Connected Bangla Speech Recognition using Artificial Neural Network

Khalil Ahammad
M.Sc. Student
Dept. of Computer
Science& Engineering
Comilla University
Bangladesh

Md. Mahfuzur Rahman
Assistant Professor
Dept. of Computer
Science& Engineering
Comilla University
Bangladesh

ABSTRACT
There are a plenty of research experiments and achieved results in various languages throughout the world regarding speech recognition. But, in Bangla language, early researchers in this field had qualified success, though the scenario is being changed in recent years. This research work aims at developing a neural network based connected digit recognition system in Bangla language. Firstly, a Bangla digit corpus has been developed comprising of male and female speakers. Speech is recorded in connected fashion and words are extracted through automatic segmentation. Then MFCC features of the segmented words are calculated and these feature values are sent as the input to the back-propagation neural network (BPNN). BPNN learning algorithm is used to train the network. The required time to train the network, number of hidden layers, error threshold and number of epochs are considered while training the network to reach the best possible recognition accuracy. This proposed system has been implemented using object oriented programming and the achieved recognition accuracy is very much satisfactory and consistent. The network has been tested for three different setups and the best recognition accuracy achieved for digit dataset is 98.46%.

Keywords
Connected Speech Recognition, Bangla Speech Recognition, Bangla Digits, Back-Propagation Learning Algorithm, MFCC.

1. INTRODUCTION
To communicate with each other in real-life, speech is probably the most efficient way. It is also possible to use speech as a useful interface to interact with machines. Consequently, speech recognition research has evolved from laboratory demonstrations to copious real-life applications. But, research on human-computer interaction using Bangla language is still inadequate as for the understanding of the spoken form of this language, a systematic and scientific effort has not been started yet [1].

Speech recognition system can be classified based on the continuity of utterances of speech contents: isolated, connected, continuous and spontaneous speech recognition systems [2]. Recognition systems are also classified based on speaker variability. Each language has its own phonemes [3]. They differ significantly from one another. So it cannot be ensured that, a system that provides good recognition accuracy for English language will also provide good recognition accuracy for Bangla language [3]. Though speech recognition research work has been started since 1930 [4], but in regard to Bangla speech recognition, it's noteworthy that several works have been done starting from 2000 [1], [5] and they all worked on isolated word recognition. For Isolated word recognition, the assumption is that the speech to be recognized comprises of a single word or phrase and to be recognized as a complete entity with no explicit knowledge or regard for the phonetic content of the word or phrase [6]. Connected and continuous speech recognition in Bangla language is still a way to go. Moreover, there is no standard Bangla corpus that can be used for Bangla speech recognition. The previous works accomplished mostly suffer the constraints of speaker dependence. In this work, an approach to connected speech recognition has been attempted. To accomplish this work, a Bangla corpus has been developed that ensures speaker variability.

A connected word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system maintains continuity of utterance which is representative of real speech [6], [7], [8], [9]. On the other hand a sentence constructed from connected words does not represent real speech as it is actually a concatenation of isolated words. Notwithstanding that it has some limitations in continuity of speech; it has some real benefits as it can be used in automatic dialing of phone calls.

In this research study, an exertion is made to develop Bangla connected speech recognizer using artificial neural network and Mel Frequency Cepstrum Coefficient (MFCC) features. For weight adjustments in BPNN, gradient descent method has been used in this proposed system. Overview of the system is presented in Fig-1.

2. METHODOLOGICAL STEPS
The methodological steps adopted for Bangla connected speech recognition system using MFCC features and Neural Network are as follows:

2.1 (Step- 01): Speech Acquisition & Dataset Development
Speech acquisition is the process of acquiring digital audio data from the analog audio data uttered by speakers. Speech acquisition is done with the following recording parameters:

1. Encoding Format: Highest quality (PCM/WAV)
2. Sampling Rate: 8kHz
3. 8 bits, mono-channel.
4. Environment: Noise-free

2.1.1 Development of Bangla Speech Corpus
In this system, two datasets of connected digits in Bangla were used, respectively for training and test, comprising of 15 male and 15 female speakers. Both training and test datasets have been acquired solely for the purpose of this work. A brief acquaintance of Bangla digits and their pronunciation is presented in Table-1.

