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

Mathematical Modelling of Engineering Problems
Vol. 12, No. 9, September, 2025, pp. 3033-3052
Journal homepage: http://iieta.org/journals/mmep

Bangla Speech Processing: Time Delay Neural Networks Enhanced by Advanced Algorithms

Md. Shafiul Alam Chowdhury[1,2*], Md. Farukuzzaman Khan[1], Shaikh Atisha Rahbath Dip[2], S M Nazmus Sadat[2], Sumaiya Tanjil Khan[2], Zarin Tasnim[2], Md. Shafikul Islam[2]

[1] Department of Computer Science and Engineering, Islamic University, Kushtia 7003, Bangladesh
[2] Department of Computer Science and Engineering, Uttara University, Dhaka 1230, Bangladesh

Corresponding Author Email: shafiul.a.chowdhury@gmail.com

https://doi.org/10.18280/mmep.120908

Received: 10 June 2025
Revised: 19 August 2025
Accepted: 25 August 2025
Available online: 30 September 2025

Keywords:
Mel Frequency Cepstral Coefficient (MFCC), Power Spectral Analysis (FFT), Linear Predictor Coefficient Analysis (LPC), Time-Delay Neural Networks (TDNN), Levenberg–Marquardt Algorithm (LMA), Bayesian Regularization Algorithm (BRA), Scaled Conjugate Gradient Algorithm (SCGA)

ABSTRACT

This study explores critical challenges in Bangla speech recognition by evaluating phoneme, word, command, and sentence-level recognition using a MATLAB-based framework. The feature extraction methods Mel-Frequency Cepstral Coefficients (MFCC), Power Spectral Analysis, and Linear Predictive Coding (LPC) are applied with Blackman, Hamming, and Hanning windowing techniques. Time Delay Neural Network (TDNN) models are trained using three optimization algorithms: Scaled Conjugate Gradient Algorithm (SCGA), Levenberg–Marquardt Algorithm (LMA), and Bayesian Regularization Algorithm (BRA). Results indicate that MFCC combined with TDNN, optimized via LMA, BRA, or SCGA, yields the highest recognition accuracy, reaching up to 94%. Six experiments are analyzed, including five from existing literature and one representing the current study. Comparative evaluation and statistical analysis, including confidence intervals, are employed to identify the most effective configuration. The findings outperform previous approaches and underscore the influence of sample size, speaker gender, and windowing methods on recognition performance. These insights offer a foundation for future improvements in Bangla speech technology.

1. INTRODUCTION

Despite remarkable global progress in automatic speech recognition (ASR), the majority of research and development has focused on high-resource languages, particularly English. Bangla (Bengali), spoken by over 300 million people worldwide, remains significantly underrepresented in speech technology initiatives [1, 2]. This disparity stems from the linguistic intricacies of Bangla, including its rich morphology, compound characters, and phonetic diversity, which pose unique challenges for accurate recognition. Historically, Bangla ASR systems have concentrated on phoneme-level, digit-level, or command-based recognition, often neglecting continuous speech and sentence-level understanding [1]. Consequently, there exists a substantial research gap, offering fertile ground for innovation in Bangla speech processing. To address these limitations, recent studies have explored advanced feature extraction techniques, such as Power Spectral Analysis (FFT), Linear Predictive Coefficients (LPC), and Mel-Frequency Cepstral Coefficients (MFCC). These features are integrated into diverse machine learning and deep learning models title- Time-Delay Neural Networks (TDNN). To further enhance recognition accuracy, optimization strategies such as Levenberg–Marquardt Algorithm (LMA), Bayesian Regularization Algorithm (BRA), and Scaled Conjugate Gradient Algorithm (SCGA) have been employed [3, 4]. These efforts mark a pivotal shift toward building robust, scalable, and inclusive ASR systems for Bangla, with promising applications.

