International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I5P121

Volume 73 Issue 5, 241-255, May 2025

Review Article

Bangla Speech Recognition: Power Spectral Analysis, LPC & MFCC as Feature Extraction Techniques in Deep Learning

Md. Shafiul Alam Chowdhury[1,2], Md. Farukuzzaman Khan[2], Mohammed Sowket Ali[3], Shahriar Ahmed[4], Md. Abdul Mannan[5*], Md. Amanat Ullah[5]

[1]*Department of Computer Science and Engineering, Uttara University, Dhaka, Bangladesh.*
[2]*Department of Computer Science and Engineering, Islamic University, Kushtia, Bangladesh.*
[3]*Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, (BAUST), Saidpur, Bangladesh.*
[4]*Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh.*
[5]*Department of Mathematics, Uttara University, Dhaka, Bangladesh.*

[*]*Corresponding Author: mannan.iu31@gmail.com*

Received: 19 November 2024 Revised: 12 May 2025 Accepted: 26 May 2025 Published: 31 May 2025

Abstract *- Speech recognition technology has already become a part of our everyday lives, and many works have been done mostly in the English language because it is an international language, but there is still more that researchers could do. Speech recognition technology has already become a part of the daily life. As can be seen, AI robots can converse with people, particularly in English. The topic of this study is speech recognition in Bangla (Bengali). To determine the highest feasible speech recognition accuracy in the Bangla (Bengali) language, several methods have been employed for pattern recognition and deep learning. Native speakers of Bangla provided the core dataset. It includes extensive experiments with Bangla phonemes, isolated words, commands, and sentences. Speech samples are subjected to feature extraction using MFCC. Simultaneously, LPC and FFT are employed. Using the maximum-likelihood approach, a multilayer feedforward deep neural network model has been utilized. A random dataset has been used to assess the model's accuracy in speech recognition. Deep learning using a neural network model and feature extraction using MFCC outperform Power spectral testing and linear predictor coefficient tests regarding recognition outcomes. The investigation found that increasing the number of speech samples affected the recognition accuracy rate, as did the speech samples from the opposing gender.*

Keywords *- Automatic Voice Recognition (AVR), Deep learning, Linear Predictor Coefficient analysis (LPC), Mel Frequency Cepstral Coefficient (MFCC), Power spectral analysis (FFT), FFNN, Zero Crossing Rate (ZCR).*

1. Introduction

Many studies on speech recognition, primarily in English, have been conducted recently. Bangla, also known as Bengali, is one of the popular languages that must be concentrated. Approximately 300 million people worldwide speak the Bangla language. Still, there is not much study being done in Bangla. One explanation is that speech recognition in Bangla is a challenging undertaking because of the language's complexity and the existence of compound letters. Research in Bangla is still restricted to individual words, commands, and comparatively simple phonemes. Some researchers used a few techniques and procedures for their research in this field. Finding any Bangla command and sentence recognition study is really challenging, though. Bangla digit recognition was done using MFCC/MFCC-LPC and HMM [2] [3], Bangla digit recognition using LPC with neural network [4], and a spectral analysis of Bangla vowels (isolated Bangla words) was carried out to measure the characteristics and format of frequencies during the feature extraction of the Bangla speech signal [1]. These findings provide evidence to conduct some basic research in the Bangla language (even at a narrow level). MFCC with a Multilayer Neural Network (MLN) was used to recognize Bangla phonemes (consonants and vowels) [5], and HMM was subsequently tested [6] [7] [8]. A system was developed to facilitate two-way communication [9]. MFCC with HMM was used to recognize Bangla words [10] [11], and it was also utilized to determine how speaker variation affected the ability to recognize Bangla words and sentences [12]. For Bangla word and command recognition, a combination of Connectionist Temporal Classification (CTC)

and consideration strategy with Recurrent Neural Network (RNN) was employed [13]. Bangla voice recognition has been the subject of some theoretical considerations and issues and fixes [14]. An investigation was carried out into the acoustic analysis of accent-specific pronunciation in the Sylheti dialect and its impact on the Bangla language of Bangladesh [15].

2. Necessary of Previous Task

Some scholars [16] concurred in 2021 regarding the limited scope of study on Bangla alphabets and speech recognition that has been carried out to now. They have identified the lack of available data as one of the primary problems. Several auditory characteristics of the processed data are incorporated into their suggested system. 39 alphabets were employed for classification in this experiment [16]. Support Vector Machines (SVM) and Multilayer Perceptron Classifiers (MLPC) are used to classify the data more precisely.

Approximately 4095 data points were employed, with MLPC and SVM achieving accuracy rates of 99.27% and 92.33%, respectively. Raw data (data set of Bangla alphabets) for this experiment was gathered in MP3 format from various sources for the intended model. The audio files were manually stored in the appropriate directories to be preprocessed. These feathers (the dataset in MP3 file format) are preserved in CSV format.

These researchers claim that the general shape of a spectral envelope (for speech signal) is succinctly defined by the MFCC approach. Thus, the MFCC (20 MFCCs) has been applied. The system in the trials was constructed using the Python programming language. During feature extraction, have considered the following issues-

1. A spectral centroid to gauge the audio's light intensity. The centroid of a spectral frame is calculated by dividing the weighted average frequency by the sum of the amplitudes.

2. A wave's ZCR is the rate at which its sign shifts, going from negative to positive or the other way around.

3. The frequency under which 85% of the amplitude distributions are localized directly correlates with the spectral roll-off.

4. Chroma Frequency: There are numerous methods for converting an audio recording into a chromagram. The mean values of an audio signal's chroma characteristics were determined by the researchers.

Spectral centroid, zero-crossing rate, spectral roll-off and Chroma frequency were calculated or derived using the built-in Python library method.

