# Isolated Bangla Word Recognition and Speaker Detection by Semantic Modular Time Delay Neural Network (MTDNN)

Md. Yasin Ali Khan, S. M. Mostaq Hossain and Mohammed Moshiul Hoque
Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Chittagong-4349, Bangladesh

***Abstract*— Speaker recognition is the identification of a person from characteristics of his/her voices and speech recognition concerns the recognizing of what is being said by the speaker. This paper presents a framework to recognize the isolated Bangla words and the corresponding speaker by proposing a semantic modular time delay neural network (MTDNN). Underlying acoustic fuzziness of human utterance and fluctuations of data due to environmental disturbance are managed by well-known Fuzzy C Means clustering technique. We have used MFCC features to recognize Bangla words and speaker detection. Experimental result with different individuals show that the proposed framework is functioning quite satisfactory with average accuracy of 82.66%.**

**Keywords**: Speech recognition, modular architecture, recurrent neural network, speaker detection.

## I. INTRODUCTION

Scientifically speech recognition is the process of converting an acoustic waveform to a suitable written equivalent of the message or information. Speaker recognition on the other hand is the task of recognizing and separating internal frequency spectrum using some processing such that intelligent agent can differ every subject accurately. One of the fundamental capabilities of human is to use natural language to express their emotions to transmit and receive information. Effective use of language is intertwined with our general cognitive abilities. Evolution makes brain an efficient dynamic structure that can learn quickly and can generalize its knowledge dramatically. An interesting field of computer science is to use this intuitive capability of human as input to computing process and transform it into different required outputs. It is difficult for computing system for recognizing speech and speaker as for higher dimensional data values, robustness and variability embedded within data, failure of non-machine learning based approach and need for huge processing speed with great memory capability.

Three types of recognition system are found in literature: speaker recognition (speaker verification and speaker identification [1], language recognition [2] and speech text recognition [3]. To keep pace with the world on different technologies like document preparation or retrieval, command and control, data entry, automated customer service and voice-control of cars and machines, robotics and many more, building computing system on Bangla Speech Recognition is mandatory. We have developed a deep hierarchical NN architecture for speaker independent detection and Bangla word recognition. For Bangla language speech and speaker recognition we have chosen modular technique. For recognition purpose, we have to use neural network. Modular approach combined with neural network achieved a higher recognition rate comparatively for foreign languages. We call hereafter the propose architecture is modular time delay neural network (MTDNN) and we define it in this way because we have used modularity concept within the architecture with time delay feedback. Neurons are connected by dynamic spiking synapses. Spiking neurons are formed into a hierarchy of Layers (Slab) [4].

## II. RELATED WORK

Research on Bangla speech processing is in preliminary stage still now. Das et al. [5] presented a speech recognition system based on neural network using LPC, MFCC, ZCR and STE features. A speaker independent method of recognizing isolated Malayan word is proposed in [6]. The development of the speech recognition system using artificial neural network and digital signal processing techniques was described in [7]. Recognition of spoken letter is described in [8]. Rahman et al. proposed a continuous Bangla speech recognition system [9]. They used back propagation neural network for recognition. Another one described an approach to the development of an automated Bangla speech recognition system as a biometrically based technology and is concerned with the speech recognition system using ANN [10].

Most of these previous works limited to thirty words and with higher training time. For speaker detection in Bangla no mentionable work have been done yet. Although there are a lot of issues to solve related to the spoken language processing in Bangla, this paper particularly will focus on an important issue-recognizing Bangla words and detecting corresponding speakers.

