Time Delay Neural Network for Bengali Speech Recognition

[I]Rubya Shaharin, [II]Uzzal Kumar Prodhan, [III]Dr. Md. Mijanur Rahman
[I]Lecturer, [II]Assistant Professor, [III]Associate Professor
Dept. of Computer Science & Engineering
[I]University of Information Technology and Sciences, Dhaka, Bangladesh
[II,III]Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh

Abstract

This paper introduces with the Bengali Speech Recognition system by Time Delay Neural Network (TDNN). This system includes feature extraction. The input speeches for this system are recorded from a single speaker and extract features by using Binary feature extraction method. After feature extracting, it has been recognized by Time delay Neural Network (TDNN) a popular speech recognition technique. Train TDNN with the Conjugate Fletcher-Reeves Update training algorithm to evaluate this system coupled with binary feature extraction and TDNN. Set of speech data are used to evaluate the performance of this system. The system has average accuracy 97%. This system is suitable for isolated Bengali speech word recognition.

Keywords

Artificial Neural Network, Bengali Speech Recognition, Isolated Word, LPC, MFCC, MLP, TDNN

I. Introduction

To be easiest life speech recognition plays vital role. In the demand of modern technology speech recognition is more challenging research area for the researchers. After the 1950s speech recognition efficiency increased day by day.

Although speech recognition started since 1930s, Bengali speech recognition research work has been started since around 2000. There are from 6800 to 6900 distinct languages in the modern world. Bengali language is one of them. Approximately 8% of people speak in Bengali in the world [3]. So it is very important to improve the efficiency of Bengali speech recognition.

For 40 years, Artificial Neural Networks (ANNs) have been used for difficult problems in pattern recognition [Viglione, 1970]. In 1989 waibel et al introduce special neural network architecture for phoneme recognition which is known as Time Delay Neural Network (TDNN) [2]. The architectural properties are almost same to the MLP only TDNN has an extra property is known as tapped delay line. For the training criteria artificial neural network is different from the conventional computing system. Conjugate Fletcher-Reeves Update training algorithm is the best choice of TDNN training [1]. We use this algorithm to train TDNN to recognize Bengali speech. The aim of this paper is to recognize Bengali speech using Time Delay Neural Network (TDNN). To evaluate this system we use isolated Bengali speech word as inputs of TDNN,

This paper organized as follows: Section I describes the introduction and arrangement of this paper. Section II describe about the TDNN. Section III about speech recognition approaches. Section IV introducing some speech feature extraction techniques. Section V discusses the methodological steps of this work. Section VI shows the experimental result and Section VII concludes this paper.

II. TDNN

The basic TDNN architecture [6,7] [8] is the four layers feed forward neural network model with the property shift invariant connection. The layers are Input layer, first hidden layer, second hidden layer, and a output layer. In the input layer the time series of Mel-scaled 16 channel FFT spectrum are fed into. One pattern contains 15 frames (vectors). So the input layer has 15X16 units. The analysis window which consists of few analysis frames, basically the size of analysis window at input layer is 3. The first 3 frame are connected to one frame of hidden layer. Then window is shift one frame forward and present window contains 2, 3 and 4 frame of input layer.

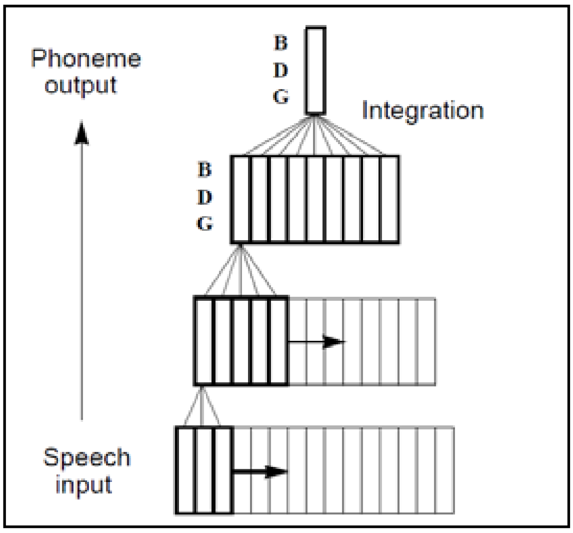

Fig. 1: TDNN architecture for phoneme recognition

The second window is connected to second frame of first hidden layer. The shifting of analysis window at the input layer repeated until reaching the last analysis frame at the input layer. A total 13 analysis frames are formed in the first hidden layer based on the 15 analysis frames at the input layer. Every analysis window has its own copy of weights when it is connected to the analysis frame at the next layer. Every neuron in one analysis window at the input layer is fully connected to all neurons in one analysis frame at the first hidden layer.