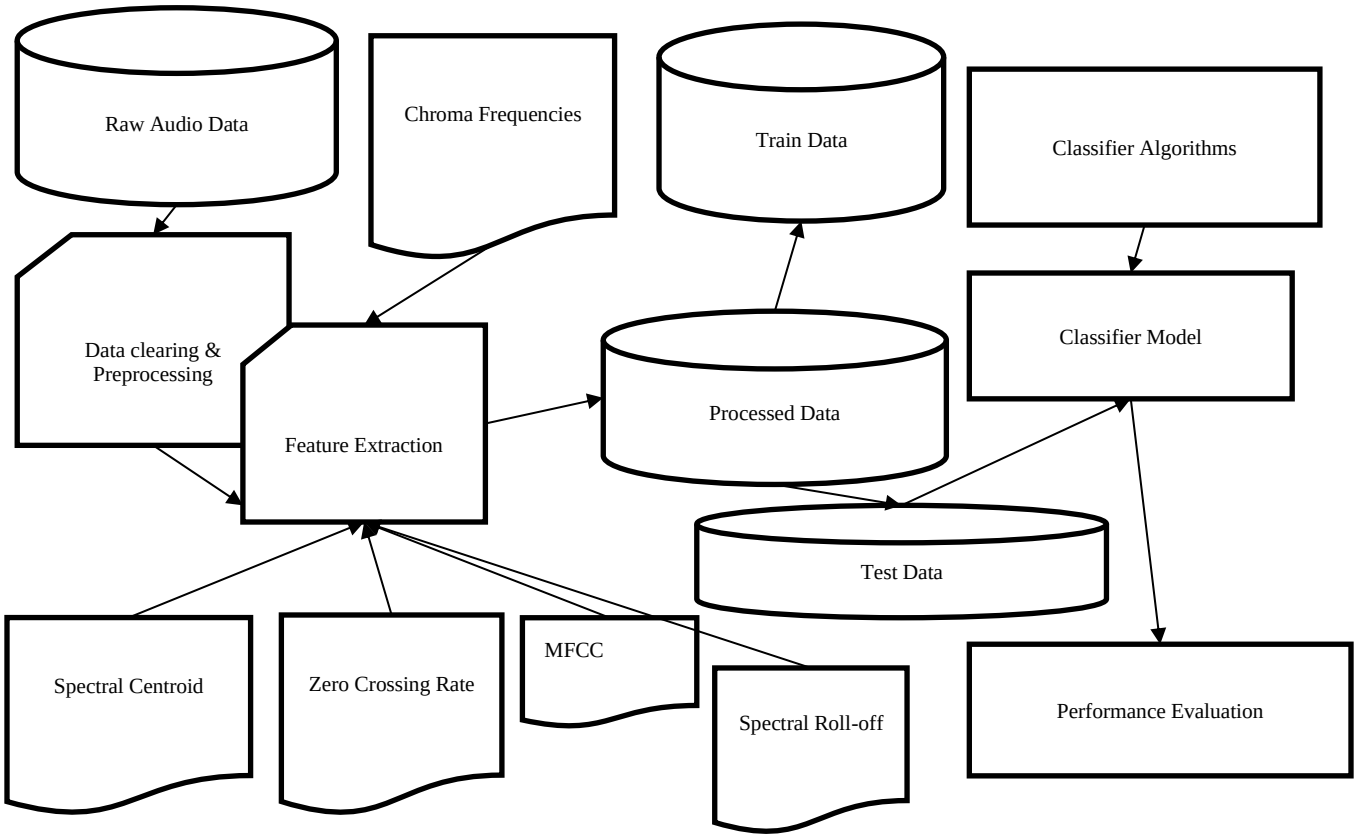

Fig. 1 System flow architecture by the researchers [16]

Note: All feature values from raw audio files are included in a CSV file and basically are processed data. An MP3 file's many frequency domain and time domain attributes have been examined to produce a unique dataset, with the information saved in a CSV format. The classifier models are trained and tested using this file. The MLPC with SVM and, alternatively, MLPC with ANN were used in the classification experiment. The SVM classifier performs better than the NN classifier when using the same dataset type. [18]. About 100 speakers, or 4095 MP3 files, have been utilized in the dataset. Table 1 shows the experiment's alphabets (28 consonants and 11 Bangla vowels, for a total of 39 alphabets that are categorized into distinct groups).

Following feature extraction (70%: 30%) and labeling, the resultant dataset was divided into "Training" and "Testing" sets. Support Vector Machine Classifiers (SVMC) and Multilayer Perceptron Classifiers (MLPC) have been evaluated for performance on prepared datasets. The dataset's accuracy was 92.33% for the Support Vector Machine Classifier and 99.27% for the Multilayer Perceptron Classifier (Table 2).

Table 1. Alphabets used in the experiment [16]

Vowels		Consonants	
Bangla Alphabet	**Corresponding pronunciations**	**Bangla Alphabet**	**Corresponding pronunciations**
অ	O	ঝ	Jho
ই	I	ঞ	Ngeo
উ	U	শ	Sho
ঋ	Ri	ট	To
আ	A	ঠ	Tho
ঈ	Ii	ড	Do
ঊ	Uu	ধ	Dho
এ	E	ন	No
ও	Oo	র	Ro
ঐ	Oi	ত	Too
ঔ	Ou	থ	Thoo
Consonants		দ	Doo
ক	Ko	ধ	Dhoo
খ	Kho	ল	Lo
গ	Go	প	Po
ঘ	Gho	ফ	Pho
ঙ	Ngo	ব	Bo
চ	Cho	ভ	Bho
ছ	Chho	ম	Mo
জ	Jo	হ	Ho

Table 2. The experiment results for both MLPC and SVMC [16]

	Accuracy	**Precision**	**Recall**	**F1-Score**
MLPC	0.99	0.99	0.99	0.99
SVMC	0.92	0.93	0.92	0.91

Table 3. Bangla vowel recognition accuracy [17]

অ, আ, ই			
Male		**Female**	
LPC	Cepstrum	LPC	Cepstrum
75%	77%	71%	72%

According to this study, the audio features fit well with ANNs' MLPC and SVM. The technology performs well in a small-scale test. Another investigation was conducted [17] applying the LPC and cepstrum-based formant estimation techniques in the bangla vowels "অ ", "আ " and " ই ". The recognition accuracy lies between 71% and 75% both for LPC and Cepstrum analysis. Both male and female speaker's data were considered in the experiment. Here, they used a platform (1). The SHRUTI Bengali Continuous ASR Speech Corpus was developed by the IIT (IITKGP) [18], and Lancaster University in the United Kingdom developed the EMILLE Corpus [19].

3. Sufficient of Propose Task

Further study of the spoken Bangla language will promote technological development and provide more sophisticated systems to comprehend conversational speech. Therefore, it might be communicated with the computer in the same manner that it would with any human, and it would be able to respond with its logical answers. All of this is possible, and

further research has been done. The MFCC, LPC and FFT with the NN method have been applied for feature extraction and demonstrate the developed system's performance by computing the error histogram, performance evaluation with epochs, confusion matrix, receiver operating characteristics (ROC), etc.

Table 4. Bangla-recorded audio samples

Eight Bangla phonemes with properties (1.018 seconds to 1.201 seconds of duration)							
অ (/O/)	আ (/A/)	ই (/I/)	উ (/OO/)	এ (/EA/)	ও (/O/)	ঐ (/OI/)	ক (/KO/)
(Short) Vowel, Oral, Compact, Grave	(Long) Vowel, Oral, Compact	(Short) Vowel, Oral, Diffuse, Acute	(Short) Vowel, Oral, Diffuse, Grave	(Complex) Vowel, Oral, Diffuse, Acute	(Complex) Vowel, Oral, Diffuse, Grave	(Complex) Vowel, Oral, Diffuse, Grave	Consonant, Oral, Compact, Unvoiced, Grave, Lax
Eight Bangla words (1.201 seconds in length)							
অংক (Math)	আমি (I)	ইলিশ (Ilish)	উট (Camel)	কলা (Banana)	খরগোশ (Rabbit)	গরু (Cow)	ঘড়ি (Clock)
Eight Bangla commands (1.802 seconds to 2.716 seconds of time duration)							
এই কাজ কর (Do this job)	দরজা খোল (Open the door)	টেবিল পরিষ্কার কর (Clean the table)	বাম দিক যাও (Go to the left)	পশ্চিম দিক সরো (Move toward the west)	অফিস যাও (Go to the office)	এই চেয়ার আনো (Bring this chair)	জানালা বন্ধ কর (Close the window)
Six Bangla sentences (2.011 seconds to 3.213 seconds of duration)							
আমরা কলা খাই (We eat bananas)		কলা ভালো ফল (Banana is a good fruit)	ফল স্বাস্থের জন্য ভাল (Fruit is good for health)	তারা তিন বন্ধু (They are three friends)	তিন বন্ধু খেলা করে (Three friends play)		তিন বন্ধু খায় (Three friends eat)

Eight hundred utterances (speech samples) of eight Bangla phonemes (vowel and consonant), eight Bangla words, eight Bangla commands, and six Bangla sentences in native male and female recorded real voice signals (main dataset) from various age groups (table 4) have been examined. To identify patterns in Bangla speech sounds, power spectral analysis (FFT), LPC and MFCC analysis have been utilized. The deep neural network and simulation are constructed using the features obtained in FFT, LPC, and MFCC. For speech recognition, a feedforward supervised deep neural network (two hidden layers) with a maximum-likelihood condition is used to construct and test the network model using random data. Finding the most effective method for accurately identifying Bangla speech and analyzing the produced systems' performance are the goals. After feature extraction, the experiment's data for the Bangla phoneme 12708, word 22592, command 64952, and phrase is 57892. Following feature extraction from 282 to 2822, data frames were produced.

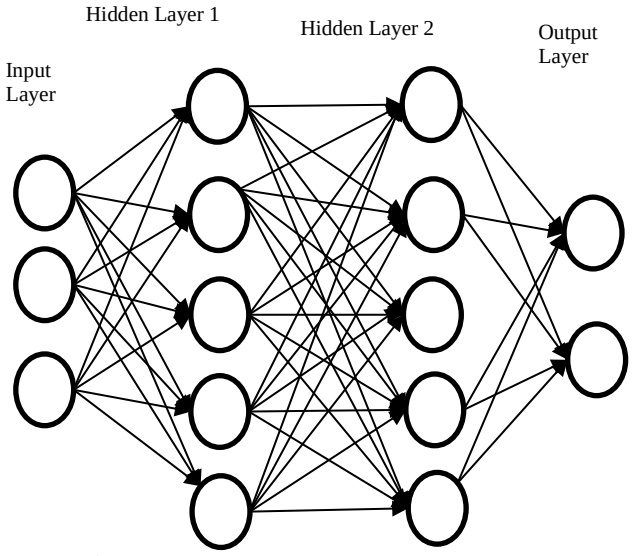

Fig. 2 Feedforward deep neural network model with two hidden layers

The deep neural network model has been developed using input data as input (train the model). Random data samples are used for 70% of training, 15% for validation, and 15% for testing following the creation of the model. We applied the model to build and examine random data for accuracy in speech recognition. Tables 5, 6, 7, and 8 present the findings.