## III. PROBLEM FORMULATION

Speech recognition is the process of transforming an acoustic signal into a set of words that was captured by an electronic microphone. This paper will investigate the problem of recognizing isolated Bangla words and corresponding speakers. We have taken speaking mode as isolated words, speaking style as read speech and enrollment as speaker-dependent. Here vocabulary is considered small, SNR is low (less than 10 dB) and voice-cancelling microphone is used as transducer. Different words result different utterance

978-1-4673-9930-2/15/$31.00 ©2015 IEEE

sequence of basic phonemes, variations in the time scale of the patterns and variations in spectrum shape. Again every human utterance of the same word results frequency variations for unique vocal cord vibration mechanism. Here we have considered Glottal Pulse Model for Voiced Signals. The Liljencrants-Fant (LF) model [11] of the derivative of the glottal pulse is defined as Eq.1

$$v(t) = \begin{cases} E_0 e^{at} \sin \omega t & \text{if } v \le t < T_e, \\ E_1 (e^{-\beta(t-T_e)} - e^{-\beta(T_e-T_c)}) & \text{if } T_e \le t < T_c, \\ 0 & \text{if } T_c \le t < T_0 \end{cases} \quad (1)$$

where a segment of less than 3/4 of a period of a sine wave, with a frequency of $\omega$ and an exponential envelop of $E_0 e^{\omega}$.

The speech recognition problem can be stated as follows: suppose we have a stream feature X extracted from the realization of a spoken sentence, and a network of pre-trained speech models $\wedge$ decode the most likely spoken word sequence $W = [w_1, w_2, ...w_N]$. Note that speech model network $\wedge$ contains models of acoustic features representation of words and can also include a model of language grammar. Formally, the problem of recognition of words from a sequence of acoustic speech features can be expressed in terms of maximization of a probability function as $[W_1, W_2, ...W_N] = maxf([w_1, w_2, ...w_N]X, \wedge)$ where f is the conditional probability of a sequence of words $W$ given a sequence of speech features $X$. Note that in practice speech is processed. In the same way for speaker sequence $P = [p_1, p_2, ...p_N]$ and corresponding probability function as $[P_1, P_2, ...P_N] = maxf([p_1, p_2, ...p_N]X, \wedge)$.

## IV. PROPOSED SYSTEM ARCHITECTURE

In order to recognize uttered Bangla words and detect corresponding speaker, we have to use modular time delay approach with neural network. Fig.1 shows the conceptual model of our proposed architecture.

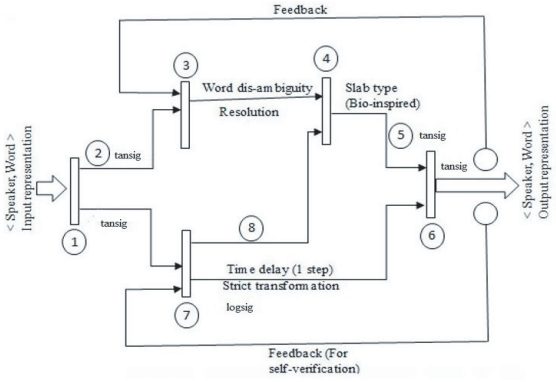


Fig. 1. Conceptual model of proposed modular time delay neural network (MTDNN).

### A. Functional Description of MTDNN

We use learning bias variables for every layer. Before processing we map every input within range $[+1, -1]$ and remove constant rows for fast processing. At the start of training the system initialize all layers of the network using the Nguyen-Widrow method [12]. A total of 15 neurons are set at input layer, next two layer levels having 30 neurons, and layer before output layer have 18 neurons. For processing first two hidden layers the system uses tan sigmoidal function and then uses log sigmoidal in the next level. Finally collecting internal processed information into a layer where those are processed using tan sigmoidal function. Biases are learned using perceptron learning rule. Biases are initialized calculating middle point of inputs. Output contains 53 neurons with processing function of tansigmoidal. Layer weights are initialized by layer-by-layer network initialization function. Based on the dot product, using synaptic weights inputs are collected within neurons. One hidden layer send processed information to the next upper layer and 1 step time is used for synchronization.

### B. Semantic Integration

For modular, we integrated several modules semantically to acquire local forms (neighborhood of an equilibrium state of the network) of controllability and observability [13]. The semantics and purposes of the modules of Fig.1 are described in the following:

- **1**− Distribution module for verification.
- **2**− Adaptive transformation; it is necessary as human utterance may change due to fear, emotion, sleep etc. But frequency remain same.
- **3**− Word sense module.
- **4**− $< Word - Speaker >$ encryption-decryption module.
- **5**− $< Speaker - Word >$ internal encoding.
- **6**− Decision module.
- **7**− Speaker identification module.
- **8**− Internal representation of speaker.

