IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Mathematical Modelling of Engineering Problems
Vol. 12, No. 7, July, 2025, pp. 2387-2404
Journal homepage: http://iieta.org/journals/mmep

Bangla Speech Processing: An Analytical Study of Feature Extraction and Recognition Methods

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

[1] Department of Computer Science and Engineering, Uttara University, Dhaka 1230, Bangladesh
[2] Department of Computer Science and Engineering, Islamic University, Kushtia 7003, Bangladesh
[3] Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur, Nilphamari 5310, Bangladesh
[4] Department of Information and Communication Technology, Islamic University, Kushtia 7003, Bangladesh
[5] Department of Mathematics, Uttara University, Dhaka 1230, Bangladesh

Corresponding Author Email: mannan.iu31@gmail.com

https://doi.org/10.18280/mmep.120718

Received: 10 April 2025
Revised: 5 June 2025
Accepted: 11 June 2025
Available online: 31 July 2025

Keywords:
Linear Predictor Coefficient (LPC), Mel Frequency Cepstral Coefficient (MFCC), Power Spectral Analysis (FFT), Time Delay Neural Network (TDNN), Hamming Window (HM), Hanning Window (HN), Blackman Window (BL), Levenberg-Marquardt Algorithm (LMA)

ABSTRACT

Speech recognition has always been an interesting yet challenging task for researchers, especially when working with Bangla, which is complex due to its linguistic structure. This research is extensive in scale, encompassing Bangla phonemes, isolated Bangla words, commands, and sentences in the experiments. Bangla speech recognition is a comparison analysis in large scale that focuses on different feature extraction techniques, recognition tools, window frame feature, other methods and techniques applied. A system is developed by writing code in MATLAB. Mel Frequency Cepstral Coefficient (MFCC), Power Spectral Analysis (FFT), and Linear Predictor Coefficient Analysis (LPC) methods are utilized as feature extraction techniques. Time delays neural network (time series) and a two-layer feed forward hidden neural network are used as speech recognition tools. The maximum likelihood method is also incorporated to enhance the accuracy of speech recognition. Blackman, Hamming, and Hanning Window frame techniques are applied in parallel during feature extraction to observe their influences on speech recognition accuracy. The datasets gathered from native speakers. MFCC as a feature extraction technique, combined with two-layer Feed Forward Neural Network (FFNN) or TDNN as speech recognition tools, outperforms FFT and LPC with the deep learning tools. The study discovered that both the quantity of speech samples, the opposite gender's voice, and different windowing techniques all had an impact on the recognition accuracy rate. This study will encourage researchers to conduct further research to advance Bangla speech recognition.

1. INTRODUCTION

Bangla or Bengali as a regional language has received less research attention compared to English and other widely spoken languages. However, 300 million people worldwide currently speak Bangla. This study is addressing a crucial and unexplored field. A worthy and essential first step is the comparison analysis that focuses on different feature extraction techniques and recognition tools. The complexity of the Bangla language and the use of compound characters make speech recognition in this language extremely difficult. Mel Frequency Cepstral Coefficient (MFCC), Power Spectral Analysis (FFT), and Linear Predictor Coefficient (LPC) are all effective feature extraction techniques that we are taking into consideration. Combining them with other recognition models, like as the Time Delay Neural Network (TDNN) (time series) and the Feed Forward Neural Network (FFNN), will give a thorough grasp of what functions best for Bangla voice recognition. This study could greatly progress this area and offer more reliable solutions to Bangla speakers around the world.

2. AN OVERVIEW OF THE LANGUAGE OF BANGLA

Bangla is a vibrant and diverse language with a rich cultural heritage. In addition to Bangladesh and West Bengal, there are also significant Bangla-speaking communities in places like Assam, Tripura, and the Andaman and Nicobar Islands in India. The journey from Vedic Sanskrit to Bangla is indeed a testament to the dynamic nature of languages and their ability to adapt and evolve over centuries. The influence of dialects like Magadhi and Ardha-Magadhi, followed by Magadhi-Apabhransa, showcases the rich tapestry of linguistic and cultural changes that shaped the language we know today. This historical progression also highlights the deep roots of Bangla in the broader Indo-Aryan language family. It's incredible to think about how these ancient dialects and languages have

contributed to the development of Bangla, which continues to thrive and evolve in the modern era [1]. It is a clear and insightful breakdown of the historical phases of the Bangla language, by dividing Bangla's history into Formative, Middle, and Modern periods, we can appreciate the significant developments in each era. During the Formative period, the language began to crystallize, laying the foundation for its distinct identity. The Middle period saw the further development of its literary traditions and cultural expressions. Finally, the Modern period has been marked by the standardization and globalization of Bangla, making it one of the most widely spoken languages in the world.

The influence of other Southeast Asian language families, such as Tibeto-Burman and Austro-Asiatic, adds to the richness and diversity of Bangla. This blending of linguistic elements has helped shape Bangla into a unique and multifaceted language [2]. Bangla is a rich tapestry of influences from various linguistic families. Its classification as an Indo-European language primarily accounts for its core structure and a significant portion of its vocabulary. However, the contributions from Tibeto-Burman and Austro-Asiatic languages have added depth and diversity to Bangla, especially in terms of vocabulary and certain grammatical features [3]. The Bangla script, derived from the ancient Brahmi alphabet, shares a close relationship with the Devanagari alphabet before diverging in the 11th century AD. The coexistence of Chaltibhasa (informal speech) and Sadhubhasa (formal speech) within Bangla showcases the language's versatility and adaptability to different contexts. Chaltibhasa is more casual and commonly used in everyday conversations. It simplifies vocabulary and often contracts pronouns and verbs, making it easier and quicker to communicate. Its widespread use reflects the dynamic nature of spoken Bangla and how it evolves to suit the needs of daily interactions. Sadhubhasa, on the other hand, has a more formal and classical tone. Influenced heavily by early Bangla poetry, it became the standard for literature, business, and formal communication by the 19th century. Sadhubhasa's rich and elaborate style made it suitable for written texts, but it was less practical for everyday speech. By the 21st century, Chaltibhasa had emerged as the dominant style not only for casual conversations but also for contemporary literature, reflecting the changing preferences and communication needs of Bangla speakers. The 1936 spelling reforms initiated by the University of Calcutta were a significant step towards standardizing Bangla. However, the lack of consensus among different institutions, like the Bangla Academy in Dhaka and Visva-Bharati University in West Bengal, has added layers of complexity to this effort. The fact that many publishers and newspapers adopt their own styles further complicates the landscape. These differing approaches can sometimes hinder the creation of a unified, standardized form of Bangla, despite the genuine efforts made by various groups. It's a testament to the dynamic and evolving nature of languages that such diversity exists, but it also poses challenges for achieving uniformity. The continued work of language researchers and institutions in this field is crucial for navigating these complexities and striving towards a more standardized Bangla. The multitude of efforts by different groups, while well intentioned, has indeed created a fragmented approach to standardizing the Bangla language. This fragmentation has slowed the development of a uniform standard, making it difficult to achieve consistency. The lack of comprehensive research in rule-based or stochastic processing to address syntactic ambiguities is a significant hurdle. However, the emergence of language technology is a promising development. By combining linguistic and technical expertise, researchers are working towards innovative solutions to standardize and improve Bangla. This multidisciplinary approach can potentially address the complexities and nuances of the language, leading to more effective standardization efforts. It's an exciting time for language researchers, as they explore new ways to harmonize and advance the Bangla language through technological and linguistic advancements.

It's clear that the journey to standardization is ongoing, but the efforts in Language Technology offer hope for meaningful progress [4].

3. HISTORICAL VIEW OF SPEECH ANALYSIS

Acoustic-phonetics played a crucial role in the foundational stages of automatic speech recognition (ASR). Understanding the elements of speech and their realization in spoken language was essential for early researchers. In 1952, Davis et al. [5] of Bell Laboratories built a system for isolated digit recognition for a single speaker, using the spectral resonances during vowel regions of each digit. In 1956, Olson and Belar [6] in RCA Laboratories tried to recognize ten syllables of a single taker. Forge's work at MIT Lincoln Laboratory in 1959 was groundbreaking in the field of speech recognition. By focusing on a speaker-independent system, they were addressing a major challenge in the field: the variability in speech among different individuals [7]. Later, the 1960s saw significant advancements in speech recognition technology, with a particular focus on developing specialized hardware. Japanese laboratories made notable contributions during this period, including the pioneering work of Suzuki and Nakata at the Radio Research Lab in Tokyo [8]. Sakai and Doshita [9]'s work at Kyoto University on the phoneme recognizer was a significant advancement in the field of speech recognition. By incorporating a speech segmenter, they were able to dissect the speech signal into different portions for more precise analysis and recognition, NEC Laboratories also made significant contributions to the field of speech recognition in the 1960s. Their work on digit recognition was particularly noteworthy [10]. In 1959, Fry [11] developed a system aimed at recognizing phonemes, which are the distinct units of sound in a language. Their system specifically focused on recognizing four vowels and consonants in the English language, they used statistical syntax in speech recognition for the first time. In 1960, Vintsyuk first proposed the use of dynamic programming for time-alignment between two utterances in order to derive a meaningful matching score. Dynamic programming (DP) has played a crucial role in advancing ASR since the late 1970s [12].

This excerpt highlights the early development directions in speech-recognition research during the 1970s, with IBM and Bell Laboratories representing two distinct approaches [12]. IBM's Focus on Speaker-Dependent Systems Led by Jelinek. IBM's effort was centered on developing a voice-activated typewriter, with a system designed to respond to a single user or a small group of users who could "train" the system to recognize their speech patterns. This approach was highly speaker-dependent, meaning it relied on the voice of a particular user, and thus would not function well for people outside of this specific training set. And Bell Laboratories' Focus on Speaker-Independent Systems that could handle

multiple speakers, including those with different regional accents or speech characteristics. The goal was to make the system speaker-independent, meaning it could work for any user without the need for individual training [13]. Reddy's contributions at Carnegie Mellon University were indeed pivotal in the evolution of speech recognition technology. His work introduced innovative concepts that have had a lasting impact on the field. Reddy was one of the first to advocate for dynamic phoneme tracking, which involves continuously monitoring and recognizing phonemes in a stream of speech. Later, Reddy proposed a knowledge integration approach to speech recognition and understanding [14], while the difference in goals led to different realizations rapid development of statistical methods in the 1980s, namely the Hidden Markov Model (HMM) framework [15], had caused a certain degree of convergence in the system design. During 1990 till 2000 were a transformative decade for ASR technology, marked by significant advancements and the integration of new methodologies. The widespread adoption of HMMs for modeling speech dynamics and the integration of neural networks into ASR systems began to gain traction. Neural networks offered a way to model complex, non-linear relationships in speech data, leading to improvements in recognition accuracy. Advances in language modeling, including the use of n-gram models and later, more sophisticated techniques like neural language models, helped improve the context-awareness of ASR systems, Techniques for adapting ASR systems to individual speakers became more refined, allowing for better performance in speaker-dependent applications [16]. The 2020 period was a pivotal time for ASR, with innovations that continue to influence the field today to Deep Neural Networks (DNNs). The integration of DNNs into ASR systems revolutionized the field. DNNs provided a way to model complex, non-linear relationships in speech data, leading to significant improvements in recognition accuracy [17]. The shift towards end-to-end models, which consolidated various components of ASR into a single neural network, streamlined the recognition process and improved performance [18].