2. OVERVIEW OF BANGLA LANGUAGE

Bangla, also known as Bengali, is a linguistically rich and culturally vibrant language spoken by around 300 million people across Bangladesh, West Bengal, and substantial communities in Assam, Tripura, and the Andaman and Nicobar Islands of India. As a member of the Eastern Indo-Aryan branch of the Indo-European language family, Bangla has evolved through a complex historical trajectory. Its development traces back to ancient vernaculars, such as Magadhi Prakrit, Ardha-Magadhi, and Apabhramsha, which themselves emerged from Vedic Sanskrit. These dialects played a pivotal role in shaping the phonological, syntactic, and lexical features of modern Bangla. The language's evolution reflects centuries of cultural exchange, religious movements, and literary innovation, culminating in a distinct linguistic identity that continues to influence regional speech patterns and modern computational linguistics, including automatic speech recognition systems [5-8].

3. HISTORICAL PERSPECTIVE OF SPEECH RESEARCH

Acoustic-phonetics was key to early ASR, helping researchers understand speech elements and their realization in spoken language. Forgie et al. [9] pioneered speech recognition by developing a system to automatically identify spoken digits. Their work laid the foundation for future voice-based technologies by demonstrating early success in acoustic pattern analysis. Then, Forge and Forgie's [10] groundbreaking research at MIT Lincoln Laboratory propelled the field forward, shaping its future innovations. By focusing on a speaker-independent system, they tackled the challenge of speech variability among individuals, while Sakai and Doshita's [11] groundbreaking work at Kyoto University on the phoneme recognizer advanced speech recognition by incorporating a speech segmenter to dissect signals for more precise analysis [11]. Fry [12] developed a phoneme recognition system focusing on four English vowels and consonants, pioneering the use of statistical syntax in speech recognition, while IBM, led by Jelinek, advanced speaker-dependent systems. IBM focused on a speaker-dependent voice-activated typewriter, requiring users to train the system to recognize their speech patterns. Boll [13] proposed a speaker-independent isolated word recognition system using clustering, dynamic time warping, and vector quantization. The method improved recognition accuracy and efficiency across different speakers. The 2020 period was pivotal for ASR, as the integration of deep neural networks (DNNs) revolutionized the field by enabling the modeling of complex, non-linear relationships in speech data, significantly enhancing recognition accuracy [14].

4. LITERATURE REVIEW OF BANGLA SPEECH RECOGNITION

Bangla speech recognition has experienced notable progress over the past decade, primarily fueled by advancements in deep learning and the emergence of Bangla language datasets though these resources remain relatively scarce. Initial approaches relied on rule-based methods and classical machine learning techniques, but recent research has shifted towards DNNs [15]. ASR research in the language remains limited in quality, with studies on phoneme recognition in 40 native speakers showing MFCC outperforming Linguistic Feature techniques [16], while a Bengali speech corpus was developed to enhance continuous automatic speech recognition systems for Bengali language users [17]. A study on Bangla phoneme recognition explored Hidden Markov Models (HMMs) with single and multi-layer neural networks, aiming to enhance precision by analyzing the strengths and weaknesses of different neural network topologies [18]. Rahman and Khatun [19] developed a speaker-independent system for recognizing isolated Bangla words using MFCC for feature extraction and Euclidean distance for classification. Tested on 600 words, it achieved 84.28% accuracy for multi-speaker input, demonstrating effective performance across different speakers. Nahid et al. [20] introduced a Bengali speech recognition system using a double-layered LSTM-RNN model. It processes MFCC features to predict phonemes, which are then filtered to reconstruct words. Tested on the Bangla-Real-Number dataset and achieved 13.2% of word error rate.

A medium-sized Bangla speech corpus, featuring 40 native speakers from diverse regions, was developed to compare acoustic features for word recognition, with experiments showing that MFCC-based methods outperform others in word correct rate (WCR) [21]. Kibria et al. [22] developed SUBAK.KO, a large Bangladeshi Bangla speech corpus for automatic speech recognition. Using RNN with CTC, the system showed improved accuracy over existing datasets, supporting robust LVCSR and regional accent coverage. Gender-Independent (GI) ASR, designed to reduce gender influence using acoustic and local features, outperformed MFCC-based methods with fewer mixture components, improving efficiency [23]. The READ system for Bangla phoneme recognition claimed 98.35% accuracy for vowel phonemes but did not account for Bangla consonants or accent variations between West Bengal and Bangladesh [24]. To address challenges such as phonetic complexity, speaker variability, and limited annotated corpora, researchers have developed medium-scale datasets and leveraged advanced machine learning strategies. A comprehensive survey [1] underscored crucial design considerations, including vocabulary size, speaker dependency, and classification methods, while highlighting the pivotal role of dataset quality and model selection in improving recognition accuracy. These developments have markedly enhanced the effectiveness of Bangla ASR systems, enabling a wide range of applications—from transcription services to voice-controlled interfaces and accessibility technologies [1].