### C. Modularity Making Task Simple

Modular Neural Network [14] approach is very effective in searching for solutions to complex problems of various fields. The neurobiologists have long believed, and appreciated the fact, that the regions of animal and human brains are organized into specialist and functionally segregated modules. In brain functionally similar modules are bound together through synchronization, feedback and lateral connections. So, developing Modular NN 3 steps to follow: task decomposition, training and multi-module decision making [15]. Modular NN is effective for modularizing complex task [16]. For its robustness it also provides hierarchical mixture of domain experts [17] while training, also provides computational efficiency and meaningful data separation and co-relation [4].

### D. Time Delay Recurrent Structure

A time delay neural network unit has the ability to relate and compare current input to the past history of events, giving network the ability to co-relate events in time, finding time invariant features [18]. In our structure, recurrent time

delay structure is used in hope to get proper system synchronization of feature identification and adaptation that can generate any finite time trajectory [19] and develop dynamic long term memory.

## V. EXPERIMENTS

We have run experiments on a general purpose computer (Windows 7), $CORE^{TM}i3$ with 3.66 GHz processor speed. We performed the experiment in multimedia lab, CUET with reasonable silent environment. Fig.2 represents the schematic set up of the experimental process.

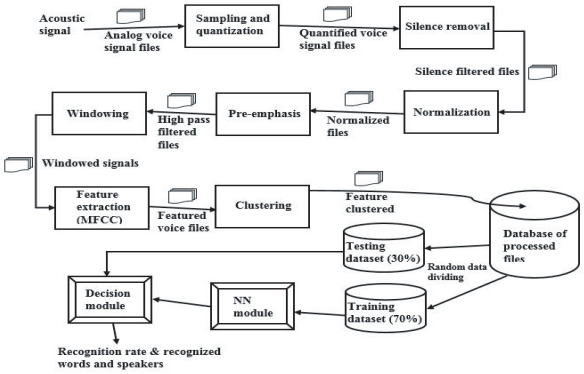


Fig. 2. Process diagram of the experiment.

### A. Acoustic Signal Capturing

Human analog voice signal is captured by a noise canceling close talk microphone using sufficient interval between start and stop point in time for processing later in digital form such as Fourier Transform.

### B. Sampling and Quantization

Sampling simply means to collect speech from human utterance to a digital storage. Here quantification simply means approximation of voiced signal into countable and computable range. This will give a platform for further analysis of natural signal. Children have very high fundamental frequency compared with male (85-180Hz) and female (165-255Hz) of adult stage. We have used a sampling rate of 22.050 kHz, 16 bit PCM. Then quantization is done for digitizing.

### C. Silence Removal

Some Physical strategy and some software based silence removing tools used. After that processed signals are stored as .wav files.

### D. Normalization

Normalization is a process of mapping approximated values within a range such that processing devices can efficiently process those collected data. Signal power is important while extracting MFCC. We use Matlab mapminmax function for -1 to +1 mapping that will enable us to compute within bipolar range only.

### E. Pre-emphasis

For spectral flattening voiced signal a low level system module is applied to remove finite precision effects as far as possible. Most used filter is first order system. Using 1 as denominator and $[1-0.9375]$ as nominator for filtering we use Matlab filter function that works as first-order high-pass filter .For spectral energy filtration, we guess mouth speech decay of 6dB per octave. After applying this filter, the output $s'(n)$ and input s(n) have the following relationship in the time domain: $s'(n) = s(n) - as(n-1)$.