4. THE PRESENT STATE OF SPEECH RECOGNITION IN BANGLA

Approximately 300 million individuals worldwide speak Bangla. However, ASR research in Bangla still lacks quality and depth. Phoneme recognition by 40 native speakers was examined using MFCC, LF, and HMM techniques, with MFCC providing higher accuracy than LF methods [19]. Record-Extract-Approximate-Distinguish (READ) is a Bangla phoneme recognition system that claims to achieve 98.35% accuracy in recognizing Bangla vowel phonemes using a specific number of speech samples. However, the Indian Bangla accent (West Bengal) differs from the Bangladeshi Bangla accent, and the study did not include any Bangla consonant phonemes in the experiments [20]. Using MFCC features extracted from a Bangla phoneme corpus, this study compares two approaches for Bangla phoneme recognition in ASR systems: a Multi-Layer Neural Network (MLN) and an HMM. The objective is to identify the most efficient method for accurately recognizing Bangla phonemes in speech. The research utilizes a primary Bangla phoneme corpus, integrating MFCC with both MLN and HMM models, and compares their performance to determine the more effective approach [21]. A medium-sized Bangla speech corpus, consisting of data from 40 native Bangla speakers, was prepared to compare the performance of different acoustic features in Bangla word recognition. MFCCs were used as input to triphone HMM-based classifiers to assess word recognition performance. Experimental results demonstrate that the MFCC-based method, with 39 dimensions, achieves higher Word Correct Rate (WCR) and Word Accuracy (WA) compared to the other methods investigated [22].

Analysis of the effects of speaker variation on performance has shown that accent-related speaker variation significantly affects WCR, WA, and Sentence Correct Rate (SCR). It is anticipated that this experimental analysis will help researchers mitigate the impact of speaker variance and improve the effectiveness of ASR systems [23]. Contextual rescoring using multi-label topic modeling has improved the performance of an end-to-end Bangla voice command recognition system. By combining Connectionist Temporal Classification (CTC), attention mechanisms, and RNN with Latent Dirichlet Allocation (LDA), the Word Error Rate (WER) was reduced from 16.7% to 12.8% [24]. While most languages have functional speech recognizers, Bangla still lacks a fully developed one. This work aims to develop a Bangla speech recognizer, focusing specifically on the automatic recognition of Bangla real numbers. Its performance is analyzed using the CMU Sphinx-4 API and the Bangla Unicode-based writing software, Avro [25].

A Bangla sentence recognition system using HMM was developed, involving feature extraction with MFCC and a recognition phase for sentence identification. Separate HMMs were trained using labeled data, and classification was performed by selecting the best match during training and testing [26]. Spectral analysis of Bangla vowels, based on vocal tract properties, is crucial for speech synthesis and recognition. The experiment classified vowels into distinct regions and illustrated vowel space in both acoustic and articulatory dimensions. Spectral characteristics and formant frequencies were measured from isolated Bangla words spoken by male and female speakers and tested in utterance synthesis [27]. The database includes training data comprising 3,824 Bangla digit utterances (25 male and 25 female speakers) and a test set of 1,985 utterances (26 male and 26 female speakers), divided into clean1, clean2, clean3, and clean4 groups. A Mel-LPC-based front-end, known for its auditory-like frequency resolution and computational efficiency, was used. Cepstral coefficients were obtained via a generalized autocorrelation function, avoiding bilinear transformation and employing a first-order all-pass filter for frequency warping. Recognition accuracy for the test sets was 98.11%, 98.05%, 97.94%, and 97.63%, respectively [28]. Mixed transform models for feature extraction were proposed, converting 1-D isolated words into 2-D forms. The first model applied 2-D FFT, Radon transform, 1-D IFFT, and 1-D discrete wavelet transforms, while the second used discrete multi circular let transforms. Recognition tasks employed dynamic time warping, achieving accuracies of 91% and 89% with wavelet transforms (Db1 and Db4) and 87%-93% with multi circular let transforms for 9 sub-bands [29]. Low-resource languages like Swahili lack adequate speech datasets for spoken digit recognition. This study developed a Swahili spoken digit dataset and explored cross-lingual, multilingual, and language-independent pre-training methods. The proposed approach integrates target language data during pre-training, optimizing recognition even with limited training

data [30]. Some scholars in 2021 concurred that the study of Bangla alphabets and speech recognition remains limited. They identified the lack of available data as a primary challenge. Their proposed system incorporates several auditory characteristics of the processed data. In this experiment, 39 alphabets were used for classification. Support Vector Machines (SVM) and Multilayer Perceptron Classifiers (MLPC) were employed to enhance classification precision. Approximately 4,095 data points were used, with MLPC and SVM achieving accuracy rates of 99.27% and 92.33%, respectively [31].

In another study, the Bangla vowels অ, আ, and ই were analyzed using Linear Predictive Coding (LPC) techniques and cepstrum-based formant estimation. Both LPC and cepstrum analysis showed recognition accuracy ranging from 71% to 75%. Data from both male and female speakers were considered in the experiment [32]. A CSV file containing the values of each feature extracted from raw audio recordings was used as the processed data. Various frequency and time domain attributes of an MP3 file were examined to create a unique dataset, which was saved in CSV format. Classifier models were trained and tested using this file. SVM and MLPC (a subclass of artificial neural networks) were employed in the experiment for classification. The SVM classifier outperformed the decision tree, particularly when the dataset was balanced. However, it performed worse than the neural network classifier, with deep learning techniques outperforming all others by a significant margin [33]. Lancaster University in the United Kingdom established the EMILLE Corpus. Large written corpora in Bengali, Gujarati, Hindi, Punjabi, Sinhala, Tamil, and Urdu were produced by EMILLE. According to the Baker and McEnery survey, these South Asian languages were the most requested by the LE community. As the project progressed, other languages were added to the collection. Additionally, EMILLE created spoken corpora for Bengali, Gujarati, Hindi, Punjabi, and Urdu—languages with a significant enough UK presence to justify the development of spoken corpus collections [34].

5. THE SCOPE OF THIS RESEARCH

The scope of this research is to investigate various features in Bangla Speech Signals and methods to recognize them. The target is to find a feature extraction method and a recognition method, which are suitable for Bangla Speech recognition. The developed system has been described for Bangla phoneme, isolated word, command and sentence's feature extraction and recognition. The intention is to provide a speech recognition system that could recognize Bangla speech with a high percentage. Necessary code in written and run in MATLAB followed by the necessary algorithms. Recorded necessary Bangla speech sample (primary dataset) for analysis within the developed system. This study covers discussions about the scope of the research, Bangla phoneme word, command and sentence sample, short-time energy calculation and silence removal procedure, Hamming, Hanning and Blackman Window framing, pre-processing, feature extraction in FFT, LPC and MFCC, forming of training and target data, FFNN, TDNN+LMA algorithm, experiment results, statistical test report (confidence intervals), concluding remarks and future scope.

6. THE STUDY'S NOVEL CONTRIBUTIONS

- A comparative analysis of various feature extraction methods and speech recognition/deep learning tools.
- A comprehensive experiment encompassing Bangla phonemes, isolated words, commands, and sentences in a single framework.
- In addition to feature extraction and deep learning tools, emphasis is also placed on the variability of frame windowing techniques, including Hamming, Hanning, and Blackman.
- Addressing the scarcity of Bangla datasets for experimentation; this dataset is made available for research purposes.
- Comparison of results with prior work, such as the READ system's 98.35% vowel accuracy [20] versus the phoneme accuracy achieved in this study.
- A discussion of recent and past research on Bangla speech recognition [19-34], aiming to overcome the limitations of previous studies.

7. DIFFICULTIES IN SPEECH RECOGNITION

Speech recognition, while incredibly powerful and useful, is a complex field that presents numerous challenges. Here are some of the key complexities involved:

7.1 Acoustic variability

- Speaker Variability: Differences in accent, dialect, gender, age, and speaking style can significantly affect the performance of speech recognition systems.
- Background Noise: Ambient noise, overlapping speech, and environmental sounds can interfere with the clarity of the spoken input.
- Microphone Quality: The quality and type of microphone used for recording can impact the accuracy of speech recognition.

7.2 Linguistic challenges

- Homophones: Words that sound the same but have different meanings and spellings (e.g., "to," "two," and "too") can be challenging to distinguish.
- Context and Ambiguity: Understanding the context in which words are used is crucial for accurate recognition, especially for words that have multiple meanings.
- Speech Disfluencies: Natural speech often includes pauses, fillers (like "um" and "uh"), and corrections, which can complicate recognition.

7.3 Technical issues

- Real-Time Processing: Processing speech in real time requires significant computational resources and efficient algorithms.
- Data Scarcity: High-quality, annotated speech data is essential for training accurate models, but such data can be scarce, especially for less widely spoken languages.
- Model Complexity: Building and training models that can accurately recognize and interpret speech involves complex machine learning techniques and large datasets.

7.4 Ethical and social considerations

- Privacy: Ensuring that speech data is collected and used in a way that respects user privacy is a critical concern.
- Bias: Speech recognition systems can exhibit biases based on the data they are trained on, which can lead to unequal performance across different demo-graphic groups.
- Accessibility: Making speech recognition technology accessible to people with speech impairments or non-standard speech patterns is an ongoing challenge.

8. SPECIFIC GAPS IN BANGLA ASR RESEARCH AND OBJECTIVES OF THIS STUDY

There are some specific gaps in Bangla ASR research noticed and how this study addresses them:

- Limited Standard Datasets: Bangla ASR research lacks widely available, well-annotated standard datasets (Corpus), forcing researchers to rely on their own primary datasets.
- Speaker Variability Challenges: Bangla has diverse accents, dialects, and pronunciation variations that pose difficulties in creating a highly generalized ASR model.
- Contextual Modeling Limitations: Existing research mainly focuses on Bangla phoneme and isolated word recognition, while Bangla command and sentence-level recognition are underexplored.

8.1 Objectives and the study contribution

- It aims to develop or enhance a structured dataset for Bangla speech recognition, ensuring greater accessibility and usability for future research. Own dataset developed (data samples more than 1500 of Bangla phoneme, word, command and sentence) and successfully utilized in the experiment.
- It incorporates data (primary dataset/ corpus) from multiple speakers across different regions (in Bangladesh) to improve model adaptability to linguistic diversity (male and female from different age group).
- It integrates contextual learning mechanisms, such as language models and deep learning approaches, to enhance sentence-level recognition. Utilized number of feature extraction methods (FFT, LPC, MFCC) and bangla recognition models (FFNN, TDNN) in one place for the experiment to see the difference, which is significant contribution to the Bangla ASR research.
- This investigation motivates future researchers to conduct more research in Bangla speech recognition.

By addressing these challenges, this study aims to be contributes making Bangla ASR more robust, scalable, and applicable in real-world scenarios.

9. METHOD OF THE EXPERIMENT

The study focuses on the speech signals of males and females (of various ages) for specific Bangla phoneme, isolated words, commands, and sentences. The method uses a variety of windowing techniques (Hamming, Hanning, and Blackman Windows), feature extraction approaches, and speech recognition tools to assess the system's accuracy in recognizing Bangla speech for both male and female voices. A basic dataset of roughly 1500 utterances (speech samples) was gathered from various age groups (Table 1).