The first hidden size at one time is comprised of an analysis window with the size of 5 analysis frames. The first analysis window at first hidden layer is connected to the first analysis frame at the second hidden layer. The analysis window is shifted by one analysis frame in order to connect to the second analysis frame at the second hidden layer. The shifting of analysis window at first hidden layer is repeated until reaching its last analysis frame. Thus a total of 9 analysis frames are formed at second hidden layer based on 13 analysis frames at first hidden layer. Every analysis window has its own copy of weights when it is connected to the next hidden layer. Every neuron at one analysis window of first hidden layer is fully connected to all neurons in one analysis frame at the second hidden layer.

The second hidden layer has 9 analysis frames, with each analysis frame consisting of 3 hidden neurons. The neurons across all the analysis frames at the second hidden layer are connected to the corresponding output neuron at the output layer. Thus the hidden neuron number at the second hidden layer is set in a way that equals to the output neuron number at the output layer. The first neuron at the second hidden layer across all activated analysis frames is connected to the first output neuron at output layer. The same approach is applied to the second until the last output neuron.

III. Speetch Recognition

Basically there are three approaches [4] to speech recognition.

A. Acoustic Phonetic Approach
B. Pattern Recognition Approach
C. Artificial Intelligence Approach

The following figure 2 shows the approaches of speech recognition.

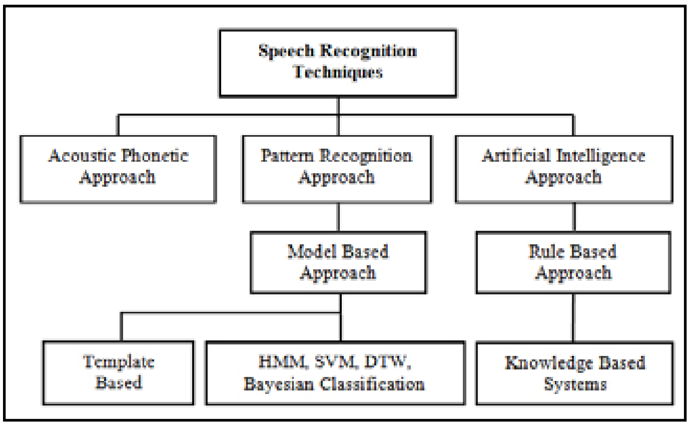
Speech Recognition Techniques
Acoustic Phonetic Approach
Pattern Recognition Approach
Artificial Intelligence Approach
Model Based Approach
Rule Based Approach
Template Based
HMM, SVM, DTW, Bayesian Classification
Knowledge Based Systems

Fig. 2: Speech recognition/ Classification Techniques [5]

A. Acoustic Phonetic Approach

Acoustic phonetic approach for speech recognition is based on finding speech sound and providing appropriate labels these sounds. The basis of acoustic phonetic approach based on the fact that, there exist finite and exclusive phonemes in spoken language and these are broadly characterized by a set of acoustic properties that are demonstrated in the speech signal over time. With speaker and co articulation effect, the acoustic properties of phonetic units are highly variable, it is assumed in this approach that, the criteria governing the instability are straightforward and can be readily learned by machine.

B. Pattern Recognition Approach

Two essential steps involves in pattern recognition approach are, pattern training and pattern comparison. Using a well formulated mathematical framework and initiates consistent speech pattern representation for reliable pattern comparison, from a set of labeled training samples through formal training algorithm is essential feature of this approach. In this, there exist two methods: Template base approach and stochastic approach. Stochastic model are more suitable approach to speech recognition as it uses probabilistic models to deal with undetermined or incomplete information. There exits many methods in this approach like HMM, SVM, DTW, VQ etc, among these Hidden Markov Model is most popular stochastic approach today.

C. Artificial Intelligence Approach

Combination of acoustic phonetic approach and pattern recognition approach makes The Artificial Intelligence approach. Acoustic phonetic knowledge is used to developed classification rules for speech sound where template based methods provide less insight about human speech processing[10], but these methods have been very productive in the design of a diversity of speech recognition system.

Artificial Neural Network method is more reliable method for this approach. Artificial neural networks (ANN): An artificial neural network contains potentially large number of simple processing element that is called units or neurons, which impact each other's performance via a network of excitatory or repressive weights [9].