4. Importance of the Research Contribution

There is no such standard dataset available for Bangla speech recognition. Either have to depend on the researcher's own primary dataset or other researcher if they agree to outsource their dataset. The key contribution to this research:

- Own primary dataset collected from recorded audio files (phoneme, word, command, and sentence in parallel) for all the experiments.
- Gathered different feature extraction methods and models (to build the system) in one place for experiments to see the difference, significantly contributing to the research on phoneme, word, command and sentence recognition and classification in Bangla.
- This investigation motivates future researchers to conduct more research in Bangla speech recognition.

5. Significance of the Procedure for Experimental

5.1. Calculating Short-Term Energy and Eliminating Silence [20]

Speech signals are separated into rectangular window frames of 16 milliseconds. The Short-Time Energy (STE) calculation algorithm [20] [21] [22] has been used to calculate the energy of the sound signal for improved processing and to remove the speech signal's silent region, which contains less energy. To eliminate frames with energy below 2% of the maximal energy, normalization has been done.

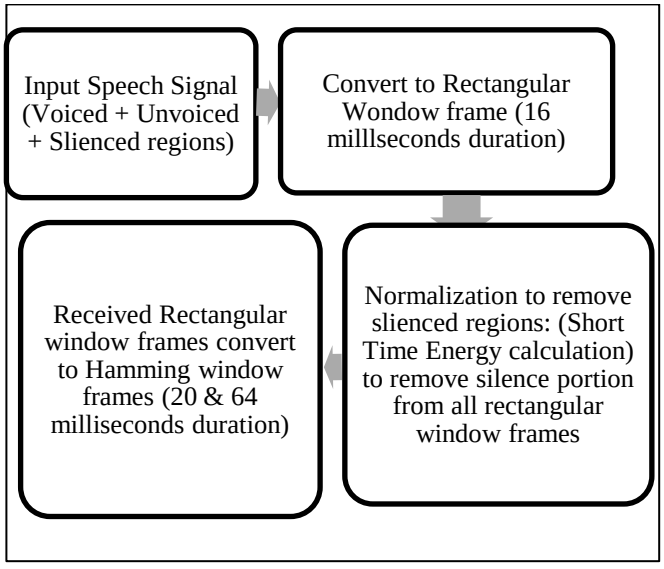

Fig. 3 Short-time energy calculation and silence removal

Equation (1) defines the rectangular window as the most basic window [21] [22]:

$$w[p] = sin(\pi p/N) = cos(\pi p/N - \pi/2) \quad (1)$$

$$where (0 \le p \le N)$$

5.2. Hamming the Frame of Windows

The Hamming window [23] is defined as follows:

$$w(p) = 0.54 - 0.46\, cos(2\pi p/N) \quad (0 \le p \le N) \quad (2)$$

The length of Window, $L = N + 1$

The audio samples have been examined in the following figure.

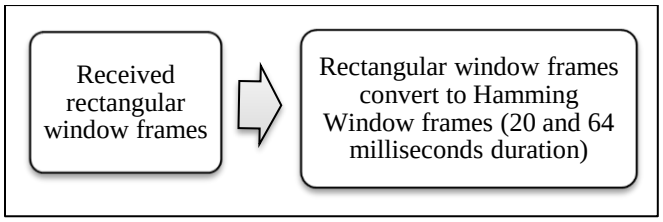

Fig. 4 Hamming window framing

5.3. Preprocessing

The voiced sections of the speech signal have a negative spectral slope, which has been compensated for by pre-emphasis.

A typical signal pre-emphasis is defined by equation 3 [24]:

$$y(p) = s(p) - Cxs(p-1) \quad where\ .9 \le C \le 1 \quad (3)$$

A filter with all zeros has been used to perform the pre-emphasis [24].

$$Preprocessing = w(p) + y(p)$$

The frame variable includes all of the frames accessible from the framing function, and each frame must undergo preprocessing. The spoken signal has been divided into many frames, avoiding the frame overlapping and zero-padding procedures used throughout. The finer structure of the spectrum may be resolved with zero padding, but the resolution cannot be improved.

5.4. Result and Discussion

The result includes-

- Characteristics of Pattern Recognition.
- Focus of the experiment.
- Expectations for the performance of the system.

6. Characteristics of Pattern Recognition

The framing signal has been converted to a sampling frequency in the frequency domain. The signal spectrum or the energy content of humans mostly lies between 20 Hz and 4 kHz (except musical voice). The spectrum (frame) of the frequency domain each time has been divided into four segments (0 kHz – 1 kHz, 1 kHz – 2 kHz, 2 kHz – 3 kHz, and 3 kHz – 4 kHz). In the experiment, a four-dimensional feature vector's mean absolute value has been taken for each frame [25] to store the mean of each band.

Power spectral analysis (FFT) has been applied to each feature. For linear predictor coefficient analysis (LPC), each frame was divided into seven segments that received seven-dimensional feature vectors [25], and LPC was applied to each feature. Similar to MFCC analysis, each frame has been divided into thirteen segments, thirteen-dimensional feature vectors have been received [25], and MFCC is then applied to each feature. These features are added to a single variable. This makes it easier to handle the dataset for training and testing.

6.1. Feature Extraction Methods

A Discrete Fourier Transform (DFT) has been applied that computes the speech signal's power spectrum [26]. A real-point Fast Fourier Transform (FFT) has been applied to increase the efficiency of speech signals. For recognition purposes, it has produced a reasonable feature space. The predictor variables in the LPC method have been applied for error reduction optimization over a small portion of the phonetic frame.

An autoregressive (AR) model has been utilized in the case of the Linear Predictor Coefficient (LPC) method. Here, q = 0, MA and ARMA [27] are used by LPC [28].

The MFCC method has been considered the third choice for the feature extraction experiment [29].

The analytical expression is defined by the equation 4:

$$y_{mel} = 2595\, log_{10}(1 + f_{Hz}/700) \quad (4)$$

The mel scale finds extensive use in speech coding and recognition. This nonlinear scale is vital for speech coding and reduces the sample space with minimal perceptual loss. To calculate the MFCC, begin with the FFT, a series of critical band filters evenly spaced along the mel scale smooths and normalizes the signal into a minimum of coefficient sets. Each time the coefficient's logarithm has been powered, the signal will be in a minimum phase. The discrete cosine transform is finally used to derive the MFCC.