Table 1. Bangla recorded audio sample

Category	Bangla	Properties	Duration (second)
Phoneme	অ (/O/)	(Short) Vowel, Oral, Compact, Grave	1.018 - 1.201
	আ (/A/)	(Long) Vowel, Oral, Compact	1.018 - 1.201
	ই (/I/)	(Short) Vowel, Oral, Diffuse, Acute	1.018 - 1.201
	উ(/OO/)	(Short) Vowel, Oral, Diffuse, Grave	1.018 - 1.201
	এ(/EA/)	(Complex) Vowel, Oral, Diffuse, Acute	1.018 - 1.201
	ও(/O/)	(Complex) Vowel, Oral, Diffuse, Grave	1.018 - 1.201
	ঐ (/OI/)	(Complex) Vowel, Oral, Diffuse, Grave	1.018 - 1.201
	ক (/KO/)	Consonant, Oral, Compact, Unvoiced, Grave, Lax	1.018 - 1.201
Category	**Bangla**	**English Meaning**	**Duration (second)**
Isolated Word	অংক	Math	1.201
	আমি	I	1.201
	ইলিশ	Ilish (Fish)	1.201
	উট	Camel	1.201
	কলা	Banana	1.201
	খরেগাশ	Rabbit	1.201
	গরু	Cow	1.201
	ঘড়ি	Clock	1.201
Command	এই কাজ কর	Do the job	1.802 - 2.716
	দরজা খোলো	Open the door	1.802 - 2.716
	টেবিল পরিস্কার কর	Clean the table	1.802 - 2.716
	বাম দিক যাও	Move toward the left	1.802 - 2.716
	পশ্চিম দিক সরো	Move toward the west	1.802 - 2.716
	অফিস যাও	Go to the office	1.802 - 2.716
	এই চেয়ার আনো	Bring this chair	1.802 - 2.716
	জানালা বন্ধ কর	Close the window	1.802 - 2.716
Sentence	আমরা কলা খাই	We eat bananas	2.011 - 3.213
	কলা ভালো ফল	Banana is a good fruit	2.011 - 3.213
	ফল স্বাস্থ্যের জন্য ভালো	Fruit is good for health	2.011 - 3.213
	তিন বন্ধ খেলা করে	They are three friends	2.011 - 3.213
	তারা তিন বন্ধু	Three friends play	2.011 - 3.213
	তিন বন্ধু খায়	Three friends eat	2.011 - 3.213

9.1 Short-time energy calculation and silence removal

All speech signals were segmented into 16-millisecond rectangular window frames. To enhance the processing of the sound signal, short-time energy (STE) calculation [35-37] was employed to eliminate the silence regions of the speech signal, which contain less energy. Additionally, the energy calculation of the sound signal was performed. Energy normalization was applied to each frame, discarding those with energy levels less than 2% of the maximum energy (Figure 1).

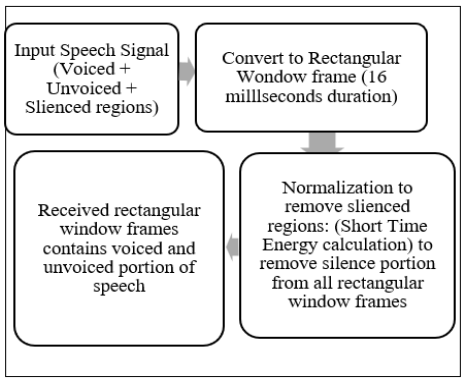

Figure 1. Short-time energy calculation and silence removal

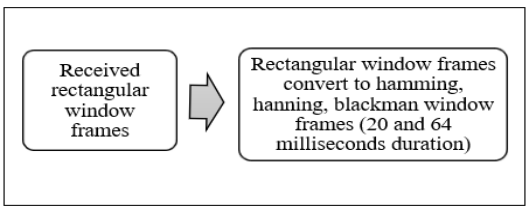

Figure 2. Hamming, Hanning, Blackman Window frame

The rectangular window is the simplest window defined by the Eq. (1):

$$w[n] = \sin(\pi n / N) = \cos(\pi n / N - \pi / 2),.....(0 \le n \le N) \quad (1)$$

The corresponding $w0(n)$ function is a cosine without $\pi/2$ phase offset [36, 37].

9.2 Hamming Window framing

The Hamming Window [38] is defined by the following Eq. (2):

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

The window length $L=N+1$.

Let $\boldsymbol{L}$ represent the window length as a positive integer, and $\boldsymbol{W}$ be the Hamming Window column vector. The Hamming Window, with a length matching the size of the frame, was applied. The speech signal was then analyzed to extract a set of features representing the spectral envelope.

9.3 Preprocessing

Pre-emphasis is applied to compensate for the negative spectral slope of the voiced portions of the speech signal.

A typical signal pre-emphasis is defined by Eq. (3) [39]:

$$y(n) = s(n) - Cxs(n-1) \quad (3)$$

where, the constant C generally falls between 0.9 and 1.0.

The pre-emphasis was performed by using an all-zero filter [39]. Three different pre-processing approaches were used:

- Pre-processing = Hamming Window + Pre-emphasis
- Pre-processing = Hanning Window + Pre-emphasis
- Pre-processing = Blackman Window + Pre-emphasis

Preprocessing of each frame has been performed, with the frame variable containing all frames generated by the framing function (Figure 2). Although zero padding can help to resolve the finer structure of the spectrum, it did not improve the resolution in this experiment. Therefore, zero-padding and frame overlapping techniques have been avoided during segmenting the entire speech signal into multiple frames.

The choice of window length in crucial especially in speech processing which essential due to the time-variant nature of speech signals. Windowing segments the signal into short frames where characteristics remain stable, aiding accurate feature extraction. Shorter windows (5-25 ms) capture rapid phoneme changes but may cause spectral distortion. Longer windows (25–64 ms) improve frequency resolution but can blur transient speech features, making phoneme distinction harder. That is the reason two types of window lengths were used in the experiment [40].

10. EXPERIMENTS AND RESULTS

The code is written in MATLAB. After extracting the feature of speech data samples split into 70% for training, 15% for validation, and 15% for testing purpose. The training data is presented to the network during training, and the network adjusts based on its error. During validation, the data is used to assess network generalization and to stop training when generalization ceases to improve. In testing, the data does not influence training and instead provides an independent measure of network performance during and after training.

Experiments has been conducted using Bangla phonemes, isolated words, commands, and sentences. These tests included dataset collected from a diverse group of speakers, encompassing both male and female participants across different age groups. The results of each experiment are presented in detail. For feature extraction in parallel FFT, LPC and MFCC methods are applied. For framing two different lengths of the window frame (20 & 64 milliseconds length of Hamming, Hanning and Blackman Window) were separately used for all experiments.

10.1 Experiment using two-layer FFNN and TDNN

Two-layer FFNN and TDNN with Levenberg Marquardt Algorithm (LMA) have been applied. The observation showcased that TDNN is slightly better than FFNN.

Table 2 about Bangla Phoneme Recognition in FFNN and TDNN (with LMA). MFCC + TDNN combination provided slightly better recognition accuracy compared to MFCC + FFNN. The TDNN likely handled temporal dependencies in speech more effectively.

It has been noticed that the recognition accuracy especially for Bangla phonemes showcased higher (even 100%) when the dataset was restricted to either male or female speakers, as opposed to a mixed-gender dataset. This suggests that gender-specific models may offer a performance advantage by

minimizing intra-class variability caused by differences in vocal tract characteristics, pitch, and articulation patterns. The observed disparity underscores the potential benefits of gender-dependent training strategies in speech recognition systems, particularly for languages with distinct phonetic variations across genders.

In Table 2, considering Blackman, Hamming, and Hanning Window types with window lengths of 20 and 64 milliseconds, the average recognition accuracy for TDNN+MFCC reaches an impressive 97.5%, while FFNN+MFCC achieves 96.5%. This suggests that TDNN slightly outperforms FFNN for single male speakers, although the difference is minimal. However, in mixed-gender scenarios, TDNN+MFCC records an accuracy of 86.5%, compared to FFNN+MFCC's 92.16%, indicating a slight underperformance. This variation highlights the impact of factors such as gender diversity, increased sample size, age differences, and speech accents on recognition accuracy. Figure 3 shows the architecture of TDNN using MFCC.

Table 2. Bangla phoneme recognition in FFNN and TDNN (with LMA)

Speaker (No. of Phoneme: 08)	Feature Extraction Methods	Window Length (in milliseconds)	Recognition Percentage (FFNN)	Recognition Percentage (TDNN, LMA)
Single male speaker/No. of utterances recognize out of 40	FFT	20 Ms. (HM)	95%	92%
		20 Ms. (HN)	95%	95%
		20 Ms. (BL)	83%	85%
		64 Ms. (HM)	85%	75%
		64 Ms. (HN)	82%	83%
		64 Ms. (BL)	82%	85%
	LPC	20 Ms. (HM)	83%	85%
		20 Ms. (HN)	77%	83%
		20 Ms. (BL)	77%	80%
		64 Ms. (HM)	90%	65%
		64 Ms. (HN)	88%	90%
		64 Ms. (BL)	88%	90%
	MFCC	20 Ms. (HM)	**100%**	**100%**
		20 Ms. (HN)	**100%**	**100%**
		20 Ms. (BL)	**100%**	**100%**
		64 Ms. (HM)	95%	100%
		64 Ms. (HN)	92%	90%
		64 Ms. (BL)	92%	95%
Single female speaker/No. of Utterances recognize out of 40	FFT	20 Ms. (HM)	72%	72%
		20 Ms. (HN)	75%	75%
		20 Ms. (BL)	77%	75%
		64 Ms. (HM)	92%	62%
		64 Ms. (HN)	90%	90%
		64 Ms. (BL)	90%	90%
	LPC	20 Ms. (HM)	95%	90%
		20 Ms. (HN)	87%	90%
		20 Ms. (BL)	77%	77%
		64 Ms. (HM)	72%	87%
		64 Ms. (HN)	72%	72%
		64 Ms. (BL)	75%	72%
	MFCC	**20 Ms.** (HM)	**100%**	**100%**
		20 Ms. (HN)	**100%**	97%
		20 Ms. (BL)	**100%**	97%
		64 Ms. (HM)	97%	97%
		64 Ms. (HN)	97%	97%
		64 Ms. (BL)	97%	97%
12 male-female speakers/No. of utterances recognize out of 480	FFT	20 Ms. (HM)	60%	66%
		20 Ms. (HN)	60%	66%
		20 Ms. (BL)	60%	65%
		64 Ms. (HM)	54%	59%
		64 Ms. (HN)	54%	59%
		64 Ms. (BL)	54%	59%
	LPC	20 Ms. (HM)	54%	61%
		20 Ms. (HN)	54%	61%
		20 Ms. (BL)	54%	61%
		64 Ms. (HM)	55%	70%
		64 Ms. (HN)	54%	55%
		64 Ms. (BL)	54%	55%
	MFCC	20 Ms. (HM)	**99%**	80%
		20 Ms. (HN)	99%	80%
		20 Ms. (BL)	97%	97%
		64 Ms. (HM)	86%	88%
		64 Ms. (HN)	86%	88%
		64 Ms. (BL)	86%	86%

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 3. Bangla word recognition in FFNN and TDNN (with LMA)