5. THE SCOPE OF THIS RESEARCH

This research investigates feature extraction and recognition techniques for Bangla speech signals, aiming to develop a high-accuracy speech recognition system and perform a comparative analysis of recognition methods. It focuses on phoneme, isolated word, command, and sentence-level recognition using primary (1,500 samples from male and female speakers across diverse age groups) datasets. Key algorithms were implemented in MATLAB, alongside essential pre-processing techniques including short-time energy calculation, silence removal, and window framing with Hamming, Hanning, and Blackman windows. Feature extraction methods FFT, LPC, and MFCC were employed to construct training and target datasets. The study evaluates advanced neural network models, including TDNN combined with LMA, BRA, and SCGA optimization techniques, and presents detailed experimental outcomes. Also, a total of six experiments are showcased five drawn from prior research and one representing the current study. These experiments are systematically compared to assess performance differences. Statistical analyses, including confidence intervals, are conducted to rigorously evaluate and identify the most effective approach among them. It concludes with meaningful insights and recommendations for future research directions.

6. NOVEL CONTRIBUTIONS

- An insightful comparative study of diverse feature extraction techniques and TDNN-powered speech recognition tools (LMA, BRA, SCGA algorithms), implemented within a unified experimental framework that seamlessly integrates Bangla

phonemes, isolated words, commands, and sentences for comprehensive linguistic analysis.

- To explore feature extraction and deep learning tools, this study emphasizes the dynamic variability of frame windowing techniques Hamming, Hanning, and Blackman for enhanced precision, while providing a curated Bangla dataset to address resource scarcity in experimental research.
- Provide a systematic performance comparison between traditional methods (e.g., statistical classifiers, template matching) and modern neural approaches.
- A critical analysis of both contemporary and historical research in Bangla speech recognition, aimed at addressing the limitations identified in earlier studies.

7. SPEECH RECOGNITION COMPLEXITIES

Speech recognition is a powerful yet highly complex technology that faces a range of challenges:

- **Acoustic Variability:** Speech recognition accuracy is shaped by speaker differences such as accent, gender, and age as well as background noise and microphone quality, all of which impact audio clarity.
- **Linguistic Challenges:** Homophones, ambiguous context, and speech disfluencies hinder recognition by blurring distinctions between similar-sounding words and meanings.
- **Technical Issues:** Real-time speech recognition requires powerful computing, efficient algorithms, and large annotated datasets to overcome data scarcity and model complexity.
- **Ethical and Social Considerations:** Speech recognition systems must protect user privacy, reduce demographic bias, and improve accessibility for those with atypical speech.

8. SPECIFIC GAPS IN BANGLA ASR RESEARCH

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

- Bangla ASR research struggles with limited annotated datasets and diverse regional speech patterns, hindering development of consistent, generalized models.
- Bangla ASR research mainly focuses on phoneme and word-level recognition, while command and sentence-level processing remain limited, yet essential for advanced applications requiring strong contextual modeling.

9. OBJECTIVES OF THE STUDY

- This research advances Bangla speech recognition by developing a 1,500-sample dataset including phonemes, words, commands, and sentences from male-female Bangladeshi native speakers enhancing model adaptability and recognition accuracy across genders and age groups.
- This research enhances Bangla sentence-level recognition through contextual learning and a unified comparison of feature extraction and ASR models advancing the field and promoting future innovation.
- By addressing these challenges, this study aims to strengthen the robustness, scalability, and real-world applicability of Bangla ASR systems.