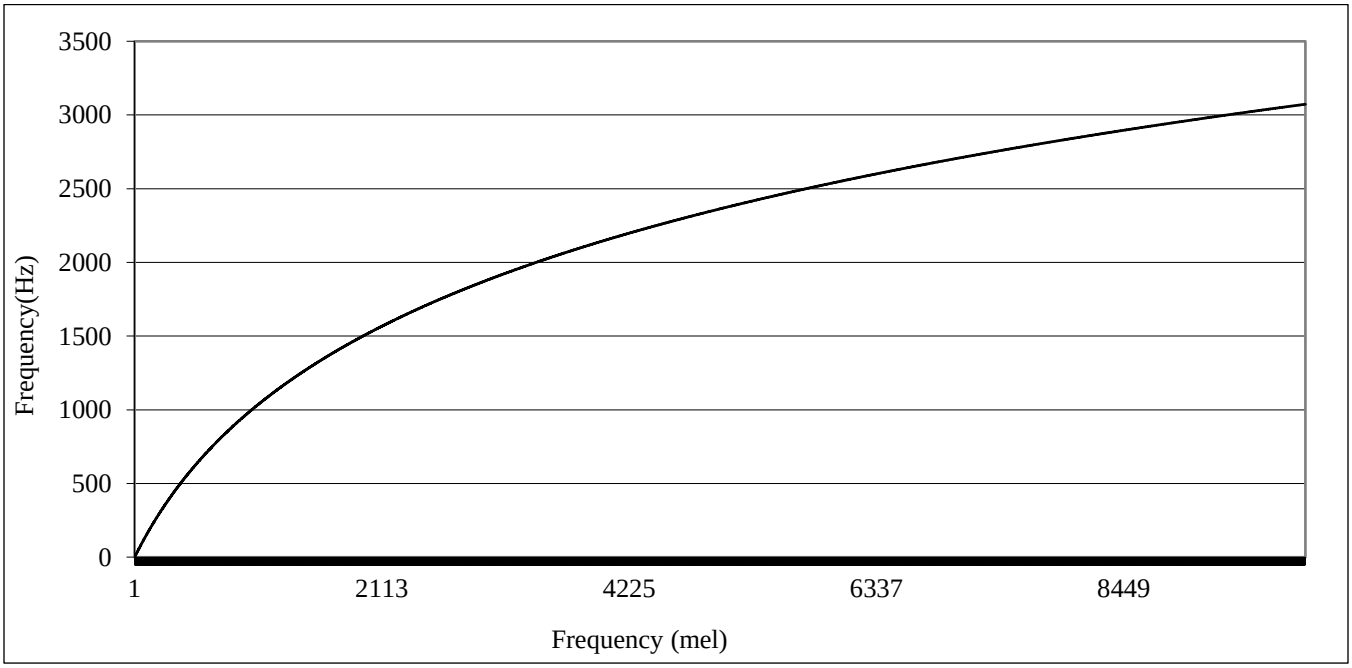

Fig. 5 Mel scale transformation

7. Focus of the Experiment

In order to search for feature extraction and recognition methods, the experiments were conducted with parallels for Bangla phonemes, words, commands, and sentences. For framing, two different lengths of the window frame (20 and 64 milliseconds length of Hamming window), respectively, were applied to all experiments. Some hidden neuron samples have also been taken (MATLAB) to train the deep neural network model. Six males-females (native speakers) of different age groups were taken for the experiment. After the

feature extraction for Bangla, audio signal data was received from 12708 to 64952, and data frames were received from 282 to 2822. To train the model for training, 70%, validation, 15% and for testing purposes, 15% random data samples were used. Finally, the model was tested using random data (Tables 5, 6, 7 and 8).

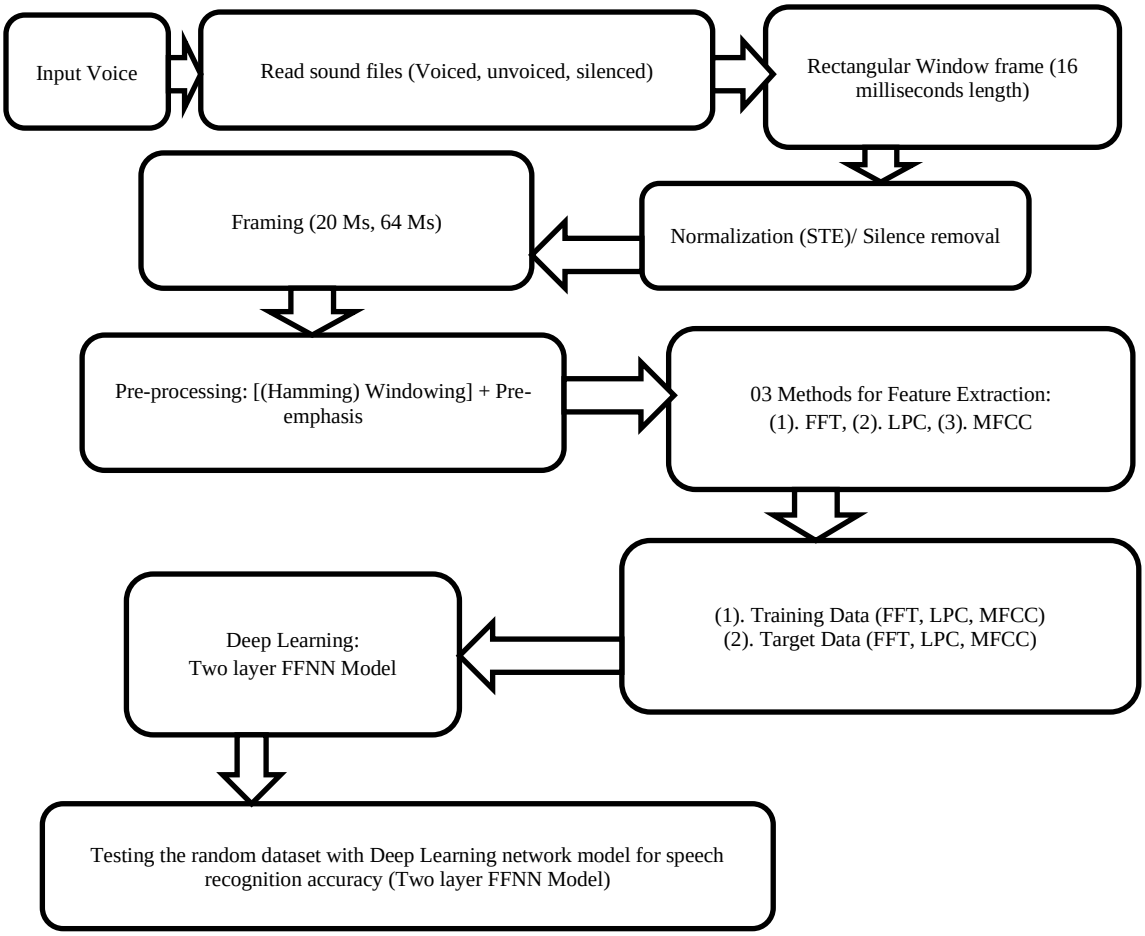

Fig. 6 Proposed system architecture

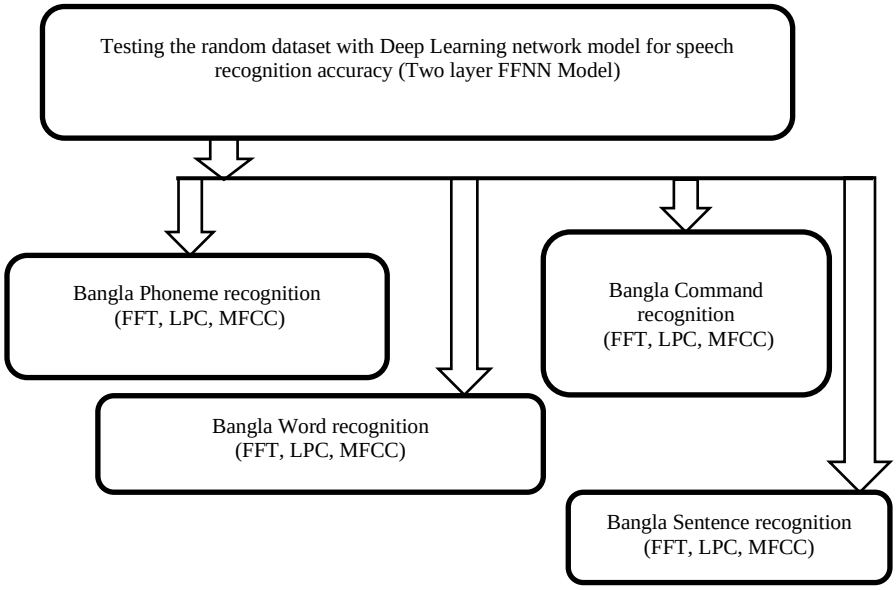

Fig. 7 Speech recognition (Deep learning model)