Speaker (No. of Word: 08)	Feature Extraction Methods	Window Length (in milliseconds)	Recognition Percentage (FFNN)	Recognition Percentage (TDNN, LMA)
Single male speaker/No. of utterances recognize out of 40	FFT	20 Ms. (HM)	65%	65%
		20 Ms. (HN)	67%	67%
		20 Ms. (BL)	65%	65%
		64 Ms. (HM)	70%	60%
		64 Ms. (HN)	70%	65%
		64 Ms. (BL)	67%	65%
	LPC	20 Ms. (HM)	87%	80%
		20 Ms. (HN)	87%	87%
		20 Ms. (BL)	87%	87%
		64 Ms. (HM)	55%	45%
		64 Ms. (HN)	57%	47%
		64 Ms. (BL)	57%	45%
	MFCC	20 Ms. (HM)	**100%**	97%
		20 Ms. (HN)	97%	97%
		20 Ms. (BL)	97%	**100%**
		64 Ms. (HM)	92%	92%
		64 Ms. (HN)	90%	90%
		64 Ms. (BL)	90%	90%
Single female speaker/No. of utterances recognize out of 40	FFT	20 Ms. (HM)	75%	70%
		20 Ms. (HN)	70%	75%
		20 Ms. (BL)	72%	72%
		64 Ms. (HM)	72%	57%
		64 Ms. (HN)	75%	75%
		64 Ms. (BL)	72%	75%
	LPC	20 Ms. (HM)	87%	82%
		20 Ms. (HN)	87%	82%
		20 Ms. (BL)	87%	87%
		64 Ms. (HM)	87%	82%
		64 Ms. (HN)	87%	87%
		64 Ms. (BL)	82%	87%
	MFCC	20 Ms. (HM)	**100%**	97%
		20 Ms. (HN)	97%	**100%**
		20 Ms. (BL)	97%	97%
		64 Ms. (HM)	87%	95%
		64 Ms. (HN)	87%	95%
		64 Ms. (BL)	87%	95%
10 male-female speakers/No. of utterances recognize out of 400	FFT	20 Ms. (HM)	50%	51%
		20 Ms. (HN)	50%	51%
		20 Ms. (BL)	50%	50%
		64 Ms. (HM)	60%	44%
		64 Ms. (HN)	60%	60%
		64 Ms. (BL)	60%	60%
	LPC	20 Ms. (HM)	43%	53%
		20 Ms. (HN)	43%	53%
		20 Ms. (BL)	43%	53%
		64 Ms. (HM)	47%	47%
		64 Ms. (HN)	43%	53%
		64 Ms. (BL)	43%	53%
	MFCC	20 Ms. (HM)	93%	75%
		20 Ms. (HN)	93%	93%
		20 Ms. (BL)	75%	93%
		64 Ms. (HM)	**94%**	75%
		64 Ms. (HN)	93%	93%
		64 Ms. (BL)	93%	93%

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 4. Bangla command recognition in FFNN & TDNN (with LMA)

Speaker (No. of Command: 08)	Feature Extraction Methods	Window Length (in milliseconds)	Recognition Percentage (FFNN, LMA)	Recognition Percentage (TDNN, LMA)
Single male speaker/No. of utterances recognize out of 40	FFT	20 Ms. (HM)	40%	47%
		20 Ms. (HN)	40%	47%
		20 Ms. (BL)	40%	47%
		64 Ms. (HM)	70%	70%
		64 Ms. (HN)	70%	70%
		64 Ms. (BL)	70%	70%

Single female speaker/No. of utterances recognize out of 40	LPC	20 Ms. (HM)	70%	75%
		20 Ms. (HN)	70%	75%
		20 Ms. (BL)	70%	70%
		64 Ms. (HM)	65%	72%
		64 Ms. (HN)	72%	65%
		64 Ms. (BL)	65%	72%
	MFCC	20 Ms. (HM)	**100%**	**97%**
		20 Ms. (HN)	97%	97%
		20 Ms. (BL)	97%	97%
		64 Ms. (HM)	**100%**	100%
		64 Ms. (HN)	97%	97%
		64 Ms. (BL)	97%	97%
	FFT	20 Ms. (HM)	32%	47%
		20 Ms. (HN)	30%	30%
		20 Ms. (BL)	30%	30%
		64 Ms. (HM)	25%	37%
		64 Ms. (HN)	25%	37%
		64 Ms. (BL)	25%	35%
	LPC	20 Ms. (HM)	35%	65%
		20 Ms. (HN)	35%	37%
		20 Ms. (BL)	35%	37%
		64 Ms. (HM)	50%	37%
		64 Ms. (HN)	50%	52%
		64 Ms. (BL)	50%	52%
	MFCC	20 Ms. (HM)	**90%**	85%
		20 Ms. (HN)	87%	92%
		20 Ms. (BL)	87%	92%
		64 Ms. (HM)	87%	92%
		64 Ms. (HN)	87%	87%
		64 Ms. (BL)	87%	87%
	FFT	20 Ms. (HM)	16%	22%
		20 Ms. (HN)	16%	22%
		20 Ms. (BL)	16%	16%
		64 Ms. (HM)	32%	25%
		64 Ms. (HN)	32%	32%
		64 Ms. (BL)	25%	32%
10 male-female speakers/ No. of utterances recognize out of 400	LPC	20 Ms. (HM)	21%	24%
		20 Ms. (HN)	21%	24%
		20 Ms. (BL)	21%	21%
		64 Ms. (HM)	32%	45%
		64 Ms. (HN)	21%	21%
		64 Ms. (BL)	21%	21%
	MFCC	20 Ms. (HM)	**83%**	57%
		20 Ms. (HN)	80%	80%
		20 Ms. (BL)	80%	80%
		64 Ms. (HM)	72%	60%
		64 Ms. (HN)	72%	72%
		64 Ms. (BL)	72%	72%

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 5. Bangla sentence recognition in FFNN and TDNN (with LMA)

Speaker (No. of Sentence: 06)	Feature Extraction Methods	Window Length (in milliseconds)	Recognition Percentage (FFNN)	Recognition Percentage (TDNN, LMA)
Single male speaker/No. of utterances recognize out of 30	FFT	20 Ms. (HM)	67%	73%
		20 Ms. (HN)	67%	73%
		20 Ms. (BL)	67%	73%
		64 Ms. (HM)	70%	70%
		64 Ms. (HN)	70%	70%
		64 Ms. (BL)	70%	70%
	LPC	20 Ms. (HM)	53%	76%
		20 Ms. (HN)	53%	65%
		20 Ms. (BL)	53%	65%
		64 Ms. (HM)	50%	73%
		64 Ms. (HN)	50%	53%
		64 Ms. (BL)	50%	53%
	MFCC	20 Ms. (HM)	**93%**	**90%**
		20 Ms. (HN)	90%	90%
		20 Ms. (BL)	99%	90%
		64 Ms. (HM)	90%	96%
		64 Ms. (HN)	90%	96%

Single female speaker/No. of utterances recognize out of 30	FFT	64 Ms. (BL)	90%	97%
		20 Ms. (HM)	50%	53%
		20 Ms. (HN)	50%	53%
		20 Ms. (BL)	50%	53%
		64 Ms. (HM)	63%	50%
		64 Ms. (HN)	63%	50%
		64 Ms. (BL)	63%	50%
	LPC	20 Ms. (HM)	43%	57%
		20 Ms. (HN)	50%	50%
		20 Ms. (BL)	50%	50%
		64 Ms. (HM)	47%	57%
		64 Ms. (HN)	50%	43%
		64 Ms. (BL)	50%	57%
	MFCC	20 Ms. (HM)	**80%**	80%
		20 Ms. (HN)	80%	80%
		20 Ms. (BL)	80%	80%
		64 Ms. (HM)	77%	**93%**
		64 Ms. (HN)	80%	80%
		64 Ms. (BL)	77%	80%
10 male-female speakers/No. of utterances recognize out of 300	FFT	20 Ms. (HM)	43%	43%
		20 Ms. (HN)	43%	57%
		20 Ms. (BL)	50%	50%
		64 Ms. (HM)	45%	47%
		64 Ms. (HN)	45%	47%
		64 Ms. (BL)	45%	47%
	LPC	20 Ms. (HM)	35%	49%
		20 Ms. (HN)	35%	49%
		20 Ms. (BL)	35%	49%
		64 Ms. (HM)	31%	44%
		64 Ms. (HN)	35%	49%
		64 Ms. (BL)	35%	49%
	MFCC	20 Ms. (HM)	**89%**	77%
		20 Ms. (HN)	90%	**93%**
		20 Ms. (BL)	90%	90%
		64 Ms. (HM)	54%	65%
		64 Ms. (HN)	54%	65%
		64 Ms. (BL)	54%	65%

Note: Hamming=HM, Hanning=HN, Blackman=BL

READ is a Bangla phoneme recognition system that was used for a Bangla phoneme recognition experiment. In this study, Bangla vowels with a West Bengal accent (which differs from the Bangladesh accent) were specifically considered [20]. The system utilized the MFCC method for feature extraction and has been reported to achieve good accuracy in Bangla phoneme recognition, based on a dataset of approximately 1,400 Bangla vowels. The present experiment (Tables 2-5) extends beyond phoneme recognition to include Bangla words, commands, and sentences. A variety of feature extraction methods, window framing techniques, window sizes, and deep learning-based recognition tools have been employed. As a result, this experiment has been conducted on a much larger scale compared to the READ experiment. When comparing the phoneme recognition aspect, a particularly Bangla vowel recognition - READ achieved a maximum accuracy of 98.35%, whereas within this experiment (Table 2), utilizing TDNN+MFCC and FFNN+MFCC, achieved up to 100% accuracy. Therefore, READ has underperformed in comparison. However, both experiments still face certain challenges, including speaker variability, availability of well format dataset, quantity of data sample etcetera.

From the experiment results (Table 3), Isolated Bangla word recognition (male-female speaker) using FFT, LPC, MFCC with FFNN provides slightly better results, 85% (average) accuracy than TDNN which was 81% (average). Here, the MFCC method for pattern recognition performs better than FFT and LPC.

From the Table 4, Bangla command recognition (male-female speaker) using FFT, LPC and MFCC with TDNN provides better results, 60% (average) accuracy than FFNN which is 57% (average). The MFCC method for pattern recognition performs better than FFT and LPC.

From the Table 5, Bangla sentence recognition (male and female speaker) using FFT, LPC and MFCC with TDNN provides better results, 66.45% (average) accuracy than FFNN which is 60.49% (average). The experiment has demonstrated that the MFCC method outperforms both Fast Fourier Transform (FFT) and LPC for pattern recognition in Bangla speech.

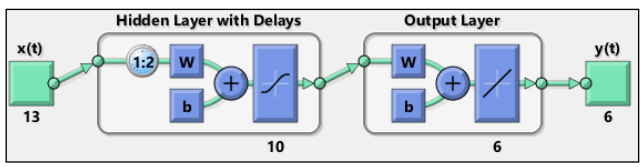

Figure 3. TDNN using MFCC

11. SYSTEM'S PERFORMANCE EVALUATION

The developed system's performance evaluation (Tables 6-9) for Bangla speech recognition (phoneme, word, command, and sentence) has been thoroughly analyzed based on experiments after feature extraction. During deep learning processes, various metrics were considered, including Best

Validation Performance, Validation Checks, Error Histogram, Regression Analysis, Time-Series Response, Error Autocorrelation, and Input-Error Cross-correlation. For Bangla phoneme, word, command, and sentence recognition, feature extraction techniques such as FFT, LPC, and MFCC were employed. Recognition was implemented using FFNN and TDNN, utilizing up to 480 samples uttered by a maximum of 12 male and female speakers. A window length of 20 and 64 MS was used for framing (Hamming, Hanning, Blackman Window). Different phonemes, words, commands, and sentences were tested to evaluate the system's performance comprehensively.