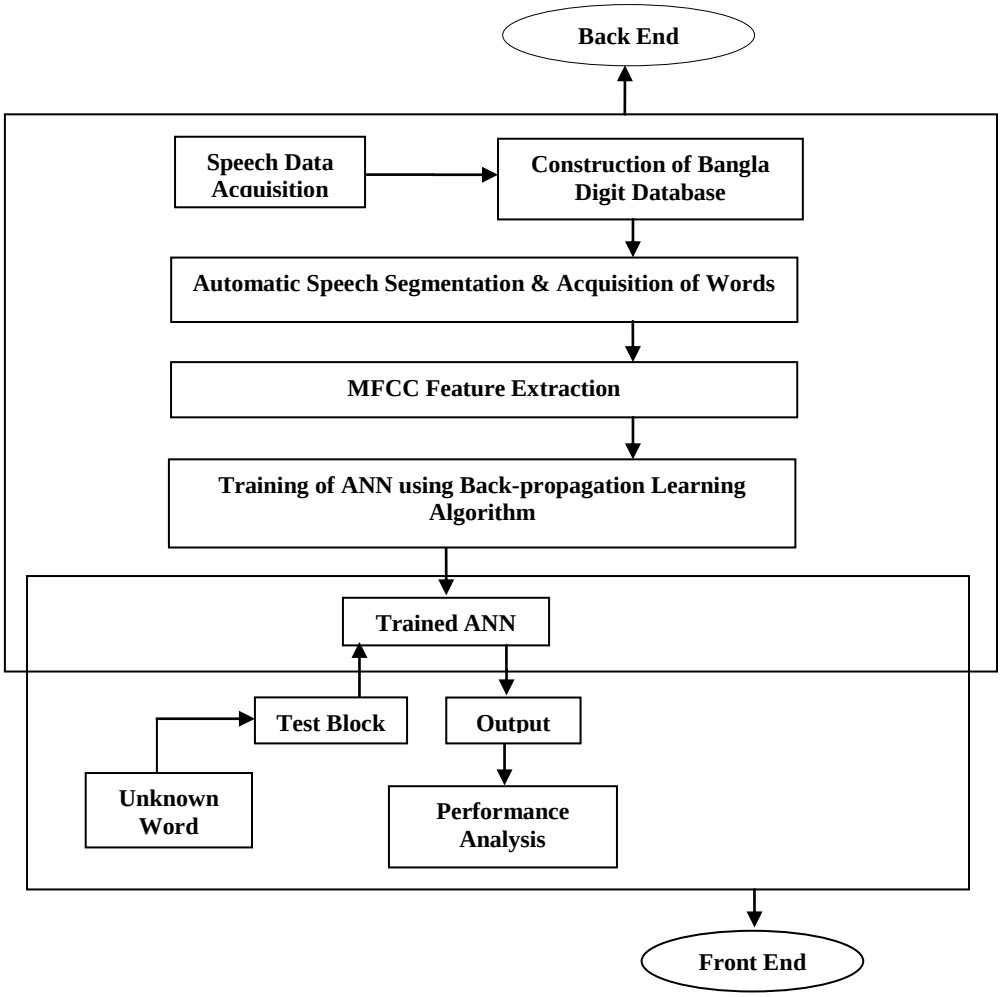

Fig-1: Overview of the System.

Table-1: Bangla digits and their pronunciation.

English Digit	*Bangla Digits*	*Pronunciation*
0 (Zero)	০ (শূন্য)	SHUNNO
1 (One)	১ (এক)	EK
2 (Two)	২ (দুই)	DUI
3 (Three)	৩ (তিন)	TIN
4 (Four)	৪ (চার)	CHAR
5 (Five)	৫ (পাঁচ)	PACH
6 (Six)	৬ (ছয়)	CHOY
7 (Seven)	৭ (সাত)	SHAT
8 (Eight)	৮ (আট)	AAT
9 (Nine)	৯ (নয়)	NOY

2.2 (Step-02): Preprocessing & Segmentation

Connected digits are subject to segmentation to acquire isolated words. A program is developed, based on trial and

error method, which can dynamically segment the intended digits and name them accordingly without any extraneous intervention.