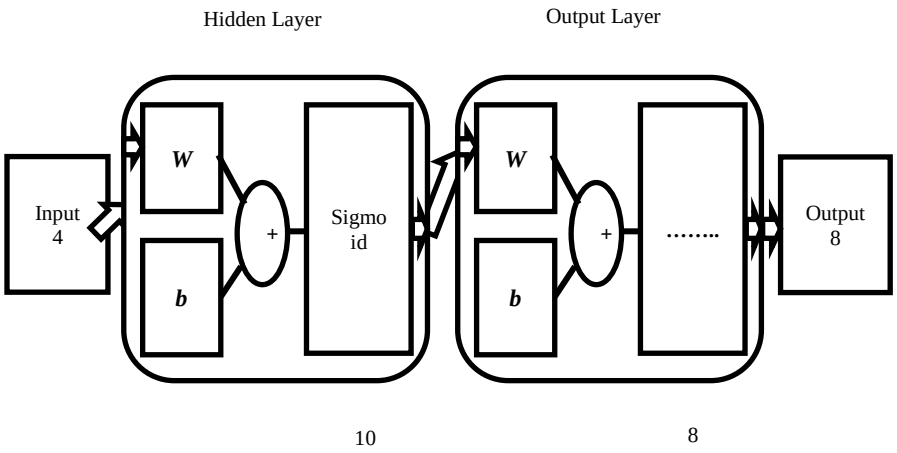

Fig. 8 Two-layer feedforward deep neural network model (using FFT feature

7.1. Bangla Phoneme, Isolated Word, Command and Sentence Recognition

Experiments in parallel have been conducted for (Table 4) applying FFT, LPC and MFCC and recognition in multi (two) layer feedforward deep neural network (hidden layer) recognition model for single male, female or up to six male-female precipitants (Tables 5, 6, 7 and 8).

Eight bangla phonemes, eight bangla words, eight bangla commands and six bangla sentences with a minimum of 30 to a maximum of 240 audio speech samples from different speakers have been used each time.

For Bangla sentence experiments, 30 to 150 speech samples were used, and around 22592 to 64952 data were received during feature extraction for each experiment. Each time, a data frame was created with a minimum of 282 to 2822.

Phoneme recognition (Table 5) using MFCC provides the best result of up to 100% when only one person is a participant. Phoneme recognition using MFCC (6 male-female speakers) the recognition rate is also good, lying between 86.66% and 99.5%, but in FFT and LPC, the percentage decreased and lies between 54.16% and 60%.

Isolated Bangla word recognition (Table 6) using MFCC provides the best results, up to 100% for a male or female speaker. Recognition using FFT is up to 75%, and LPC is up to 87.5%, which is not bad.

The word recognition rate using MFCC for 6 male-female participants is also good, that is be-tween 93.5% and 94%, but in FFT and LPC, it decreases between 43.5% and 60%. Bangla command recognition (Table 7) using MFCC provides the best result, which is up to 100% for a male person and 90% for a female person.

Recognition (for single speakers) using FFT is up to 70% and LPC 70%. The command recognition accuracy rate using MFCC for 5 male-female participants was up to 82.5%, but in FFT and LPC, it was only up to 32.5%, which is very poor.

Bangla sentence recognition (Table 5) using MFCC provides the best result, which is up to 93.33% for a male and 80% for a female. The recognition rate using FFT is up to 70%, and LPC is up to 53.33%. The command recognition accuracy rate using MFCC for 5 male-female participants is good (maximum 88.66%, whereas it is 44.66% for FFT and only 34.66% for LPC).

Table 5. Bangla Phoneme Recognition

No of Phoneme: 08	Feature Extraction Methods	Window Length (milliseconds)	Recognition Percentage	Window Length (milliseconds)	Recognition Percentage
Single male speaker (40 utterances)	FFT	20 Ms.	95%	64 Ms.	85%
	LPC	20 Ms.	82.5%	64 Ms.	90%
	MFCC	20 Ms.	**100%**	64 Ms.	95%
Single female speaker (40 utterances)	FFT	20 Ms.	72.5%	64 Ms.	92.5%
	LPC	20 Ms.	95%	64 Ms.	72.5%
	MFCC	20 Ms.	**100%**	64 Ms.	97.5%
Six male–female speakers (240 utterances)	FFT	20 Ms.	60%	64 Ms.	54.5%
	LPC	20 Ms.	54.16%	64 Ms.	55.83%
	MFCC	20 Ms.	**99.5%**	64 Ms.	86.66%

Table 6. Bangla word recognition

No of Word: 08	Feature Extraction Methods	Window Length (milliseconds)	Recognition Percentage	Window Length (milliseconds)	Recognition Percentage
Single male speaker (40 utterances)	FFT	20 Ms.	65%	64 Ms.	70%
	LPC	20 Ms.	87.5%	64 Ms.	55%
	MFCC	20 Ms.	**100%**	64 Ms.	92.5%
Single female speaker (40 utterances)	FFT	20 Ms.	75%	64 Ms.	72.5%
	LPC	20 Ms.	87.5%	64 Ms.	87.5%
	MFCC	20 Ms.	**100%**	64 Ms.	87.5%
Five male–female speakers (200 utterances)	FFT	20 Ms.	50.5%	64 Ms.	60%
	LPC	20 Ms.	43.5%	64 Ms.	47%
	MFCC	20 Ms.	93.5%	64 Ms.	**94%**

Table 7. Bangla command recognition

No of Command: 08	Feature Extraction Methods	Window Length (milliseconds)	Recognition Percentage	Window Length (milliseconds)	Recognition Percentage
Single male speaker (40 utterances)	FFT	20 Ms.	40%	64 Ms.	70%
	LPC	20 Ms.	70%	64 Ms.	65%
	MFCC	20 Ms.	**100%**	64 Ms.	**100%**
Single female speaker (40 utterances)	FFT	20 Ms.	32.5%	64 Ms.	25%
	LPC	20 Ms.	35%	64 Ms.	50%
	MFCC	20 Ms.	**90%**	64 Ms.	87.5%
Five male–female speakers (200 utterances)	FFT	20 Ms.	16.5%	64 Ms.	32.5%
	LPC	20 Ms.	21%	64 Ms.	32%
	MFCC	20 Ms.	**82.5%**	64 Ms.	72.5%

Table 8. Bangla sentence recognition

No of Sentence: 06	Feature Extraction Methods	Window Length (milliseconds)	Recognition Percentage	Window Length (milliseconds)	Recognition Percentage
Single male speaker (30 utterances)	FFT	20 Ms.	66.66%	64 Ms.	70%
	LPC	20 Ms.	53.33%	64 Ms.	50%
	MFCC	20 Ms.	**93.33%**	64 Ms.	90%
Single female speaker (30 utterances)	FFT	20 Ms.	50%	64 Ms.	63.33%
	LPC	20 Ms.	43.33%	64 Ms.	46.66%
	MFCC	20 Ms.	**80%**	64 Ms.	76.66%
Five male-female speakers (150 utterances)	FFT	20 Ms.	42.66%	64 Ms.	44.66%
	LPC	20 Ms.	34.66%	64 Ms.	31.33%
	MFCC	20 Ms.	**88.66%**	64 Ms.	54%

8. Expectations for the Performance of the System

Table 9, about Bangla phoneme, word, command, and sentence recognition, shows the performance of the system used for deep learning for Bangla speech recognition. The result reveals how good the approaches are for Bangla speech recognition. An asterisk '*' denotes mean squared error, which is an average squared difference between outputs and targets. Minor values are considered good. Zero means no error.