In the experiments, the necessary code written and executed in MATLAB. There are various evaluation metrics available in MATLAB that help to assess the model's robustness and generalizability. These evaluation metrics also ensure the developed system model it truly potential.

- **Best Validation Performance** tracks validation error during training, identifies the lowest validation error to prevent overfitting and stops training at the optimal point for better generalization.
- **Error Histogram** visualizes the distribution of prediction errors, helps detect **biases** in error patterns, and ideally shows errors centered around zero for **balanced generalization**.
- **Validation Checks** stop training when validation error increases to prevent overfitting, confirm model performance on validation data before final testing, and ensure adaptation to new datasets without losing accuracy.
- **Regression Analysis (R)** measures the correlation between predicted and actual values, with R values close to 1 indicating strong predictive ability and ensuring the model generalizes well across different datasets.
- **Time-Series Response** assesses the model's reaction to sequential data, ensures adaptability to trends and fluctuations, and validates accuracy in forecasting tasks like speech or financial predictions.
- **Error Autocorrelation** analyzes error relationships over time, ensuring they remain uncorrelated to avoid systematic bias and enhance robustness against repetitive errors.
- **Input-Error Cross-Correlation** analyzes whether input variables excessively influence errors, where high correlation indicates bias, while minimal correlation ensures fair and generalizable predictions across diverse input conditions.

Table 6. Bangla phoneme recognition in TDNN

08 Different Phonemes (uttered 480 times)	Feature Extraction Methods	Window Length (in milliseconds)	*Performance Evaluation with Epoch=E	*Training State (Gradient, Epoch=E)	*Error Histogram (Bins=B)	**Regression Analysis (R)	**Time-Series Response (Error) (R)	*Error Autocorrelation (Correlation)	*Input-Error Cross-Correlation (Error)
12 male-female speakers	FFT	20 Ms. (HM)	0.0702, E70	0.0002, E76	0.0366, B20	0.6025	-0.3534	0.0795	-0.0025
		20 Ms. (HN)	0.0702, E81	0.0002, E59	0.0365, B20	0.6026	-0.3537	0.0796	-0.0025
		20 Ms. (BL)	0.0702, E68	0.0002, E81	0.0367, B20	0.6024	-0.3529	0.0799	-0.0025
		64 Ms. (HM)	0.0737, E60	0.0003, E66	0.109, B20	0.6400	-0.4065	0.0725	-0.0009
		64 Ms. (HN)	0.0747, E44	0.0003, E68	0.101, B20	0.6500	-0.4062	0.0726	-0.0009
		64 Ms. (BL)	0.0747, E71	0.0003, E87	0.111, B20	0.6300	-0.4069	0.0725	-0.0009
	LPC	20 Ms. (HM)	0.0716, E127	0.0013, E133	0.0661, B20	0.5966	-0.3873	0.0652	-0.0072
		20 Ms. (HN)	0.0715, E171	0.0012, E109	0.0671, B20	0.5969	-0.3870	0.0653	-0.0072
		20 Ms. (BL)	0.0617, E111	0.0013, E121	0.0761, B20	0.5976	-0.3971	0.0653	-0.0074
	MFCC	64 Ms. (HM)	0.0641, E75	0.0022, E81	0.0098, B20	0.6308	-0.3562	0.0518	-0.0179
		64 Ms. (HN)	0.0641, E57	0.0022, E99	0.0098, B20	0.6407	-0.3563	0.0528	-0.0179
		64 Ms. (BL)	0.0641, E79	0.0022, E77	0.0098, B20	0.6307	-0.3570	0.0517	-0.0179
		20 Ms. (HM)	0.0503, E22	0.0022, E28	0.0317, B20	0.7346	-0.4075	0.0377	-0.0381
		20 Ms. (HN)	0.0503, E31	0.0022, E19	0.0314, B20	0.7347	-0.4059	0.0376	-0.0382
		20 Ms. (BL)	0.0503, E19	0.0022, E24	0.0318, B20	0.7347	-0.4081	0.0377	-0.0380
		64 Ms. (HM)	0.0449, E20	0.0083, E26	0.0330, B20	0.7950	-0.4295	0.0178	-0.1737
		64 Ms. (HN)	0.0448, E31	0.0083, E33	0.0340, B20	0.7949	-0.4287	0.0179	-0.1741
		64 Ms. (BL)	0.0449, E22	0.0083, E24	0.0329, B20	0.8010	-0.4301	0.0181	-0.1743

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 7. Bangla word recognition in TDNN

08 Different Words (uttered 400 times)	Feature Extraction Methods	Window Length (in milliseconds)	*Performance Evaluation with Epoch=E	*Training State (Gradient, Epoch=E)	*Error Histogram (Bins=B)	**Regression Analysis (R)	**Time-Series Response (Error) (R)	*Error Autocorrelation (Correlation)	*Input-Error Cross-Correlation (Error)
10 male-female speakers	FFT	20 Ms. (HM)	0.0866, E18	0.0102, E24	0.0184, B20	0.4728	-0.3557	0.0808	-0.0056
		20 Ms. (HN)	0.0865, E22	0.0102, E33	0.0189, B20	0.4733	-0.3498	0.0809	-0.0051
		20 Ms. (BL)	0.0867, E31	0.0103, E19	0.0190, B20	0.4730	-0.3571	0.0810	-0.0049
		64 Ms. (HM)	0.0803, E30	0.0048, E36	0.0927, B20	0.5755	-0.1686	0.0530	-0.0030
		64 Ms. (HN)	0.0804, E29	0.0048, E41	0.0930, B20	0.5777	-0.1866	0.0529	-0.0034
		64 Ms. (BL)	0.0803, E38	0.0048, E28	0.0929, B20	0.5801	-0.1801	0.0550	-0.0040
	LPC	20 Ms. (HM)	0.0866, E62	0.0036, E68	0.0663, B20	0.4691	-0.0789	0.0826	-0.0059

		20 Ms. (HN)	0.0870, E87	0.0036, E86	0.0697, B20	0.4689	-0.0779	0.0830	-0.0058
		20 Ms. (BL)	0.0869, E71	0.0036, E74	0.0701, B20	0.4697	-0.0784	0.0829	-0.0060
		64 Ms. (HM)	0.0869, E31	0.0011, E37	0.0945, B20	0.4567	-0.1054	0.0934	-0.0110
		64 Ms. (HN)	0.0870, E42	0.0011, E55	0.0950, B20	0.4571	-0.1076	0.0937	-0.0120
		64 Ms. (BL)	0.0870, E39	0.0011, E49	0.0949, B20	0.4570	-0.1081	0.0940	-0.0118
		20 Ms. (HM)	0.0404, E41	0.0015, E47	0.0170, B20	0.3242	-0.176	0.0238	-0.0131
		20 Ms. (HN)	0.0405, E61	0.0015, E55	0.0180, B20	0.3249	-0.189	0.0241	-0.0155
	MFCC	20 Ms. (BL)	0.0405, E55	0.0015, E41	0.0179, B20	0.3247	-0.191	0.0240	-0.0161
		64 Ms. (HM)	0.0695, E14	0.0236, E20	0.0379, B20	0.6882	-0.5727	0.0150	-0.1259
		64 Ms. (HN)	0.0696, E21	0.0236, E33	0.0380, B20	0.6888	-0.5802	0.0159	-0.1307
		64 Ms. (BL)	0.0696, E34	0.0237, E27	0.0383, B20	0.6891	-0.5843	0.0160	-0.1345

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 8. Bangla command recognition in TDNN

08 Different Commands (uttered 400 times)	**Feature Extraction Methods**	**Window Length (in milliseconds)**	***Performance Evaluation with Epoch=E**	***Training State (Gradient, Epoch=E)**	***Error Histogram (Bins=B)**	****Regression Analysis (R)**	****Time-Series Response (Error) (R)**	***Error Autocorrelation (Correlation)**	***Input-Error Cross-Correlation (Error)**
		20 Ms. (HM)	0.1032, E99	0.0002, E105	0.0381, B20	0.2543	-0.6741	0.0959	0.0002
		20 Ms. (HN)	0.1043, E101	0.0002, E117	0.0391, B20	0.2550	-0.6801	0.0958	0.0002
	FFT	20 Ms. (BL)	0.1040, E89	0.0002, E98	0.0379, B20	0.2551	-0.6811	0.0964	0.0002
		64 Ms. (HM)	0.1009, E17	0.0124, E23	0.0691, B20	0.3010	-0.3116	0.0811	0.0005
		64 Ms. (HN)	0.1009, E22	0.0124, E31	0.0690, B20	0.3029	-0.3210	0.0818	0.0005
		64 Ms. (BL)	0.1009, E29	0.0124, E38	0.0700, B20	0.3030	-0.3302	0.0829	0.0005
		20 Ms. (HM)	0.1033, E62	0.0004, E68	0.0851, B20	0.2794	-0.2578	0.0903	-0.027
		20 Ms. (HN)	0.1051, E49	0.0004, E77	0.0861, B20	0.2799	-0.2581	0.0910	-0.028
10 Male-female speakers	LPC	20 Ms. (BL)	0.1034, E54	0.0004, E81	0.0859, B20	0.2789	-0.2531	0.0921	-0.028
		64 Ms. (HM)	**0.1052, E6**	0.0017, E12	0.0397, B20	0.2341	-0.0979	0.0919	-0.000
		64 Ms. (HN)	0.1057, E9	0.0017, E19	0.0411, B20	0.2349	-0.0985	0.0997	-0.000
		64 Ms. (BL)	0.1060, E11	0.0017, E22	0.0399, B20	0.2350	-0.0983	0.0981	-0.000
		20 Ms. (HM)	0.0938, E29	0.0040, E35	0.1072, B20	0.3919	-0.1656	0.0855	-0.153
		20 Ms. (HN)	0.0938, E21	0.0040, E48	0.1219, B20	0.3932	-0.1665	0.0859	-0.156
	MFCC	20 Ms. (BL)	0.0939, E44	0.0040, E51	0.1171, B20	0.3921	-0.1671	0.0860	-0.161
		64 Ms. (HM)	0.0940, E18	0.0025, E24	0.0022, B20	0.4632	-0.182	0.0783	-0.202
		64 Ms. (HN)	0.0941, E27	0.0025, E29	0.0023, B20	0.4710	-0.199	0.0791	-0.222
		64 Ms. (BL)	0.0942, E24	0.0025, E33	0.0023, B20	0.4555	-0.189	0.0789	-0.232