### F. Frame Blocking and Windowing

In order to prevent spectral discontinuities frame blocking techniques have used that produce some overlapping regions. The windowing is applied to each frame to minimize the spectral discontinuities at both end. The windowing can be expressed as Eq.2.

$$x_t'(n) = w(n)x_t(n) \quad if \quad 0 \le n \le (N-1) \tag{2}$$

where $x(n)$ = applied signal, $x^{'}(n)$ = windowed version of the signal. We use, function enframe (from VoiceBox [20] toolbox of Matlab for framing and windowing (Hamming) in single step that lead us extracting invariance properties of speech signal. It prevents non-continuity in the two ends of the audio frame, and to avoid the influence of front and back audio frames when analyzed. Non-continuity of audio frames is achieved due to more centered audio frequency frame.

### G. Feature Extraction (MFCC)

A widely used form of cepstrum is mel frequency cepstral coefficients (MFCC) [21]. These coefficients represent audio based on perception. MFCC is considered the best available approximation of human ear. Because the mel frequency cepstrum can represent a listener's response system in a better way. To obtain MFCC features, the spectral magnitude of FFT frequency bins are averaged within frequency bands spaced according to the mel scale which is based on a model of human auditory perception. The scale is approximately linear up to about 1000 Hz and approximates the sensitivity of the human ear.

Feature extraction consists of computing representations of the speech signal that are robust to acoustic variation but sensitive to linguistic content. The Mel-filter is used to find band filtering in the frequency domain with a bank of filters. The filter functions used are triangular in shape on a curve linear frequency scale. The filter function depends on three parameters: the lower frequency, the central frequency and the higher frequency. On a Mel scale the distances between the lower and the central frequencies are equal. The filter functions that have used to noise removal as presented in Eqs.3-5.

$$H(f) = 0 \quad if \quad f \le f_1 \quad and \quad f \le f_h \tag{3}$$

$$H(f) = \frac{f - f_1}{f - f_h} \quad if \quad f_1 \le f \ge f_c \tag{4}$$

$$H(f) = \frac{f_h - f}{f_h - f_c} \quad if \quad f_c \le f \ge f_g \tag{5}$$

Mel-Frequency cepstral coefficients are found from the Discrete Cosine Transform of the filter bank spectrum by using the Davis and Mermelstein formula [22] as Eq.6

$$\sum_{j=1}^{N} P_j \cos\left(\frac{i\pi}{N(j-0.5)}\right) \quad (6)$$

$P_j$ denotes the power in dB in the $j^{th}$ filter and N denotes number of samples. From theoretical basis, approximation of human ear (listening) can be done effectively using MFCC (considering sensitivity level of human ear). For Bangla Speech recognition MFCC39 reports good [23], while applying modular technique it also shows good performance. In this purpose, we use function melcepst from Voicebox.

### H. Handling Acoustic Randomness and Fluctuation

Unsupervised Fuzzy-c-means clustering technique developed by Bezdek [24] is used in feature clustering that is better than k-means suitable for handling acoustic fuzziness and variability.

### I. Training

70% samples for training chosen in random order. Among two modes (epoch wise and continuous) of training a recurrent network, it was chosen epoch wise training [25] using epoch wise back-propagation through time algorithm [26].

### J. Testing

15% for testing of speaker dependent or independent mode. And 15% for validation.

### K. Data Collection

Creative WaveStudio was used to capture the uttered voice signal from the participants. Five participants were participated in the experiment. Their average age are 22 years. A total of 525 (5 [participants]$\times$ 35 [no. of words] $\times$ 3 [no. of times uttered each word]) sample words was collected from all participants. Fig.3 presents word list that have used in experimentation.

| |
|---|
| **Mono Syllable:** দেশ, বই, বেশ, শোক, ভোট, কাজ, তার, রাত, গাছ, কেউ, আজ |
| **Di Syllable:** তিনি, পড়েন, হবে, টাকা, থাকেন, আমরা, পাতা, পুলিশ, কথা, উৎসব, যুদ্ধ, বিজয় |
| **Tri Syllable:** ধরণের, জানানোর, আমাদের, বর্তমান, ফলাফল, সুবিধা, কর্তৃপক্ষের, সৌন্দর্য, পদত্যাগ, পৃথিবীর |
| **Poly Syllable:** স্বাধীনতার, সহযোগিতা |
| **English Word:** Word, file, open, print, exit, edit, cut, copy, paste, doc1, doc2 |

Fig. 3. List of words used in the experiment.