Table 9. Bangla phoneme, word, command and sentence recognition in a deep neural network model

Bangla phoneme						
	Feature Extraction Methods	**Window Length (in milliseconds)**	***Performance Evaluation with Epoch = E**	***Training State (Gradient, Epoch = E)**	***Error Histogram (Bins or B = 20 taken)**	**Confusion Matrix (Overall accuracy = OA, Error = Er)**
	FFT	20 Ms	0.156, E128	0.003, E134	0.048	**OA53%, Er46%**

08 different phonemes (240 samples), [6 male-female speakers]		64 Ms	0.134, E66	0.011, E72	0.046	OA66%, Er33%
	LPC	20 Ms	0.172, E217	0.010, E223	0.041	OA52%, Er47%
		64 Ms	0.180, E77	0.022, E83	0.047	OA51%, Er49%
	MFCC	20 Ms	**0.094, E67**	**0.008, E73**	**0.047**	**OA83.5%,Er16.5%**
		64 Ms	0.070, E66	0.013, E72	0.037	OA84%, Er15%
Bangla word						
08 different words (200 samples), [5 male-female speakers]	FFT	20 Ms	0.196, E62	0.011, E68	0.049	OA40%, Er59%
		64 Ms	0.180, E56	0.009, E62	0,047	OA52%, Er48%
	LPC	20 Ms	0.208, E83	0.008, E89	0.035	OA39%, Er60%
		64 Ms	0.191, E102	0.011, E108	0.008	OA52%, Er48%
	MFCC	20 Ms	0.123, E89	0.014, E95	0.049	OA70%, Er29%
		64 Ms	0.125, E70	0.020, E76	0.043	OA82%, Er18%
Bangla Command						
08 different commands (200 samples), [5 male-female speakers]	FFT	20 Ms	0.242, E101	0.003, E107	0.010	OA24%, Er75%
		64 Ms	0.228, E76	0.006, E82	0.042	OA36%, Er63%
	LPC	20 Ms	0.246, E41	0.008, E47	0.037	OA24%, Er75%
		64 Ms	0.239, E33	0.008, E39	0.040	OA31%, Er68%
	MFCC	20 Ms	0.167, E239	0.009, E245	0.045	OA54%, Er45%
		64 Ms	0.175, E98	0.013, E104	0.044	OA61%, Er38%
Bangla Sentence						
06 different sentences (uttered 150 times) [5 male-female speakers]	FFT	20 Ms	0.273, E238	0.003, E244	0.013	OA29%,Er70.8%
		64 Ms	0.268, E47	0.011, E53	0.044	OA35%, Er65%
	LPC	20 Ms	0.283, E71	0.003, E77	0.022	OA28%,Er71.6%
		64 Ms	0.289, E27	0.008, E33	0.009	OA36%, Er64%
	MFCC	20 Ms	0.217, E217	0.010, E223	0.040	OA51.9%,Er48.1%
		64 Ms	0.236, E71	0.014, E77	0.002	OA52%,Er47.4%

Two asterisks '**' indicate regression R values to measure the correlation between outputs and targets.

R = 1 denotes a close relationship, whereas R = 0 denotes a random relationship.

The Receiver operating characteristics (ROC) * should be (1- values) within range 1 for all cases.

8.1. Performance Evaluation

Figure 9 shows how good the performance evaluation of the system for speech recognition is during deep neural network model training for Bangla word recognition (using the MFCC feature). The validation performance close to zero (mean squared error rate very close to zero) with only a few epochs denotes that speech recognition has considerable potential. The performance evaluation is 0.123 with Epoch E = 89.

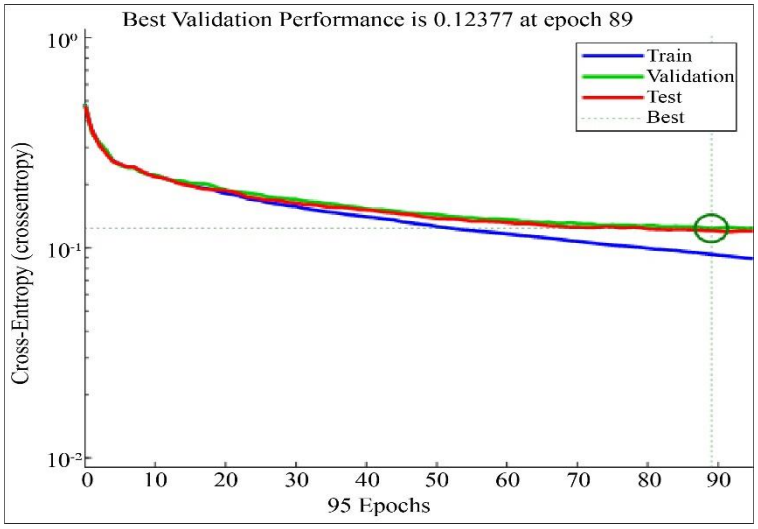

Fig. 9 Deep neural network training (performance evaluation)

8.2. Deep Neural Network Training

Figure 10 Performance evaluation of Bangla phonemes. Feature extraction using MFCC (LPC and FFT also conducted) and recognition in a deep (feedforward) neural network model where 08 different phonemes with 240 speech samples (male-female) and 20 milliseconds of window length has been taken. For all the cases, the gradient point is close to zero (0.009) with only a few epochs (E245), which indicates it is a potential approach for speech recognition.

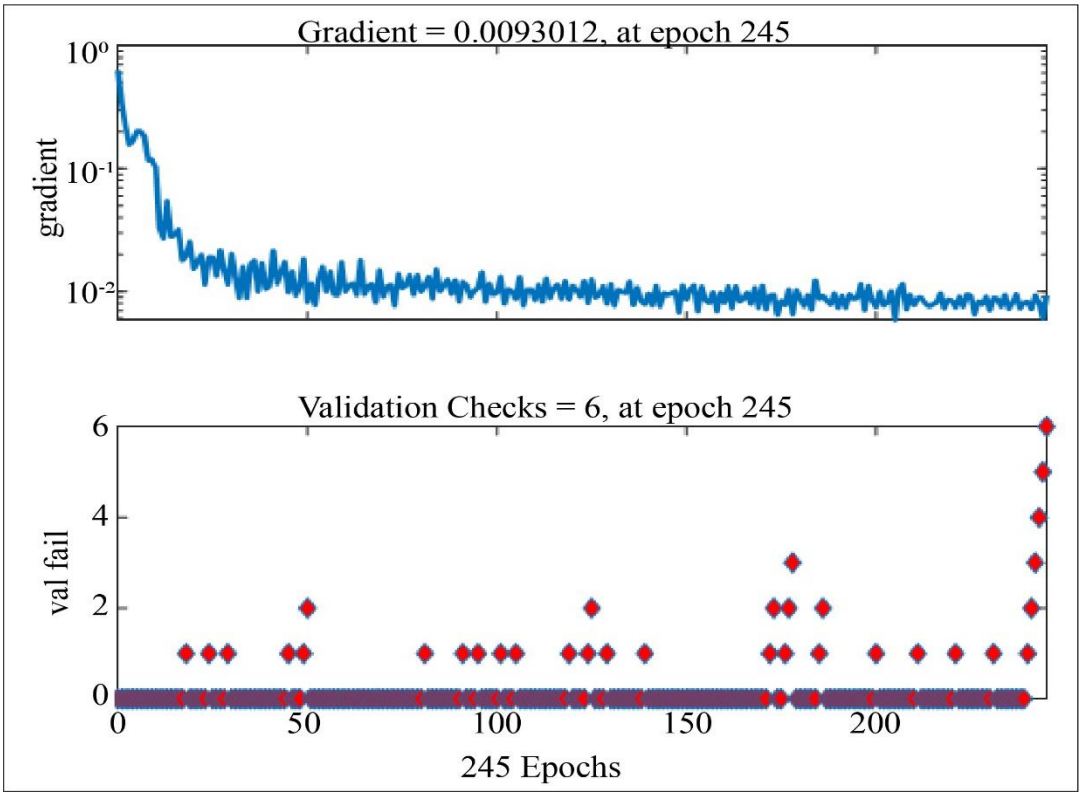

Fig. 10 Deep neural network model training (using MFCC feature)

8.3. Error Histogram

Figure 11 shows Bangla sentence recognition using MFCC + FFNN. The experiment showcases that the MFCC error histogram result is 0.040 (20 bins), very close to zero.

Therefore, the recognition is rate potential. A total of 150 speech samples from both males and females have been taken. All the speech's window frames are 20 milliseconds long.