2.3 (Step-03): Feature Extraction

The most important and common part of all recognition systems is the signal processing front-end, which converts the speech waveform to some type of parametric representation [10]. This parametric representation is then used for collection of meaningful features. The proposed system used Mel Scale Cepstral Coefficient analysis, one of the popular signal analysis techniques, to elicit meaningful features for recognition.

For each segmented digit, a MFCC feature file is generated that contains a set of floating point values. These values are used as key parameters for training of the ANN. For MFCC computation, standard approach has been followed incorporating framing, pre-emphasis, windowing, FFT and DCT.

2.4 (Step-04): Development of Neural Network

Artificial neural network is the chosen classifier for this proposed research work. The number of input neurons is selected depending on the number of features extracted for each digit. Several combinations of learning rate, number of hidden layers, error thresholds are implemented to improve accuracy. BPNN learning algorithm is used for the system.

2.4.1 Experimental Setup and Recognition Results

BPNN, as implemented in this work, uses 352 neurons in the input layer keeping it in accord with number of feature parameters in MFCC file. An optimized ANN has been obtained through minimization of errors.

The recognition accuracy is calculated using the following formula.

$$\textbf{Accuracy} = \frac{\textbf{Total recognized words}}{\textbf{Recognized} + \textbf{unrecognized words}} \times 100\%$$

Three experiments were conducted each with different settings. Summary of recognition results of the experiments is shown in Table-2 and details of digit recognition are shown in Fig.2.

Table-2: Details of Experiments and Recognition Results.(HL for No. of Hidden Layers, ET for Error Threshold, LR for Learning Rate)

Exp. No.	Experimental Parameters			No. of Digits	Recognized Digits	Overall Recognition Accuracy (%)	Training Complexity	
	HL	ET	LR				Training Time (minute)	No. of Epochs
Exp-1	30	6×10^{-7}	0.3	260	249	88.84%	22	2700
Exp-2	45	6×10^{-7}	0.3	260	256	98.46%	15	1470
Exp-3	50	6×10^{-7}	0.3	260	214	82.31%	9	776