The overall process for speech recognition are acquisition of input speech, feature extraction, design recognize then finally get the recognized output. The basic model for speech recognition is given below:

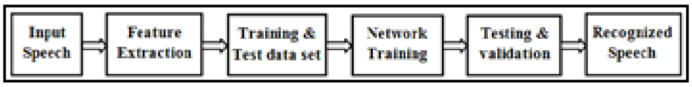

Fig. 3: Block diagram of speech recognition based on ANN

Fig.3 shows basic representation of speech recognition system in simple equation which contains feature extraction, database, network training and testing or decoding. We produce speech signal as sound file. This signal has some property which is known as feature of speech. After recording speech as signal we extract those features from this signal by which speech can identified. The extracted features are fed into the recognition tool. Some features are used for training and others some are used as testing vector. After training the network are adjusted for such types of data are recognized. And finally network is able to recognize both known and unknown speech signals. This is the procedure of speech recognition.

IV. Speech Feature Extraction Techniques

Feature extraction [4] is the most important part of speech recognition as it distinguishes one speech from other. The utterance can be extracted from a vast range of feature extraction techniques suggested and successfully utilized for speech recognition task, but extracted feature should meet some criteria while negotiating with the speech signal such as [9]:

- Easy to measure extracted speech feature
- It should not be receptive to mimicry
- It should show less variation from one speaking environment to another
- It should be balanced over time
- It should occur normally and naturally in speech

Different techniques for feature extraction are shown in following table:

Table.1: Popular Feature Extraction Methods

No	Feature Extraction Method
1	Linear prediction Coding (LPC)
2	Mel Frequency Cepstral Coefficient (MFCC)
3	Principal Component Analysis (PCA)
4	Linear Prediction Cepstral Coefficient (LPCC)
5	perceptual linear predictive coefficients (PLP)
6	Relative spectra filtering of log domain coefficients (RASTA)
7	Binary Feature Extraction method [11]

V. Methodology

We propose a speech recognition system based on TDNN with binary feature extraction technique. The proposed system is shown in figure 4.

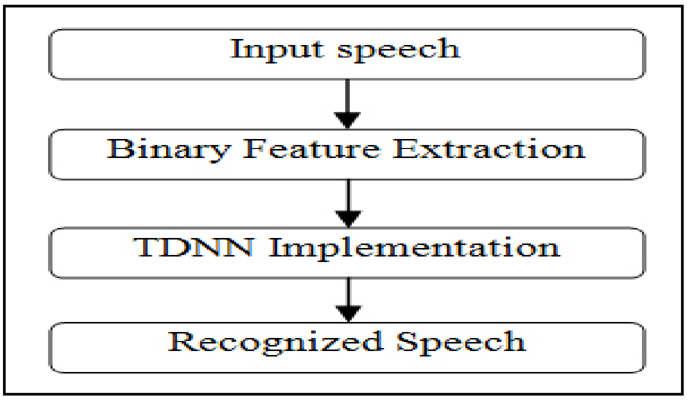

Fig. 4: Proposed System for SPEECH Recognition

A. Input Speech

For this system we have used 10 isolated Bengali speech words. The speeches are recorded from a single speaker by using MATLAB 7.10.0 (R2010a) speech recording function. Speech word is recorded for 1 second with the sampling rate 8 KHz and the number of bit per sample is 16. Each word has been recorded for 20 times that is each word has 20 samples. 200 samples are recorded for 10 words.

B. Binary Feature Extraction

Feature extraction is the most important part of all recognition system. After acquiring speech data we choose only the voiced part of speech then we create the spectrogram image from it. The spectrogram images are converted to the binary image by using Otsu's thresholding method. The binary images are too large for calculating feature that's why its need to resize. Resize the binary image into 8X6 dimension. After resizing calculate the pixel value. The values are the features of speech signals. The process of binary feature extraction is summarized by the following block diagram:

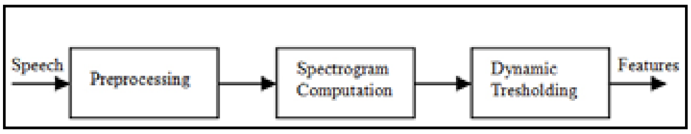

Fig. 5: Main steps of Binary Feature Extraction

C. TDNN Implementation

For Bengali speech recognition we implement a Time Delay Neural Network (TDNN) by using MATLAB with 2 hidden layers. TDNN have 100 and 40 neurons for 1st and 2nd hidden layer respectively. Train the TDNN using Fletcher-Reeves Conjugate Gradient algorithm.