Note: Hamming=HM, Hanning=HN, Blackman=BL

Table 9. Bangla sentence recognition in Time delays neural network

06 Different Sentences (uttered 300 times)	**Feature Extraction Methods**	**Window Length (in milliseconds)**	***Performance Evaluation with Epoch=E**	***Training State (Gradient, Epoch=E)**	***Error Histogram (Bins=B)**	****Regression Analysis (R)**	****Time-Series Response (Error) (R)**	***Error Autocorrelation (Correlation)**	***Input-Error Cross-Correlation (Error)**
		20 Ms. (HM)	0.1309, E19	0.0038, E25	0.0471, B20	0.2443	-0.1867	0.107	-0.0045
		20 Ms. (HN)	0.1321, E22	0.0038, E33	0.0481, B20	0.2450	-0.1888	0.111	-0.0047
	FFT	20 Ms. (BL)	0.1311, E32	0.0038, E12	0.0481, B20	0.2452	-0.1967	0.119	-0.0046
		64 Ms. (HM)	0.1280, E11	0.0098, E17	0.0240, B20	0.2958	-0.3211	0.0887	-0.0022
		64 Ms. (HN)	0.1270, E17	0.0098, E24	0.0253, B20	0.2961	-0.3279	0.0869	-0.0023
		64 Ms. (BL)	0.1269, E19	0.0098, E31	0.0249, B20	0.2960	-0.3301	0.0878	-0.0022
		20 Ms. (HM)	0.1304, E57	0.0009, E63	0.0234, B20	0.2916	-0.2259	0.1117	0.00171
		20 Ms. (HN)	0.1310, E66	0.0009, E44	0.0243, B20	0.2924	-0.2121	0.1201	0.00179
10 Male-female speakers	LPC	20 Ms. (BL)	0.1310, E75	0.0009, E51	0.0234, B20	0.3000	-0.2212	0.1199	0.00180
		64 Ms. (HM)	0.1315, E18	0.0086, E24	0.143, B20	0.3098	-0.1951	0.0952	0.00193
		64 Ms. (HN)	0.1333, E21	0.0086, E41	0.166, B20	0.3101	-0.1999	0.0961	0.00199
		64 Ms. (BL)	0.1321, E33	0.0086, E33	0.159, B20	0.3108	-0.2001	0.0967	0.00201
		20 Ms. (HM)	0.1188, E32	0.0146, E38	0.0082, B20	0.3927	-0.1716	0.0532	-0.1216
		20 Ms. (HN)	0.1190, E44	0.0147, E52	0.0083, B20	0.3929	-0.1787	0.0555	-0.1287
	MFCC	20 Ms. (BL)	0.1193, E39	0.0148, E45	0.0083, B20	0.3930	-0.1809	0.0543	-0.1333
		64 Ms. (HM)	0.1207, E12	0.0073, E18	0.0890, B20	0.4273	-0.353	0.0735	-0.1135
		64 Ms. (HN)	0.1211, E19	0.0073, E22	0.0891, B20	0.4281	-0.360	0.0740	-0.1231
		64 Ms. (BL)	0.1221, E32	0.0073, E31	0.0899, B20	0.4283	-0.369	0.0741	-0.1210

Note: Hamming=HM, Hanning=HN, Blackman=BL

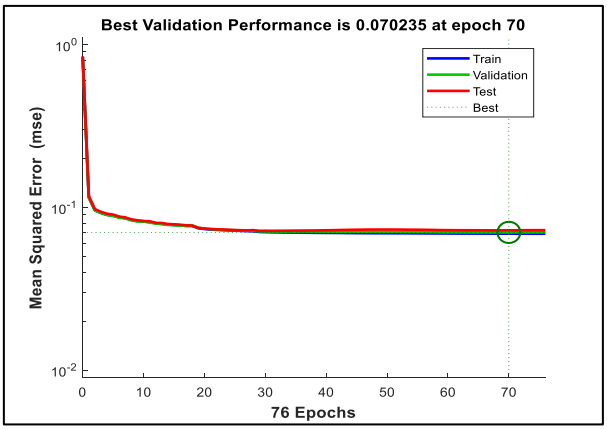

Figure 4. TDNN training (Best validation performance) FFT - Bangla phoneme

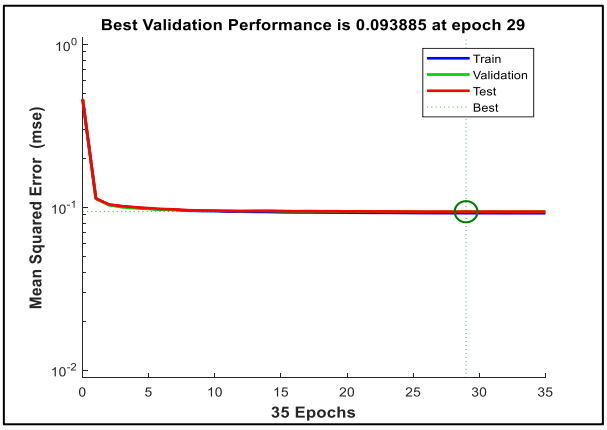

Figure 5. TDNN training (Best validation performance) MFCC - Bangla command

11.1 Best validation performance

Figures 4 and 5 represent the graphical presentation of the results during training and validation. Here Achieving Best validation performances with mean squared error rates close to zero indicates that the system is highly effective in recognizing speech accurately. The fact that this is achieved within just a few epochs speaks volumes about the efficiency and robustness of the model.

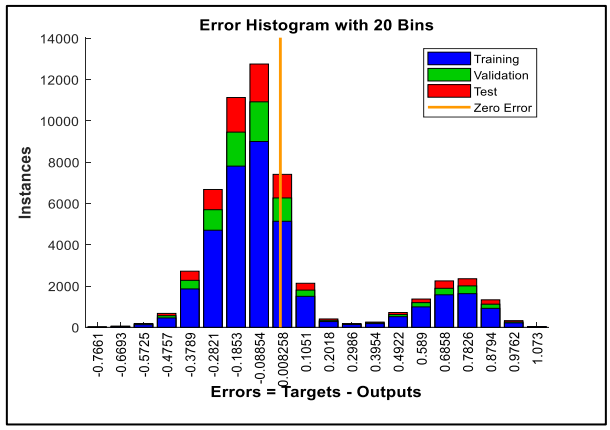

Figure 6. Neural network training (Error histogram) MFCC - Bangla sentence

11.2 Error histogram

Figure 6 shows the strong indicator of the system's potential and effectiveness. An error histogram with results close to zero for 20 bins suggests that this model has very low error rates, which is a positive sign.

11.3 Validation checks

Achieving gradient points close to zero with only a few epochs during validation checks indicates that TDNN is highly effective and well-optimized and that is observed in Figure 7.

11.4 Regression analysis

The R value, or correlation coefficient, is a critical measure of how well the model's predictions align with the actual targets. (Figure 8) An R value close to 1 signifies a strong positive correlation between the predicted outputs and the actual targets. This indicates that the model's predictions are highly accurate and closely match the true values.

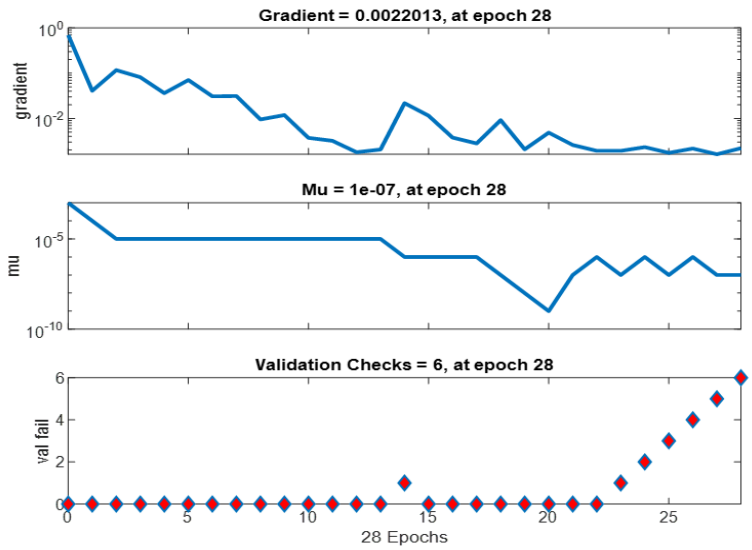

Figure 7. TDNN training- (MFCC Bangla phoneme)

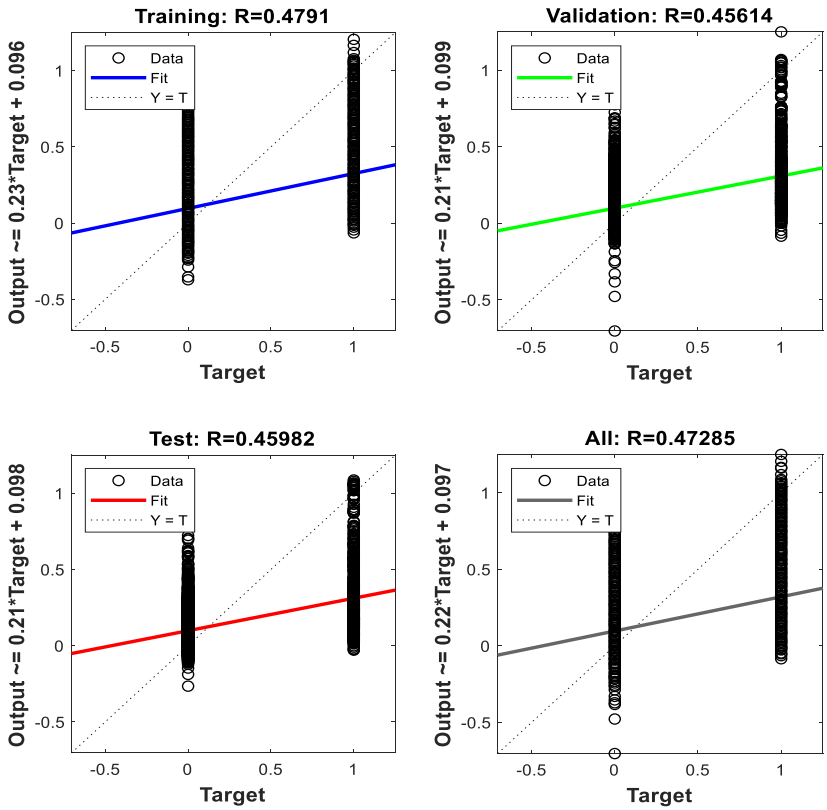

Figure 8. TDNN training (Regression analysis) FFT – Bangla word

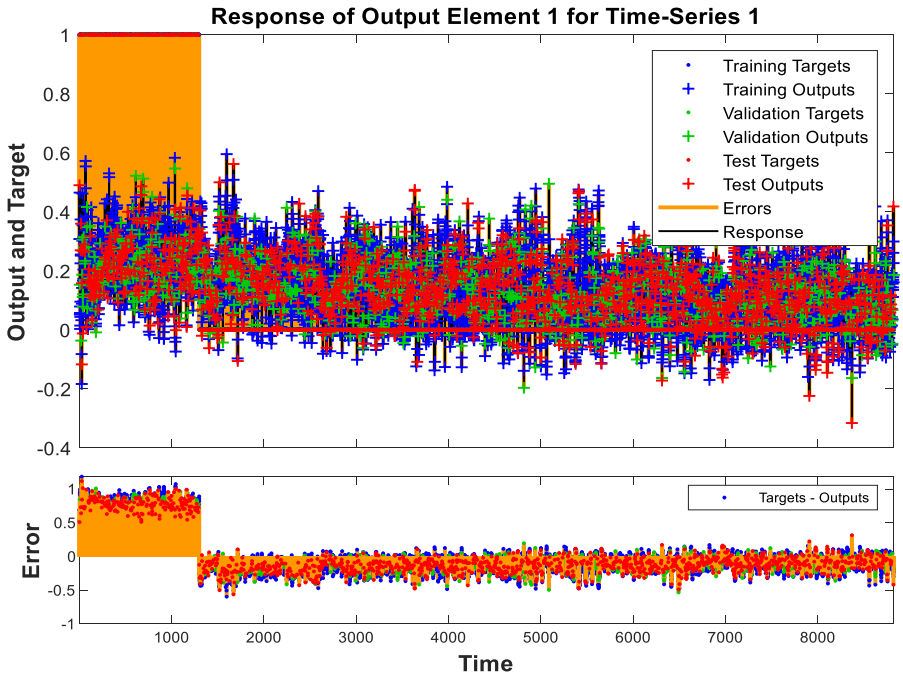

Figure 9. Neural network training (Time-series response) LPC - Bangla sentence

11.5 Time-series response

Figure 9 highlights the robustness of the TDNN system for speech recognition. The high ***R*** values from the time-series response analysis demonstrate a strong correlation between the model's outputs and the actual targets, reinforcing the effectiveness of the approach. An R value of 1 means a close relationship, 0 a random relationship.