### L. Processing

We have used hp filtering of MATLAB for high pass filtering. Then the voice signal is processed by remove constant row, mapminmax for removing constant rows from data matrix and mapping data values within the range [0-1]. Input layer transfer function is tansig. Input layer initialization function is initnw. Transfer function of layer 1, 2, 4 is tansig and initialization function is initnw and for layer 3 is logsig and initnw respectively. For all layers bias initialization and bias learning function are midpoint and learnp respectively. For input layer initialization, learning and weight functions are initlay, learnp and dotprod respectively. For all layers initialization and weight functions are initlay and dotprod respectively. Every delay inherently embedded within the MTDNN architecture are implemented at source code level using 1-step time delay. For network object initialization, performance estimation, training and data dividing function are initlay, mse (mean squared error), trainlm and dividerand. We have used 53 output layer neurons for taking speaker and word recognition output. We have trained our net object and taken different performance measurement for measuring performance of our model.

## VI. RESULTS

Fig.4 shows the initial sound frequency of word *Desh* from our word list.

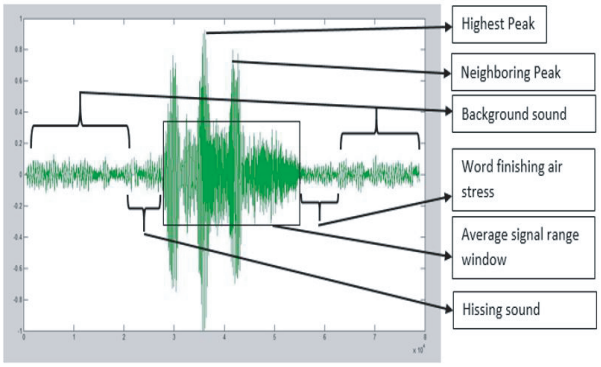


Fig. 4. Initial sound frequency of a sample word *Desh*.

In that time the frequency have many section such as highest peak, neighboring peak, background sound, finishing air stress, average signal range window and hissing sound. Then the sound is send for noise cancellation and volume reduction which shows in Fig.5.

We have taken effect between -1 to 1 where the volume reduction effect remains between -0.8 to 0.8 on scale. After that the frequency is send for windowing and Fig.6 shows that.

We have used hamming windowing function where the frame length is 256 and overlap is 192. Now the time for feature extraction and we used MFCC. Fig.7 shows that extraction.

In -2 to 2 range the frequencies are separating based on cepstrum and in lower section the accumulated effect is

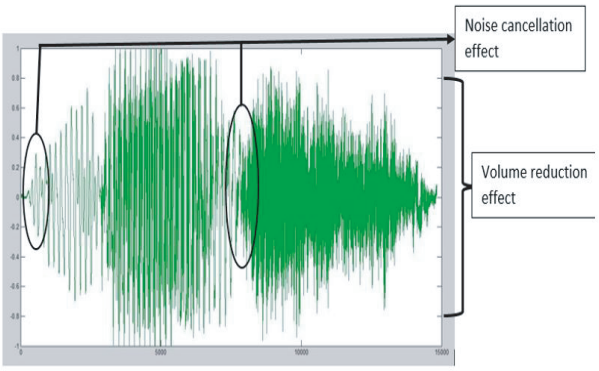


Fig. 5. After noise cancelation and volume reduction.

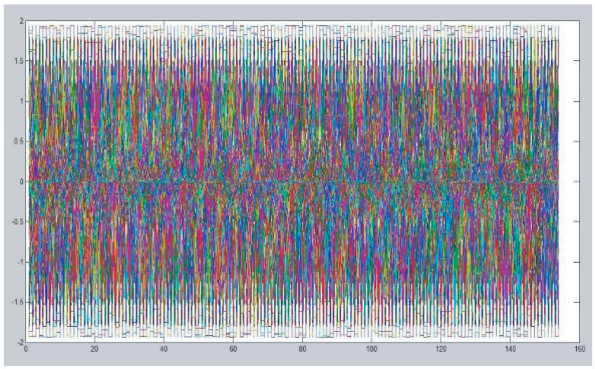

Fig. 6. After windowing.

