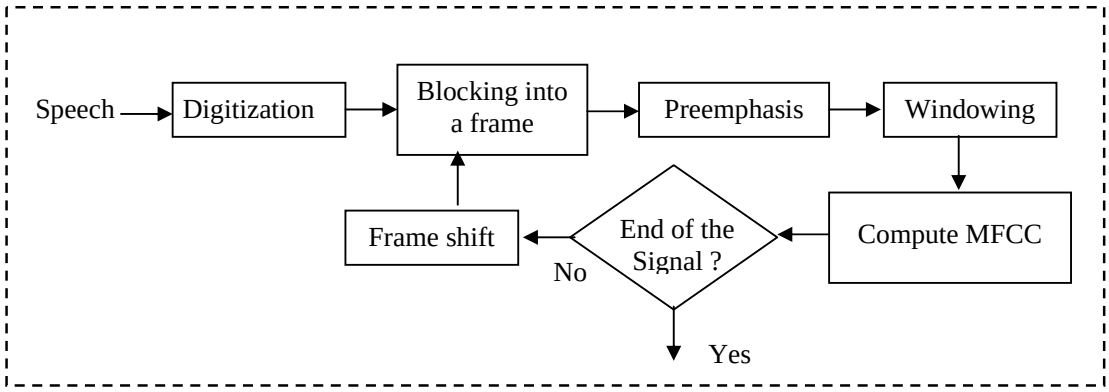


Fig. 3: Feature extraction process.

Table 2: Grouping results

Speaker ID	Group-1 (No. of Words)	Group-2 (No. of Words)	Group-3 (No. of Words)	Total No. of Words
S1	252	256	92	600
S2	260	232	108	600
S3	236	249	115	600
S4	251	242	107	600
S5	259	243	98	600
S6	257	230	113	600
Total	**1515**	**1452**	**633**	**3600**

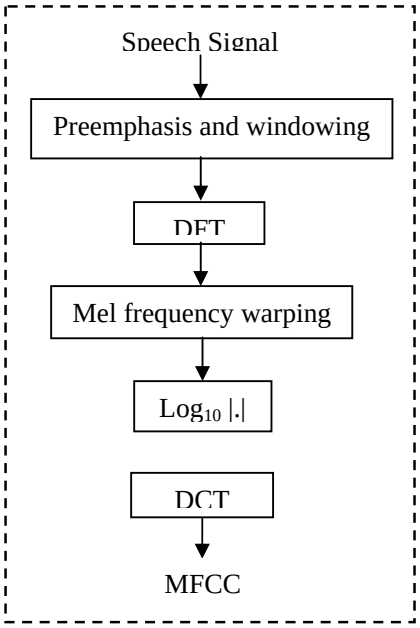


Fig. 4: Calculation of MFCC.

2.4 Speech Recognition Process

Pattern recognition is concerned with the automatic detection or classification of objects [14]. In this research, a direct comparison of the unknown speech (the speech to be recognized), with each possible reference pattern stored in the training phase and classifies the unknown speech according to the goodness of match of the patterns. The process finds the best match between the test pattern and the reference patterns. The method has two steps- namely, training of speech patterns, and recognition of patterns via pattern comparison. Several distance measurement techniques are used in pattern comparison. For simplicity, the Euclidean distance measurement technique was used to compare the test and reference patterns in this research.

3. Experimental Results

This research was aimed to develop a system to recognize speech words from a reference database. The database contains totally 3600 prerecorded Bangla speech words which were classified into three different groups. The detailed grouping result is given in Table 2.

In the recognition phase, the syllable of unknown speech word was checked and then the corresponding group was selected from the reference database. The speech words, which have no gap between two successive syllables, were considered as mono-syllabic words included in Group-1 (G-1) and so on. With the help of Euclidean distance measurement technique, the best match between the unknown pattern and the group patterns was determined and hence the decision was made.

The detailed recognition result is shown in Table 3 and the graphical representation of percentage recognition accuracy is shown in Fig. 7.