11.6 Error autocorrelation

Error autocorrelation measures the correlation of errors in the predictions over time. Lower values are better, with zero indicating no error correlation, which means the system's errors are random and not systematic. The result is graphically presented in Figure 10.

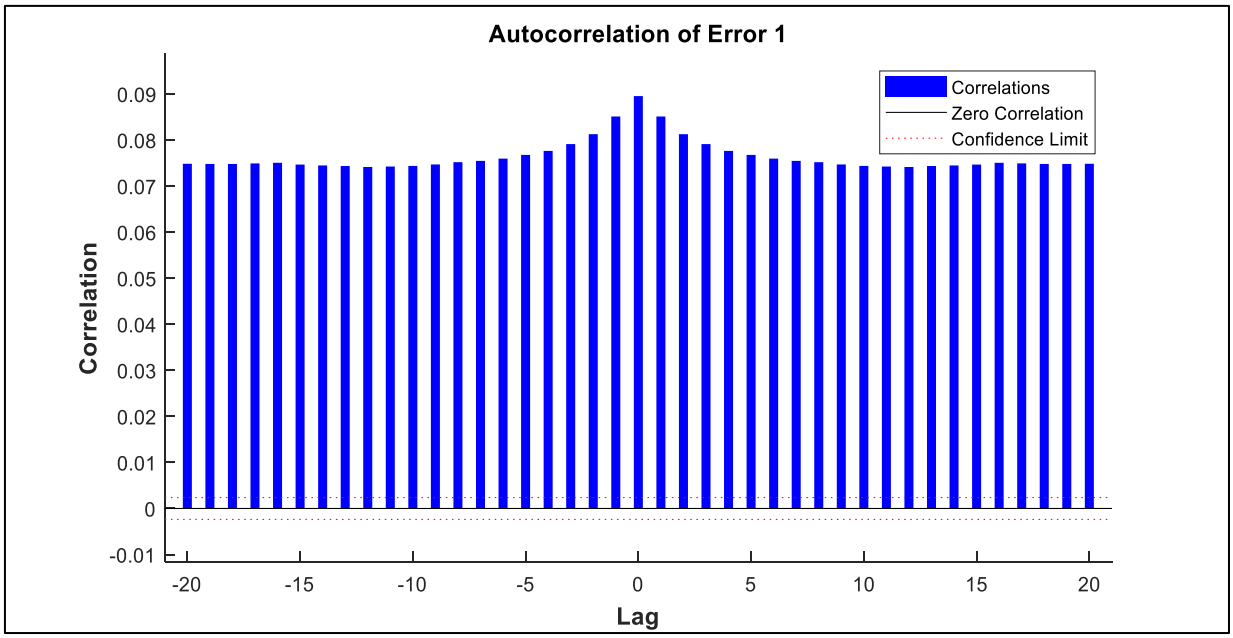

Figure 10. TDNN training for error autocorrelation (FFT – Bangla phoneme)

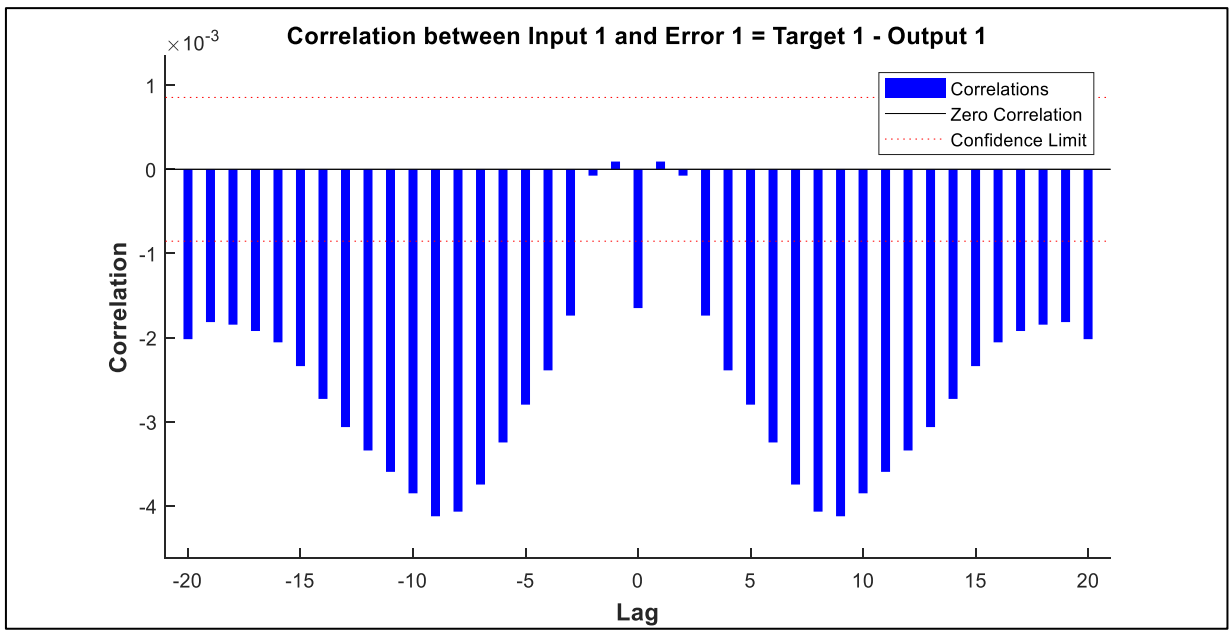

Figure 11. TDNN training (Input-error cross-correlation) FFT – Bangla word

11.7 Input-error cross-correlation

Figure 11 represents the result of Input-error cross-correlation. Here Input-error cross-correlation is a valuable metric for assessing the performance of the speech recognition system. This measures how well the errors in the system's predictions correlate with the input signal. Ideally, lower values indicate that the errors are minimal and not systematically related to the input. Achieving values close to zero signifies that the model's errors are random and not influenced by the input, which is an excellent outcome.

11.8 Statistical tests

To validate the superiority of MFCC + TDNN over other combinations like- MFCC + FFNN, LPC + TDNN, LPC + FFNN, FFT + TDNN, and FFT + FFNN, statistical tests has been conducted. The confidence intervals, p-values, and effect size calculations have been done [41]. Tables 10-13 showcase the Bangla phoneme, word, command and sentence recognition in FFNN and TDNN respectively.

11.8.1 Confidence intervals (CI)

A confidence interval [41] provides a range within which the true accuracy of each method is likely to fall. The experiment showcased confidence interval for MFCC + TDNN is significantly higher than that of other combinations, it suggests that MFCC + TDNN is the superior technique. Compute the mean accuracy and standard deviation for each method the 95% confidence interval formula:

$$CI = x \pm Z \times \frac{\delta}{\sqrt{n}}$$

where, x = mean accuracy; Z = 1.96 (for 95% confidence); σ (sigma) = standard deviation; n = sample size.

11.8.2 Hypothesis testing (p-value)

A T-test has been utilized to compare the accuracy distributions of MFCC + TDNN vs. other combinations.

- Null Hypothesis (H_0): There is no significant difference between MFCC + TDNN and other methods.
- Alternative Hypothesis (H_1): MFCC + TDNN has significantly higher accuracy than other methods.
- Compute the p-value:

If $p < 0.05$, Reject H_0, meaning MFCC + TDNN is statistically superior.

Table 10. Bangla phoneme recognition in FFNN and TDNN

Speaker (No. of Phoneme: 08)	Feature Extraction Methods	FFNN (Recognition Percentage Range, Mean Accuracy)	TDNN (Recognition Percentage Range, Mean accuracy)
Single male speaker/No. of utterances recognize out of 40	FFT	82% - 95%, 87%	75% - 85%, 85.83%
	LPC	77% - 90%, 83.83	65% - 90%, 82.17%
	MFCC	92% - 100%, 96.5%	90% - 100%, 96.5%
Single female speaker/No. of utterances recognize out of 40	FFT	72% - 92%, 82.67%	72% - 90%, 77.33%
	LPC	72% - 95%, 79.67%	72% - 90%, 81.33%
	MFCC	97% - 100%, 98.5%	97% - 100%, 97.5%
12 male-female speakers/No. of utterances recognize out of 480	FFT	54% - 60%, 57%	59% - 66%, 62.33%
	LPC	54% - 55%, 54.17%	55% - 70%, 60.5%
	MFCC	86% - 99%, 92.17%	80% - 97%, 86.5%

Table 11. Bangla word recognition in FFNN and TDNN

Speaker (No. of Word: 08)	Feature Extraction Methods	FFNN (Recognition Percentage range, Mean Accuracy)	TDNN (Recognition Percentage Range, Mean Accuracy)
Single male speaker/No. of utterances recognize out of 40	FFT	65% - 70%, 67.33%	60% - 67%, 64.5%
	LPC	55% - 87%, 71.67%	45% - 80%, 65.17%
	MFCC	90% - 100%, 94.33%	90% - 97%, 94.33%
Single female speaker/No. of utterances recognize out of 40	FFT	70% - 75%, 72.67%	57% - 75%, 70.67%
	LPC	82% - 87%, 86.17%	82% - 87%, 84.5%
	MFCC	87% - 90%, 92.5%	95% - 100%, 96.5%
10 male-female speakers/No. of utterances recognize out of 400	FFT	50% - 60%, 55%	51% - 60%, 52.67%
	LPC	43% - 47%, 43.67%	47% - 53%, 52%
	MFCC	75% - 93%, 90.17%	75% - 93%, 87

Table 12. Bangla command recognition in FFNN and TDNN

Speaker (No. of Command: 08)	Feature Extraction Methods	FFNN (Recognition Percentage Range, Mean Accuracy)	TDNN (Recognition Percentage range, Mean Accuracy)
Single male speaker/ No. of utterances recognize out of 40	FFT	40% - 70%, 55%	47% - 70%, 58.5%
	LPC	65% - 72%, 68.67%	65% - 75%, 71.5%
	MFCC	97% - 100%, 98%	97% - 100%, 97.5%
Single female speaker/ No. of utterances recognize out of 40	FFT	25% - 32%, 27.83%	30% - 47%, 36%
	LPC	35% - 50%, 42.5%	37% - 65%, 46.67%
	MFCC	87% - 90%, 87.5%	85% - 92%, 89.17%
10 male-female speakers/ No. of utterances recognize out of 400	FFT	16% - 32%, 22.83%	16% - 32%, 24.83%
	LPC	21% - 32%, 22.83%	21% - 45%, 26%
	MFCC	72% - 83%, 76.5%	57% - 80%, 70.17%

Table 13. Bangla Sentence Recognition in FFNN and TDNN

Speaker (No. of Sentence: 06)	Feature Extraction Methods	FFNN (Recognition Percentage Range, Mean Accuracy)	TDNN (Recognition Percentage Range, Mean Accuracy)
Single male speaker / No. of utterances recognize out of 30	FFT	67% - 70%, 68.5%	70%- 73%, 71.5%
	LPC	50% - 53%, 51.5%	53%- 76%, 64.17%
	MFCC	90% - 99%, 92%	90%- 97%, 93.17%
Single female speaker/ No. of utterances recognize out of 30	FFT	50% - 63%, 56.5%	50%- 53%, 51.5%
	LPC	43% - 50%, 48.33%	43%- 57%, 52.33%
	MFCC	77% - 80%, 79%	80%- 93%, 82.17%
10 male-female speakers/ No. of utterances recognize out of 300	FFT	43% - 50%, 45.17%	43%- 57%, 48.5%
	LPC	31% - 35%, 34.33%	44%- 49%, 48.17%
	MFCC	54% - 89%, 71.83%	65%- 93%, 75.83%

11.8.3 Effect size (Cohen's d)

To measure the magnitude of the difference between MFCC + TDNN and other methods:

$$d = \frac{x1 - x2}{s}$$

where, x_1 = mean accuracy of MFCC + TDNN; x_2 = mean accuracy of other methods; S = pooled standard deviation.