10. THE EXPERIMENT METHODS

This study investigates speech signals from male and female speakers across diverse age groups to evaluate the recognition accuracy of Bangla phoneme utterances, individual words, commands, and sentences. To ensure precise speech analysis, multiple windowing techniques such as Hanning, Hamming and Blackman (HN, HM, and BL) windows are applied for effective signal processing. A range of feature extraction methods is employed to capture essential speech characteristics, thereby enhancing model performance. Advanced speech recognition tools are used to assess the system's accuracy in identifying and interpreting Bangla speech, with particular attention to gender-based variations in pronunciation and articulation. A foundational dataset comprising approximately 1,500 speech samples (Table 1) has been collected from speakers of varying age groups. These samples reflect diverse linguistic attributes, enabling a comprehensive evaluation of the system's ability to recognize speech across demographic differences. By incorporating a wide range of voices, the study aims to improve the adaptability and robustness of Bangla speech recognition technology, ensuring reliable performance across real-world applications.

10.1 Short-time energy calculation and silence removal

To facilitate precise and efficient analysis of speech signals, all audio data were segmented into fixed-length rectangular window frames of 16 milliseconds (Figure 1). This segmentation strategy is grounded in the principle that short, overlapping frames can effectively capture the dynamic nature of human speech, which varies rapidly over time. By dividing the signal into these manageable units, the system is able to extract localized acoustic features while maintaining computational efficiency a critical consideration for real-time or large-scale speech processing tasks. Each frame serves as a snapshot of the speech waveform, preserving essential temporal and spectral characteristics. However, raw speech signals often contain silent or low-energy regions that do not contribute meaningful information to the recognition process. To address this, Short-Time Energy (STE) analysis was employed. STE is a widely used technique for quantifying the energy content of a signal within a short time window, making it particularly effective for identifying silent segments. By calculating the energy of each frame, the system can distinguish between voiced and unvoiced regions, allowing for the removal of frames that fall below a defined energy threshold [25-27]. These low-energy frames, typically corresponding to pauses, background noise, or weak articulations, can introduce unnecessary variability and degrade the performance of feature extraction algorithms. Their elimination ensures that only acoustically rich segments are retained for further analysis. To enhance consistency

across frames and improve the reliability of recognition, energy normalization was applied. This process scales the energy values of each frame relative to the maximum observed energy, ensuring uniformity in amplitude and reducing the influence of speaker-specific loudness variations. Following normalization, frames with energy levels below 2% of the maximum energy were systematically discarded. This threshold-based filtering ensures that the retained frames contain sufficient acoustic information to support accurate phoneme and word recognition. By focusing exclusively on high-energy, information-rich segments, the pre-processing pipeline enhances the clarity and intelligibility of the speech signal.

This multi-step pre-processing approach comprising segmentation, STE-based silence removal, energy calculation, normalization, and thresholding results in a cleaner and more representative signal. It significantly improves the robustness and accuracy of the Bangla speech recognition system by minimizing noise, reducing irrelevant variability, and emphasizing linguistically meaningful content. These enhancements are particularly valuable in real-world applications, where speech input may be affected by environmental noise, speaker variability, and inconsistent articulation.

Table 1. Bangla recorded audio samples

Category	Bangla (English Accent)	Properties		In Seconds
Phoneme	অ (/O/) আ (/A/) ই (/I/) ঊ(/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		1.018–1.201
Category	**Bangla**	**English Accent**	**English Meaning**	**In Seconds**
Isolated Word	অংক আমি ইলিশ উট কলা খরেগাশ গরু ঘড়ি	Onko Ami Ilish Ut Kola Khorgosh Goru Ghuri	Math I Ilish (Fish) Camel Banana Rabbit Cow Clock	1.201
Command	এই কাজ কর দরজা খোলো টেবিল পরিস্কার কর বাম দিক যাও পশ্চিম দিক সরো অফিস যাও এই চেয়ার আনো জানালা বন্ধ কর	Ai kaj koro Dorja kholo Table poriskar koro Bam dik jao Poschim dik soro Office jao Ai chair ano Janala bondho koro	Do the job Open the door Clean the table Move toward the left Move toward the west Go to the office Bring this chair Close the window	1.802–2.716
Sentence	আমরা কলা খাই কলা ভালো ফল ফল স্বাস্থ্যের জন্য ভালো তিন বন্ধ খেলা করে তারা তিন বন্ধু তিন বন্ধু খায়	Amra kola khai Kola valo fol Fol shaster jonno valo Tin bondhu khela kore Tara tin bondhu Tin bondhu khae	We eat bananas Banana is a good fruit Fruit is good for health They are three friends Three friends play Three friends eat	2.011–3.213