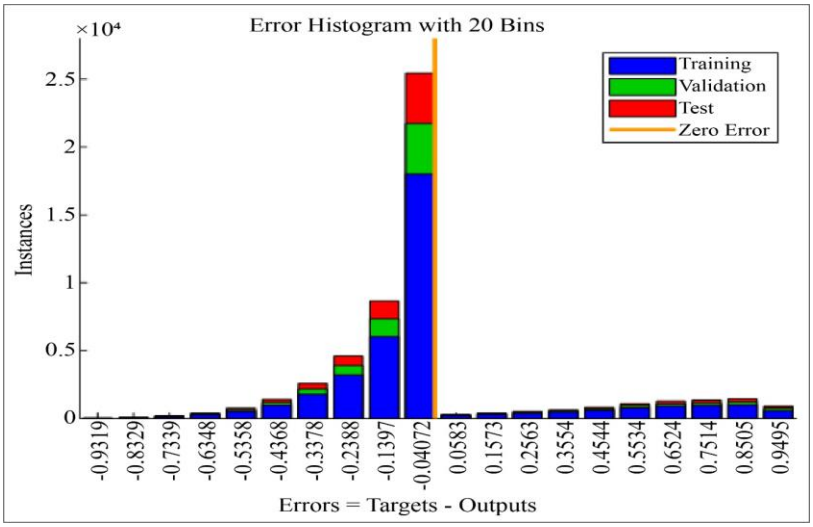

Fig. 11 Error histogram in deep neural network model training (using MFCC feature)

8.4. Confusion Matrix

Figure 12 showcases the confusion matrix for the feature extraction and recognition of eight different Bangla phonemes using MFCC in a deep neural network model (multiple samples for each alphabet utilized). In the confusion matrix results, where an Overall Accuracy (OA) rate of 83.5% and an error (Er) rate of only 16.5% were found, a random dataset was used for testing purposes.

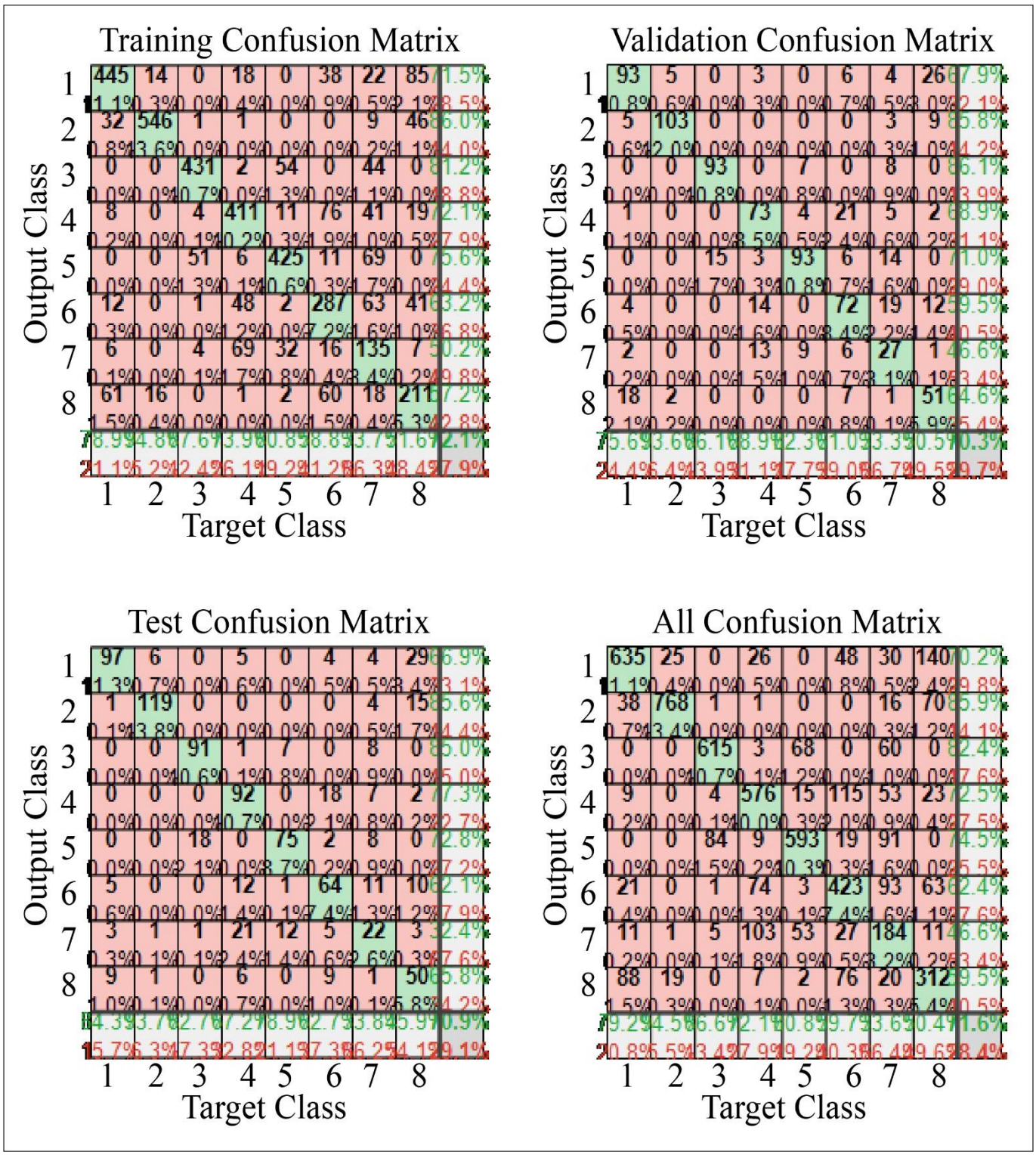

Fig. 12 Confusion matrix in deep neural network model training (using MFCC feature)

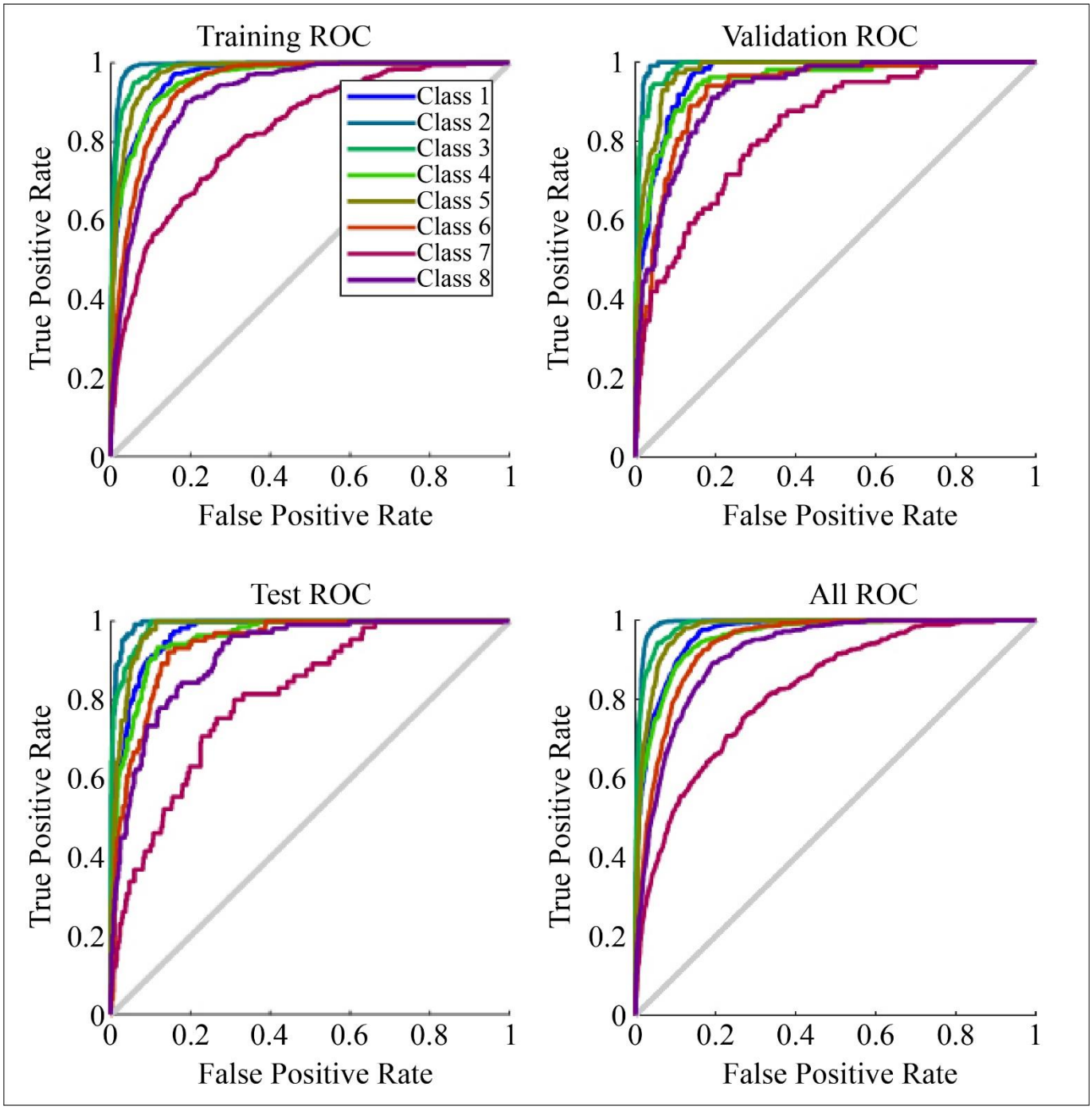

Fig. 13 ROC in Deep neural network model training (using MFCC feature

8.5. Receiver Operating Characteristics (ROC)

Figure 13 illustrates the feature extraction of Bangla phonemes by applying MFCC and feedforward deep neural network models (different phonemes) as recognition tools.