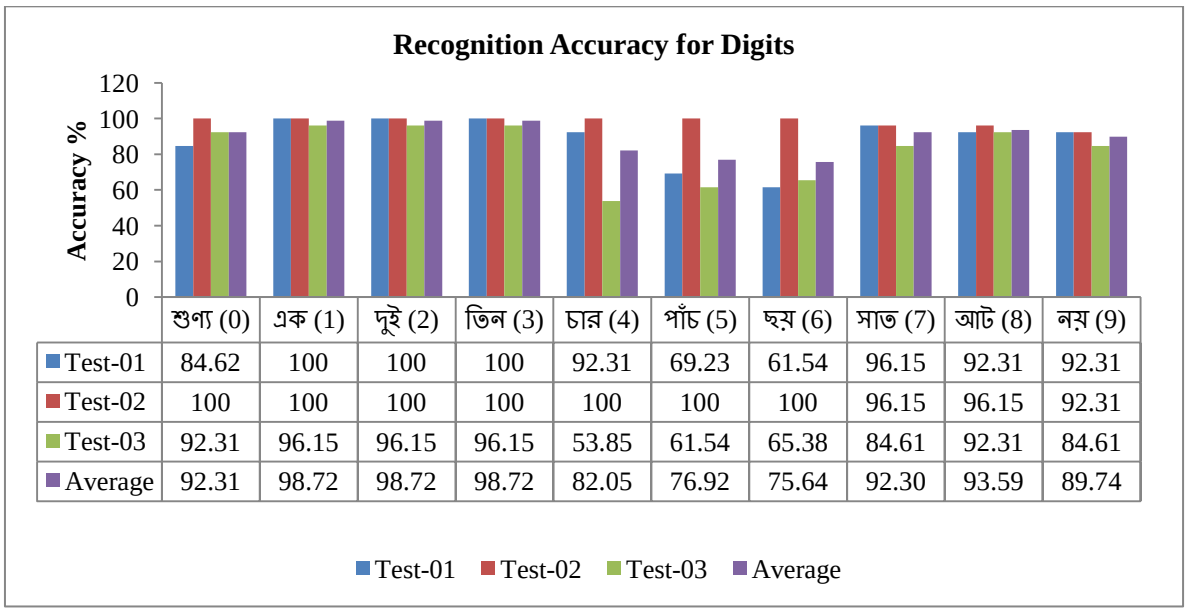

Fig 2: Recognition Accuracy for Digits

3 EXPERIMENTAL RESULT ANALYSIS

It has been marked that among the experiments conducted in this work, experiment-2 outperforms other experiments and this setup shows persistent performance even for different digits-intimating that number of hidden layers required for best performance should be chosen by trials. It doesn't necessarily depend on any predefined formula. Conspicuous deterioration of recognition results for some digits has been noticed. This is due to several Bangla digits are conflicting while they are pronounced, e.g. "CHOY (6)", "NOY (9)" and "SHAT (7)", "AAT (8)" are phonetically very close to their counterparts. Consequently, the average recognition accuracy for some digits like PACH (5)-76.92%, CHOY (6)-75.64%, CHAR (4) - 82.1% is a bit far from expectation due to phonetic similarity in Bangla language whereas some satisfactory performance has been noticed for EK (1) - 98.72%, DUI (2) - 98.72%, TIN (3) - 98.72% as they are quite different when they are pronounced. As long as Test-2 is concerned, it is also noticeable that this deterioration of recognition accuracy for phonetic variability can be eliminated as it is justified for CHAR (4), PACH(5),CHOY (6) as 100% recognition accuracy for these words has been achieved through proper setup of experimental parameters.

4 CONCLUSION

In this work, an automatic recognizer of connected Bangla digits has been developed using BPNN and MFCC feature extraction method. Neural networks are very sensitive classifiers. A small amount of changes in the network architecture may cause significant change in the output. One of the major goals of this research experiment was to optimize the network based on network parameters such as number of hidden layers, learning rate, error threshold and number of epochs. The best recognition accuracy achieved was 98.46%. It is also evident from the result is that the recognition accuracy varies over digits due to their phonetic traits. To acquire persistent performance over digits, BPNN based recognition system can be improved either by employing hybrid classifier and/or incorporating robust features. The current recognition system is desired to be updated using genetic algorithm rather than gradient descent method of weight adjustments.

5 REFERENCES

[1] Md. Ali Hossain, Md. Mijanur Rahman, Uzzal Kumar Prodhan, Md. Farukuzzaman Khan "Implementation Of Back-Propagation Neural Network For Isolated Bangla Speech Recognition", IJIST-Vol.3, No.4, July 2013

[2] Pratik K. Kurzekar , Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, "Continuous Speech Recognition System: A Review", Asian Journal of Computer Science And Information Technology 4 : 6 (2014) 62 – 66

[3] S. A. Hossain, M. L. Rahman, F. Ahmed & M. Dewan, "Bangla speech synthesis, analysis, and recognition: an overview", Proc. NCCPB, Dhaka, Bangladesh-2004.

[4] K. J. Rahman, M. A. Hossain, D. Das, A. Z. M. Touhidul Islam and Dr. M G. Ali, "Continuous Bangla Speech Recognition System", Proc. 6th Int. Conf. on Computer and Information Technology (ICCIT), Dhaka, 2003.

[5] M. R. Hassan,B. Nath &M. A. Bhuiyan, "Bengali phoneme recognition: a new approach", Proc. 6th International Conference on Computer and Information Technology, Dhaka, Bangladesh-2003.

[6] Md. Abul Hasnat, Jabir Mowla, Mumit Khan, "Isolated and Continuous Bangla Speech Recognition: Implementation, Performance and application perspective", Conference Papers (Centre for Research on Bangla Language Processing) Department of Computer Science and Engineering, BRAC University, Bangladesh, 2007.

[7] ParneetKaur, Parminder Singh, VidushiGarg, "Speech Recognition System; Challenges and Techniques", International Journal of Computer Science and Information Technologies, Vol. 3 (3), 2012, 3989-3992.

[8] V.A. Keturi, "Speech Recognition Based on Artificial Neural Networks", Helsinki University of Technology, Finland, 2004.

[9] Joseph Picone,"Continuous Speech Recognition Using Hidden Markov Models ", IEEEASSP MAGAZINE JULY 1990.

[10] Lawrence Rabiner, Biing Hwang Juang, B. Yegnanarayana,"Fundamentals of Speech Recognition", Pearson Education India, 2008 -page-60