VI. Experimental Result

The proposed system has been implemented in windows environment using MATLAB 7.10.0 (R2010a) version. For this experiment there were use 10 Bengali isolated speech words as inputs. The inputs are divided into 20 groups. Each group has 10 words (from 1 to 10). First 10 groups are used to train the network and rest of 10 groups is used to test the TDNN. The performance of this system for speech recognition is given by table 2. The table shows the recognition accuracy (%) of this system.

Table.2: Experimental Results

Group	Iteration	Time	Performance	Gradient	Accuracy (%)
1	38	2.32	0.0108	0.0447	100
2	31	1.44	0.0210	0.0970	100
3	23	2.25	0.0136	0.0899	100
4	73	2.45	0.0101	0.0112	100
5	64	3.35	0.0086	0.0488	100
6	52	2.52	0.0119	0.0102	100
7	59	3.15	0.0135	0.0583	100
8	51	3.06	0.0109	0.0577	100
9	36	2.07	0.0136	0.0372	90
10	58	3.33	0.0090	0.0376	100
11	95	5.26	0.0036	0.0907	100
12	26	1.30	0.0227	0.0991	90
13	45	2.44	0.0084	0.0590	90
14	55	3.17	0.0065	0.0105	100
15	28	1.34	0.0201	0.0796	90
16	54	3.09	0.0073	0.0620	100
17	42	2.22	0.0112	0.0417	100
18	68	3.50	0.0069	0.0503	100
19	55	3.08	0.0064	0.0480	100
20	31	1.55	0.0191	0.0684	90

VII. Conclusion

The achievement of this research is higher accuracy rate for Bengali Speech recognition. This research used binary feature extraction method for extract feature from speech data and also finds out an efficient training algorithm for train the network. The number of neuron at first hidden layer should equal to the number of training pattern. This experiment has 97% average accuracy for Bengali speech recognition. The couple of training parameters are the best parameter for speech recognition using TDNN. The accuracy is higher.

VIII. AcknowledgEment

The author would like to thanks to the Ministry of Information and Communication Technology of Bangladesh for giving financial support by providing MOICT-Fellowship.

References

[1] *Rubya Shaharin, Uzzal Kumar Prodhan, Dr. Md. Mijanur Rahman "Performance Study of TDNN Training Algorithm for Speech Recognition" International Journal of Advanced Research in Computer Science & Technology (IJARCST) Vol. 2, Issue 4 (Oct. - Dec. 2014).*

[2] *Colin Keng-Yan TAN and Kim-Teng LUA,"Learning of Word Boundaries In Continuous Speech Using Time Delay Neural Networks".*

[3] *Md. Mijanur Rahman, Fatema Khatun, and Dr. Md. Al-Amin Bhuiyan "Development of Isolated Speech Recognition System for Bangla Words", Vol. 1, Issue. 3, September - December, 2010 (IJARITAC), pp.272-278.*

[4] *Nidhi Desai,Prof.Kinnal Dhameliya,and Prof.Vijayendra Desai,"Feature Extraction and Classification Techniques for Speech Recognition: A Review",International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 12, December 2013).*

[5] *S.J.Arora and R.Singh, "Automatic Speech Recognition: A Review", "International Journal of Computer Applications", vol60-No.9, December 2012.*

[6] *Hua-Nong Ting, Boon-Fei Yong, Seyed Mostafa Mirhassani, "Self-Adjustable Neural Network for Speech Recognition", Engineering Applications of Artificial Intelligence, ELSEVIER, 2013.*

[7] *Colin Keng-Yan TAN and Kim-Teng LUA,"Learning of Word Boundaries In Continuous Speech Using Time Delay Neural Networks"*

[8] *ALEXENDER WAIBEL, TOSHIYUKI HANAZAWA, GEOFFREY HINTON, KIYOHIRO SHIKANO AND KEVIN J LANG, "Phoneme Recognition Using Time delay Neural Networks", IEEE Transaction on acoustic speech and signal processing Vol.37, No. 3 March 1989.*

[9] *Nidhi Srivastava and Dr.Harsh Dev"Speech Recognition using MFCC and Neural Networks", International Journal of Modern Engineering Research (IJMER), march 2007.*

[10] *Sai Jayram A K V, Ramasubramanian V and Sreenivas T V, "Robust parameters for automatic segmentation of speech", Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02), Vol. 1, pp. 513-516, 2002.*

[11] *Md. Mijanur Rahman and Md. Al-Amin Bhuiyan,"DYNAMIC THRESHOLDING ON SPEECH SEGMENTATION",IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.*