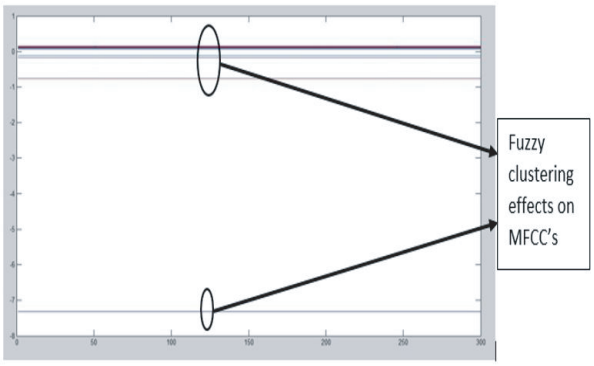


Fig. 8. After FCM clustering.

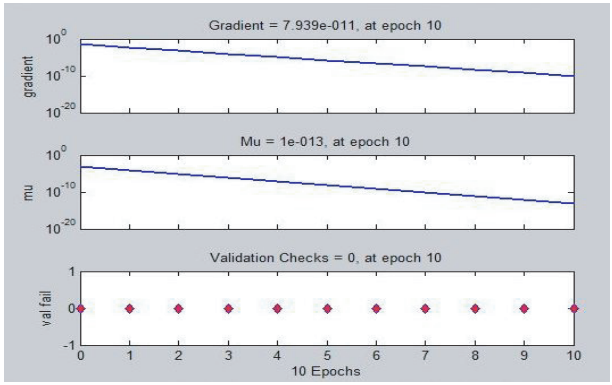


Fig. 9. One of many Batch wise training states.

covered. After the words are ready for clustering as Fig.8 follows.

Here the fuzzy c means clustering is used. The result is clustered by 12 matrix. And it will be the input to the neural network. After the help of neural network training we got the result as follows in Fig.9.

It is one of many batch wise training states. Figure shows the result of final epoch. As the more train is done the output will be more error free. Fig.10 shows the error surface of output neuron in 3 dimensions. Sum squared error, weight and bias are those dimensions.

We can show the error surface in 2 dimensions where weight and bias are used. Fig.11 shows the contour of error surface.

### A. *Comparison*

Different techniques used different sample size and different windowing technique. Thus, it is very difficult to make an exact comparison with various methods. We have compared our approach with the other approaches in terms of no. of words, features used and recognition accuracy. Evaluation is done based on the recognition accuracy (R) which was calculated with a range value from Matlab Neural Network

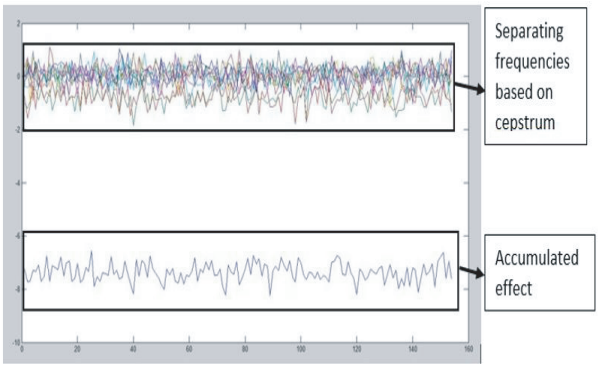


Fig. 7. After extracting MFCC.