Table 4: Recognition results

No. of speakers	No. of words in database	No. of test words	No. of accurately recognized words	Recognition rate (%)
1	600	600	576	96.00
2	1200	1200	1122	93.50
3	1800	1800	1615	89.72
4	2400	2400	2022	84.25
5	3000	3000	2436	81.20
6	3600	3600	2848	79.11
Total		**12600**	**10619**	**84.28**

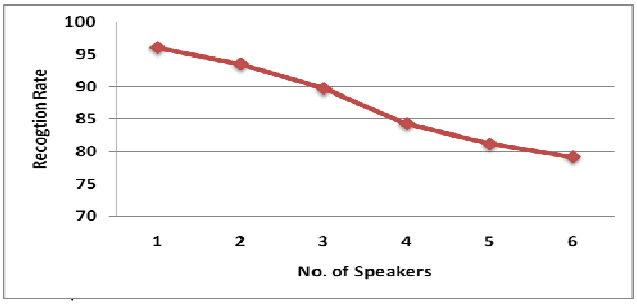


Fig. 5: Recognition rate vs number of speakers

4. Discussion

In this research the main goal was to develop system for speech recognition in Bangla Language. The feature selection and grouping of words are of 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 grouping of words enhances the recognition rate. Among the different distance measurement technique the Euclidean distance measurement technique is simple in computation and produces very good results. The table 5.1 shows that the average recognition accuracy is 84.28% with highest rate of 96%.

All of these tests were conducted with six different speakers from different age group. During speaker verification it was observed that personal speaking habit or style changes the sound of a speech. Speeds of utterance, loudness variation were also the sources of errors. Characteristics of microphone, other recording instruments and environment also affect the result. These problems may be eliminated if the speakers were phonetically trained, recording instruments should have constant settings and the environment should be noise free.

5. Conclusion

Although the developed system produces reasonable results for isolated words, it may develop a recognition system using continuous speech signals. The system did not employ any knowledge (syntactic or semantic) of linguistics. Inclusion of such knowledge will increase the recognition performance. For syllable-based grouping constant thresholds have been used. If we could use dynamic threshold for grouping it might produce more accurate grouping, which in turn will produce better recognition results. Future work must be able to handle the variability in loudness, speed and noise. An efficient system should be fully speaker-independent. So the future researchers should employ speakers of different ages and genders. Future system should also employ more powerful recognition tools like Gaussian Mixture Model (GMM), Time-Delay Neural Network (TDNN) and the Hidden Markov Model (HMM) to improve the system performance.

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] Md. Farukuzzaman Khan, Md. Mijanur Rahman and Md. Mostafizur Rahman, "*Development of Bangla Voice Command Driven DOS Utility System*", Journal of Aplied Science and Technology, Islamic University, Kusgtia, Bangladesh, Vol 03, No 02, P93-98, December 2003.

[4] 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.

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

[6] Tan Keng Yan, Colin, A thesis on "*Speaker Adaptive Phoneme Recognition using Time Delay Neural Network*", Computer & Information Science, National University of Singapore, 2000.

[7] S. Gokul, "Multimedia Magic", BPB Publications, B-14, Connaught Place, New Delhi-110001, ISBN 81-7029-972-1.

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

[9] Dr. Ramesh Chandra Debnath and Md. Farukuzzaman Khan, "*Bangla Sentence Recognition Using End-Point Detection*", Rajshahi University Studies: Part B, Journal of Science, Vol 32, 2004.

[10] Prabhu Raghavan, "*Speaker And Environment Adaptation In Continuous Speech Recognition*", Technical Report CAIP-TR-227, The State University of New Jersey, Piscataway, New Jersey 08855-1390, June, 1998.

[11] Jean-Claude Junqua & Jean-Paul Haton, "*Robustness in Automatic Speech Recognition: Fundamentals and Applications*", Kluwer Academic Publishers, Dordrecht, Netherlands, 1997.

[12] F. Jelinek, L. R. Bahl, and R. L. Mercer, "*Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech*", IEEE Trans. Information Theory, IT-21, pp 250-256, 1975.

[13] 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.

[14] Earl Gose, Richard Johnson Baugh, Steve Jost, "*Pattern Recognition and Image Analysis*", Prentice-Hall of India Private Limited, New Delhi-110001, 2002.