The 240 samples have been taken from male and female speakers. The receiver operating characteristic (ROC) results, especially for all ROC curves, are a true positive rate (TPR) close to 1 compared to the False Positive Rate (FPR).

9. Conclusion

Bangla words, sentences, commands, and phonemes are all taken into consideration for analysis. The validation performance assessment (epoch) technique, error histogram, confusion matrix, Receiver Operating Characteristics (ROC), and other metrics are used to assess the outcomes. The tools and procedures were effective for recognizing Bangla words and phonemes with several participants. Sometimes, the accuracy of the recognition is decreased by increasing the

number of participants and voice samples. The deep neural network model's pattern recognition and supervised training performed well for Bangla words and phonemes (male and female). The accuracy rate is influenced by the voice of the opposing gender and the various window frames. With the exception of MFCC, FFT and LPC are not effective feature extraction techniques for command and sentence recognition for larger numbers of participants. Time Delays Neural Network (TDNN), or HMM, can be used for voice recognition, and delta-MFCC can be used as a feature extraction technique for improved outcomes in a large-scale experiment.

References

[1] Syed Akhter Hossain, M. Lutfar Rahman, and Farruk Ahmed, "Spectral Analysis of Bangla Vowels," *Pakistan Section Multitopic Conference*, Karachi, Pakistan, pp. 1-5, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ghulam Muhammad, Yousef A. Alotaibi, and Mohammad Nurul Huda, "Automatic Speech Recognition for Bangla Digits," *12th International Conference on Computers and Information Technology*, Dhaka, Bangladesh, pp. 379-383, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[3] Umme Muslima, and M. Babul Islam, "Experimental Framework for Mel-Scaled LP Based Bangla Speech Recognition," *16th International Conference on Computer and Information Technology*, Khulna, Bangladesh, pp. 56-59, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[4] Anup Kumar Paul, Dipankar Das, and Md. Mustafa Kamal, "Bangla Speech Recognition System Using LPC and ANN," *7th International Conference on Advances in Pattern Recognition*, Kolkata, India, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[5] Qamrun Nahar Eity et al., "Bangla Speech Recognition using Two Stage Multi-Layer Neural Networks," *2010 International Conference on Signal and Image Processing*, Chennai, pp. 222-226, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[6] Mohammed Rokibul Alam Kotwal et al., "Bangla Phoneme Recognition Using Hybrid Features," *International Conference on Electrical and Computer Engineering (ICECE 2010)*, Dhaka, Bangladesh, pp. 718-721, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[7] Foyzul Hassan, Mohammed Rokibul Alam Kotwal, and Mohammad Nurul Huda, "Bangla Phonetic Feature Table Construction for Automatic Speech Recognition," *16th International Conference on Computer and Information Technology*, Khulna, Bangladesh, pp. 51-55, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[8] Mohammed Rokibul Alam Kotwal et al., "Bangla Phoneme Recognition for Different Acoustic Features," *2010 International Conference on Computer Applications and Industrial Electronics*, Kuala Lumpur, Malaysia, pp. 543-547, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[9] Rhythm Shahriar et al., A Communication Platform between Bangla and Sign Language, *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, Bangladesh, pp. 1-4, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[10] Mohammad Mahedi Hasan et al., "Bangla Triphone Hmm Based Word Recognition," *2010 IEEE Asia Pacific Conference on Circuits and Systems*, Kuala Lumpur, Malaysia, pp. 883-886, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[11] Md. Shahadat Hossain et al., "Evaluation of Bangla Word Recognition Performance using Acoustic Features," *International Conference on Computer Applications and Industrial Electronics*, Kuala Lumpur, Malaysia, pp. 490-494, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[12] Bulbul Ahamed et al., "Effect of Speaker Variation on the performance of Bangla ASR," *2013 International Conference on Informatics Electronics and Vision (ICIEV)*, Dhaka, Bangladesh, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nafis Sadeq et al., "Bangla Voice Command Recognition in End-To-End System Using Topic Modeling Based Contextual Rescoring," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] S.M. Saiful Islam Badhon et al., "State of art Research in Bengali Speech Recognition, *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Shafkat Kibria et al., "Acoustic Analysis of Accent-Specific Pronunciation Effect on Bangladeshi Bangla: A Study on Sylheti Accent," *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, Sylhet, Bangladesh, pp. 1-4, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Md Gulzar Hussain et al., "Classification of Bangla Alphabets Phoneme Based on Audio Features Using MLPC and SVM," *2021 International Conference on Automation Control and Mechatronics for Industry 4.0 (ACMI)*, Rajshahi, Bangladesh, pp. 1-5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tonmoy Ghosh et al., "Formant Analysis of Bangla Vowel for Automatic Speech Recognition," *Signal and Image Processing International Journal*, vol. 7, no. 5, 2016. [Google Scholar] [Publisher Link]

[18] Biswajit Das, Sandipan Mandal, and Pabitra Mitra, "Bengali Speech Corpus for Continuous Automatic Speech Recognition System," *2011 International Conference on Speech Database and Assessment (Oriental COCOSDA)*, Hsinchu, Taiwan, pp. 51-55, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[19] The Emille Corpus, Lancaster University, 2025. [Online]. Available: http://www.lancaster.ac.uk/fass/projects/corpus/emille/

[20] Muhammad Asadullah, and Shibli Nisar, "A Silence Removal and Endpoint Detection Approach for Speech Processing," *Sarhad University International Journal of Basic and Applied Sciences*, vol. 4, no. 1, pp. 10-15, 2016. [Google Scholar] [Publisher Link]

[21] John R. Deller, John H.L. Hansen, and John G. Proakiset, *Discrete-Time Processing of Speech Signals*, Wiley, 1999. [Google Scholar] [Publisher Link]

[22] Marina Bosi, and Richard E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Springer, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[23] Alan V. Oppenheim, Ronald W. Schafer, *Discrete-Time Signal Processing*, Pearson Education, 3rd ed., 2010. [Google Scholar] [Publisher Link]

[24] R. Vergin, and D. O'Shaughnessy, "Pre-Emphasis and Speech Recognition," *Proceedings 1995 Canadian Conference on Electrical and Computer Engineering*, Montreal, QC, Canada, pp. 1062-1065, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[25] Simon Haykin, and Michael Moher, *Communication Systems*, John Wiley and Sons Inc., 5th ed., 2009. [Google Scholar] [Publisher Link]

[26] Richard A. Haddad, and Thomas W. Parsons, "Digital Signal Processing: Theory, Applications and Hardware," *Computer Science Press*, New York, USA, 1991. [Google Scholar] [Publisher Link]

[27] Jean-Claude Junqua, and Jean-Paul Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, Springer US, pp. 1-440, 1996. [Google Scholar] [Publisher Link]

[28] Lawrence R. Rabiner, *Digital Processing of Speech Signals*, Prentice-Hall, 1978. [Google Scholar]

[29] Md. Sahidullah, and Goutam 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. [CrossRef] [Google Scholar] [Publisher Link]