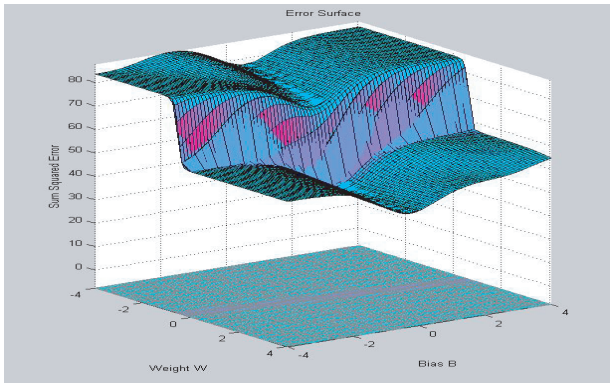


Fig. 10. Error surface of output neuron.

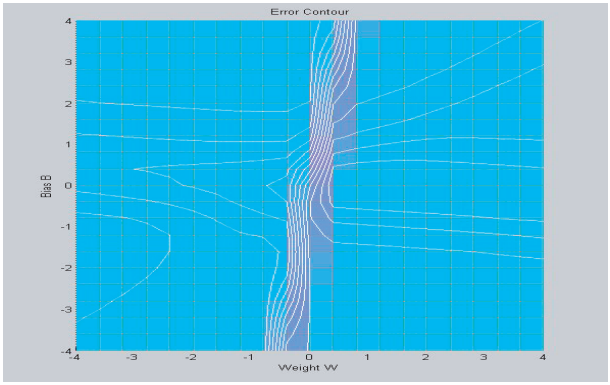


Fig. 11. Contour of error surface.

Tool using eq 7.

$$R = \frac{trained\ performance\ value}{max\ range\ value\ -\ min\ range\ value} * 100\% \quad (7)$$

. Table I illustrates the summary of the comparison. The result indicates that the proposed approach is somewhat better compared to the other methods.

TABLE I

COMPARISON OF PROPOSED METHOD WITH OTHERS

.

| Authors | No. of words | Feature | Recognition rate |
|---|---|---|---|
| Roy et al. [7] | 5 | FFT | 82% |
| Islam et al. [10] | 30 | FFT | 78.85% |
| Proposed Work | 35 | MFCC | 82.66% |

## VII. CONCLUSION

The main objective of our work is to design a framework for recognizing Bangla words and detecting corresponding speaker. For this purpose, we have proposed a modular time delay neural network. Although the performance of the current implementation is quite low it would be much better if noise-canceling microphone be used in perfect silent environment. Though speaker volume frequency remain same most of the time, but their utterance style may vary due to fear, emotion, sleepy mode etc. Future research can use positive and negative samples for training in an unsupervised way or different architectures like convolutional, scaly etc. Instead of CPU if GPU be used or other clustered computer technology be used, then overall system performance will increase significantly. Getting perfect silent room was slightly difficult. Air flow, surrounding noise, hissing sound etc. disturbed the process and taking speech signal from different people was difficult also.

## REFERENCES


[1] C. S. Kumar, V. P. Mohandas, H. Li, *Multilingual Speech Recognition: A Unified Approach*, in Proc. 9th International Conference On Speech and Computer (SPECOM), Lisbon, Portugal, 2005, pp. 3357-3360.

[2] S. C. Kumar, L. Haizhou, *Language Identification System for Multilingual Speech Recognition Systems*, in Proc. 9th Int. Conf. on Speech and Computer (SPECOM), St. Petersburg, Russia, 2004.

[3] M. Bin, G. Cuntai, L. Haizhou, L. Chin-Hu, *Multilingual Speech Recognition with Language Identification*, International Conference on Spoken Language Processing (ICSLP), Denver, Colorado, 2002.

[4] P. Poirazi, C. Neocleous, C. S. Pattichis, C. N. Schizas, *Classification Capacity of a Modular Neural Network Implementing Neurally Inspired Architecture and Training Rules*, vol. 15, no. 3, May 2004.

[5] B. P. Das and R. Parekh, *Recognition of Isolated Words Using Features Based on LPC, MFCC, ZCR and STE with Neural Network Classifiers*, International Journal of Modern Engineering Research, vol. 2, 2012, pp. 854-858.

[6] S. Sunny, D. Das and M. G. Ali, *Development of a Speech Recognition System for Speaker Independent Isolated Malayan Words*, International Journal of Computer Science and Engineering Technology, vol. 3, no. 4, April 2012, pp. 69-75.