A higher Cohen's d (above 0.8) indicates a strong effect size, meaning MFCC + TDNN is significantly better.

12. CONCLUSIONS AND FUTURE WORK

The investigation demonstrated that MFCC produced the best results for identifying individual words, sentences, instructions, and Bangla phonemes. When it came to voice recognition accuracy, TDNN performed marginally better than FFNN, demonstrating its ability to handle the temporal dependencies in speech signals. Bangla voice recognition has benefited greatly from the thorough methodology and meticulous testing. Several experiments have shown that the accuracy of voice recognition is impacted by the variation in window frame usage (Blackman, Hanning, and Hamming Window frames during feature extraction). The Hanning Window, in particular, performs marginally better than the Blackman and Hamming Window frames. Variability in gender and the number of participants affect the accuracy rate of voice recognition in every instance. The accuracy rate of recognition can occasionally be decreased by increasing the number of participants. As a feature extraction method, MFCC works well with neural networks for recognition accuracy, and occasionally with TDNN. As a recognition tool with a very large (primary/secondary) Bangla dataset, CNN, Vector Quantization, Dynamic Time Warping, Delta-MFCC, Perceptual Linear Prediction, PLP-Relative Spectra, or alternative feature extraction methods with variability of window frames (Bartlett, Bartlett–Hann, Planck–Bessel, Hann–Poisson, and Lanczos windows) and window lengths.

REFERENCES

[1] Islam, M.S. (2009). Research on Bangla language processing in Bangladesh: Progress and challenges. In 8th International Language & Development Conference, pp. 23-25.

[2] Chatterji, S.K. (2024). The Origin and Development of the Bengali Language: Volume One. Routledge. https://doi.org/10.4324/9781003480945

[3] Cardona, G., Jain, D.S. (2003). The Indo-Aryan Languages. Routledge, pp. 87-90. https://books.google.com/books?id=OtCPAgAAQBAJ&pg=PA87.

[4] Banglapedia: National Encyclopedia of Bangladesh. Dhaka, Bangladesh: Asiatic Society of Bangladesh. https://en.banglapedia.org/index.php?title=Main_Page.

[5] Forgie, C., Groves, M.L., Frick, F.C. (1958). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 30: 669. https://doi.org/10.1121/1.1929935

[6] Olson, H.F., Belar, H. (1956). Phonetic typewriter. The Journal of the Acoustical Society of America, 28(6): 1072-1081. https://doi.org/10.1121/1.1908561

[7] Forgie, J.W., Forgie, C.D. (1959). Results obtained from a vowel recognition computer program. The Journal of the Acoustical Society of America, 31(11): 1480-1489. https://doi.org/10.1121/1.1907653

[8] Fujisaki, H., Nakamura, N., Yoshimune, K. (1970). Analysis, Nominalization and recognition of sustained Japanese vowels. The Journal of the Acoustical Society of Japan, 26(3): 152-154. https://doi.org/10.20697/jasj.26.3_152

[9] Sakai, T., Doshita, S. (1962). An automatic recognition system of speech sounds. Studia Phonologica II, 2: 83-95.

[10] Nagata, K., Kato, Y., Chiba, S. (1964). Spoken digit recognizer for Japanese language. In Audio Engineering Society Convention 16. Audio Engineering Society.

[11] Fry, D.B. (1959). Theoretical aspects of mechanical speech recognition. Journal of the British Institution of Radio Engineers, 19(4): 211-218. https://doi.org/10.1049/jbire.1959.0026

[12] Vintsyuk, T.K. (1968). Speech discrimination by dynamic programming. Cybernetics, 4: 52-57. https://doi.org/10.1007/BF01074755

[13] Rabiner, L., Levinson, S., Rosenberg, A., Wilpon, J.A.Y.G. (1979). Speaker-independent recognition of isolated words using clustering techniques. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(4): 336-349. https://doi.org/10.1109/TASSP.1979.1163259

[14] Lesser, V., Fennell, R., Erman, L., Reddy, D. (2003). Organization of the HEARSAY II speech understanding system. IEEE Transactions on Acoustics, Speech, and Signal Processing, 23(1): 11-24. https://doi.org/10.1109/TASSP.1975.1162648

[15] Rintaluoma, T., Silven, O. (2007). Energy efficiency of mobile video decoding. In 2007 International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation, Samos, Greece, pp. 103-109. https://doi.org/10.1109/ICSAMOS.2007.4285740

[16] Rabiner, L., Juang, B.H. (2008). Historical perspective of the field of ASR/NLU. In Springer Handbook of Speech Processing, pp. 521-538. https://doi.org/10.1007/978-3-540-49127-9_26

[17] Furui, S. (1995). Speech recognition-past, present, and future. NTT Review, 7(2): 13-18.

[18] Sanchez, Z. (2025). Navigating the evolution of automatic speech recognition (ASR). ML-Blog, LaMarr Institute. https://lamarr-institute.org/blog/automatic-speech-recognition-evolution.

[19] Swarna, S.T., Ehsan, S., Islam, M.S., Jannat, M.E. (2017). A comprehensive survey on Bengali phoneme recognition. arXiv preprint arXiv:1701.08156. https://doi.org/10.48550/arXiv.1701.08156

[20] Mukherjee, H., Halder, C., Phadikar, S., Roy, K. (2017). READ-a Bangla phoneme recognition system. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, pp. 599-607. https://doi.org/10.1007/978-981-10-3153-3_59

[21] Eity, Q.N., Banik, M., Lisa, N.J., Hassan, F., Hossain, M.S., Huda, M.N. (2010). Bangla speech recognition using two stage multilayer neural networks. In 2010 International Conference on Signal and Image Processing, Chennai, India, pp. 222-226. https://doi.org/10.1109/ICSIP.2010.5697473

[22] Hasan, M.M., Hassan, F., Islam, G.M.M., Banik, M., Kotwal, M.R.A., Rahman, S.M.M., Mohammad, N.H. (2010). Bangla triphone hmm based word recognition. In 2010 IEEE Asia Pacific Conference on Circuits and Systems, Kuala Lumpur, Malaysia, pp. 883-886. https://doi.org/10.1109/APCCAS.2010.5775010

[23] Ahamed, B., Israt, F., Chowdhury, S.M.R., Huda, M.N. (2013). Effect of speaker variation on the performance of Bangla ASR. In 2013 International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, pp. 1-5. https://doi.org/10.1109/ICIEV.2013.6572578

[24] Sadeq, N., Ahmed, S., Shubha, S.S., Islam, M.N., Adnan, M.A. (2020). Bangla voice command recognition in end-to-end system using topic modeling based contextual rescoring. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 7894-7898. https://doi.org/10.1109/icassp40776.2020.9053970

[25] Nahid, M.M.H., Islam, M.A., Islam, M.S. (2016). A noble approach for recognizing Bangla real number automatically using CMU sphinx4. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, Bangladesh, pp. 844-849. https://doi.org/10.1109/ICIEV.2016.7760121

[26] Kadir, A., Rahman, M. (2016). Bangla speech sentence recognition using hidden Markov models. International Journal of Multidisciplinary Research and Development, 3(7): 122-127.

[27] Hossain, S.A., Rahman, M.L., Ahmed, F. (2005). Spectral analysis of Bangla vowels. In 2005 Pakistan Section Multitopic Conference, Karachi, Pakistan, pp. 1-5. https://doi.org/10.1109/INMIC.2005.334396

[28] Muslima, U., Islam, M.B. (2014). Experimental framework for Mel-scaled LP based Bangla speech recognition. In 16th International Conference Computer and Information Technology, Khulna, Bangladesh, pp. 56-59. https://doi.org/10.1109/ICCITechn.2014.6997304

[29] Abou-Loukh, S.J., Abdul-Razzaq, S.M. (2013). Isolated word speech recognition using mixed transform. Journal of Engineering, 19(10): 1271-1286. https://doi.org/10.31026/j.eng.2013.10.06

[30] Kivaisi, A.R., Zhao, Q., Mbelwa, J.T. (2023). Swahili speech dataset development and improved pre-training method for spoken digit recognition. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(7): 1-24. https://doi.org/10.1145/3597494

[31] Hussain, M.G., Rahman, M., Sultana, B., Khatun, A., Al Hasan, S. (2021). Classification of Bangla alphabets phoneme based on audio features using MLPC & SVM. In 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), Rajshahi, Bangladesh, pp. 1-5. https://doi.org/10.1109/ACMI53878.2021.9528088

[32] Das, B., Mandal, S., Mitra, P. (2011). Bengali speech corpus for continuous auutomatic speech recognition system. In 2011 International Conference on Speech Database and Assessments (Oriental COCOSDA), Hsinchu, Taiwan, pp. 51-55. https://doi.org/10.1109/ICSDA.2011.6085979

[33] Ghosh, T., Saha, S., Ferdous, A.I. (2016). Formant analysis of Bangla vowel for automatic speech recognition. Signal & Image Processing: An International Journal (SIPIJ), 7(5): 1-10. https://doi.org/10.5121/sipij.2016.7501

[34] Engineering and Physical Sciences Research Council (EPSRC). (2003). The EMILLE Corpus. Lancaster University. https://www.lancaster.ac.uk/fass/projects/corpus/emille.

[35] Ittichaichareon, C., Suksri, S., Yingthawornsuk, T. (2012). Speech recognition using MFCC. In International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Pattaya, Thailand, pp. 135-138.

[36] Asadullah, M., Nisar, S. (2016). A silence removal and endpoint detection approach for speech processing. Sarhad University International Journal of Basic and Applied Sciences, 4(1): 10-15.

[37] Shih, F.Y. (2010). Image Processing and Pattern Recognition: Fundamentals and Techniques. John Wiley & Sons. https://doi.org/10.1002/9780470590416

[38] Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. (1992). Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press.

[39] Haddad, R.A., Parsons, T.W. (1991). Digital Signal Processing: Theory, Applications, and Hardware. Computer Science Press, Inc., United States, p. 636.

[40] Bäckström, T., Räsänen, O., Zewoudie, A., Pérez Zarazaga, P., Koivusalo, L., Das, S., Gómez Mellado, E., Bouafif Mansali, M., Ramos, D., Kadiri, S., Alku, P., Vali, M.H. (2022). Windowing. In Introduction to Speech Processing. https://speechprocessingbook.aalto.fi/Representations/Windowing.html.

[41] Tan, S.H., Tan, S.B. (2010). The correct interpretation of confidence intervals. Proceedings of Singapore Healthcare, 19(3): 276-278. https://doi.org/10.1177/201010581001900316