[7] K. Roy, D. Das and M. G. Ali, *Development of the Speech Recognition System Using Artificial Neural Network*, in Proc.5th International Conference on Computer and Information Technology (ICCIT02), Dhaka, Bangladesh, 2002, pp. 118-122.

[8] A. H. M. R. Karim, M. S. Rahman and M. Z. Iqbal, *Recognition of Spoken Letters in Bangla*, in Proc. of 6th ICCIT, Dhaka, 2002, pp. 213-216.

[9] K. J. Rahman, M. A. Hossain, D. Das, T. Islam, and M. G. Ali , *Continuous Bangla Speech Recognition System*, in Proc.6th International Conference on Computer and Information Technology (ICCIT03), Dhaka, Bangladesh, 2003, pp. 303-307.

[10] M. R. Islam, A. S. M. Sohail, M. W. H. Sadid and M. A. Mottalib, *Bangla Speech Recognition Using Three Layer Back-Propagation Neural Network*, in Proc. of NCCPB, Dhaka, 2005, pp. 148-153.

[11] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[12] D. Nguyen and B. Widrow, *Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights*, in Proc. Int. Joint Conf. on Neural Networks, vol. 3, 1990, pp. 21-26.

[13] A. U. Levin and K. S. Narendra, *Control of nonlinear dynamical systems using neural networks controllability and stabilization*, IEEE Tran. on Neural Networks, 1993, vol. 4, pp. 192-206.

[14] M. I. Jordan and R. A. Jacobs, *A competitive modular connectionist architecture*, in Advances in neural information processing systems, 1991, pp. 767-773.

[15] G. Auda and M. Kamel, *Modular Neural Network Classifiers: A Comparative Study*, J. of Int. and Robotic Sys., 1998, pp. 117-129.

[16] H. C. Fu, Y. P. Lee, C. C. Chiang and H. T. Pao, *Divide-and Conquer Learning and Modular Perceptron Networks*, IEEE Transactions on neural networks, 2001, vol. 12, no. 2, pp. 250-263.

[17] M. I. Jordan and R. A. Jacobs, *Hierarchical Mixture of Experts and the EM Algorithm*, Neural Computation, 1994, Vol. 6, pp.181-214.

[18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, *Phoneme Recognition Using Time-Delay Neural Networks*, IEEE Transactions on Acoustics Speech and Signal Processing, 1989, Vol. 37. No. 3, pp. 328-339.

[19] K. J. Lang and A. H. Waibel, *A Time-Delay Neural network Architecture for Isolated Word Recognition*, Neural networks, vol. 3, 1990.

[20] Voicebox: Speech processing toolbox for matlab. "http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html".

[21] N. J. Lisa, Q. N. Eity, G. Muhammad, M. N. Huda, C. M. Rahman, *Performance Evaluation of Bangla Word Recognition Using Different Acoustic Features*, International Journal of Computer Science and Network Security (IJCSNS), vol. 10, no. 9, 2010.

[22] M. Sahidullah and G. Saha, *Design, Analysis and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition*, Speech Communication, vol. 54, no. 4, pp. 543-565, 2012.

[23] M. F. Khan and R. C. Debnath, *Comparative Study of Feature Extraction Methods for Bangla Phoneme Recognition*, in Proc. Int. Conference on Computer and Information Technology, 2002, Dhaka, Bangladesh, pp. 257-261.

[24] J. C. Bezdek, R. Ehrlich, W. Full, *FCM: The Fuzzy C-Means Clustering Algorithm*, Computers and Geosciences, 1984, vol. 10, no. 2-3, pp. 191-203.

[25] R. J. Williams and D. Zipser,*Gradient Based Learning Algorithms for Recurrent Networks and Their Computational Complexity*, Gradient Based Learning, 1995.

[26] R. J. Williams and J. Peng,*An Efficient Gradient Based Algorithm for On-line Training of Recurrent Network Trajectories*, Neural Computation, 1990, vol. 2, pp. 490-501.