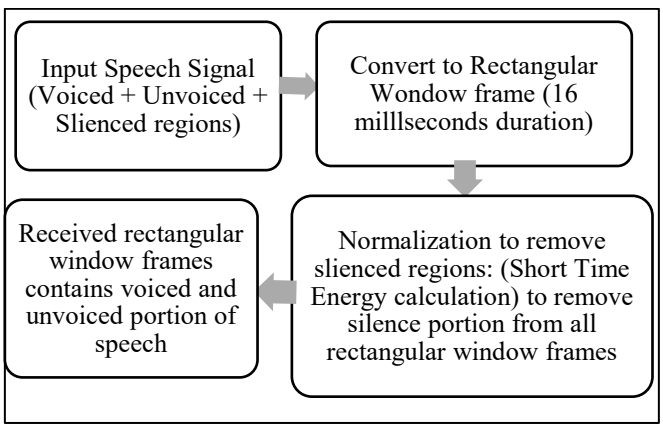

Figure 1. Short-time energy calculation and silence removal

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

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

The corresponding $w0(n)$ function is a cosine without $\pi/2$ phase offset [26, 27].

10.2 Hamming window framing

The Hamming window [28] is defined by the following Eq. (2):

$$w(n) = 0.54 - 0.46\, cos(2\pi n/N), (0 \le n \le N) \tag{2}$$

The window length $L = N+1$.

Let L denote the window length, defined as a positive integer, and w represent the Hamming window column vector utilized for signal processing (Figure 2). The Hamming window, known for its smooth tapering at the edges, was applied to each frame to minimize spectral leakage, a common issue in frequency analysis that can distort the representation of signal components. The window length was carefully chosen to align with the frame size, ensuring optimal segmentation and preserving the integrity of the speech signal during analysis. Following the windowing process, the speech signal was subjected to spectral analysis to extract key features critical for accurate recognition. Among these, the spectral envelope was a primary focus. This feature captures the overall shape of the frequency spectrum and reflects variations in energy distribution across different frequency bands. The spectral envelope provides a detailed acoustic profile of the speech signal, making it instrumental in distinguishing between phonemes and improving the precision of Bangla speech recognition models.

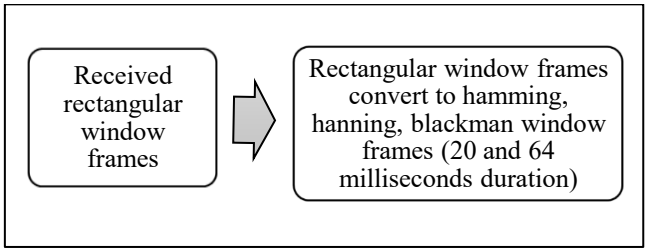

Figure 2. Hamming, Hanning, Blackman window frame

By integrating windowing techniques with spectral feature extraction, the system achieves a more nuanced understanding of speech dynamics. This combination enhances the model's ability to interpret complex speech patterns, ultimately contributing to more robust and accurate recognition performance across diverse linguistic inputs.

10.3 Pre-processing

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) [29]:

$$y(n) = s(n) - Cxs(n-1) \tag{3}$$

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

The pre-emphasis was performed by using an all-zero filter [29]. Three different pre-processing approaches were used:

Pre-processing = (Hamming/Hanning/Blackman) Window+Pre-emphasis

Each frame of the speech signal underwent a detailed pre-processing phase, with the variable frame storing all individual segments generated by the framing function. This step is essential in preparing the raw signal for subsequent analysis, as it transforms the continuous waveform into discrete, time-localized units suitable for feature extraction. While zero-padding is a common technique used to enhance the spectral representation by artificially increasing the length of the signal and thereby improving the frequency domain resolution it was found to be ineffective in this particular experiment. Specifically, zero-padding did not contribute to a meaningful improvement in spectral resolution or feature clarity. Consequently, both zero-padding and frame overlapping were intentionally omitted during the segmentation process. This decision was made to preserve the natural temporal boundaries of the speech signal and to avoid introducing artifacts that could compromise the integrity of the extracted features (Figure 3 is about the internal architecture of TDNN).

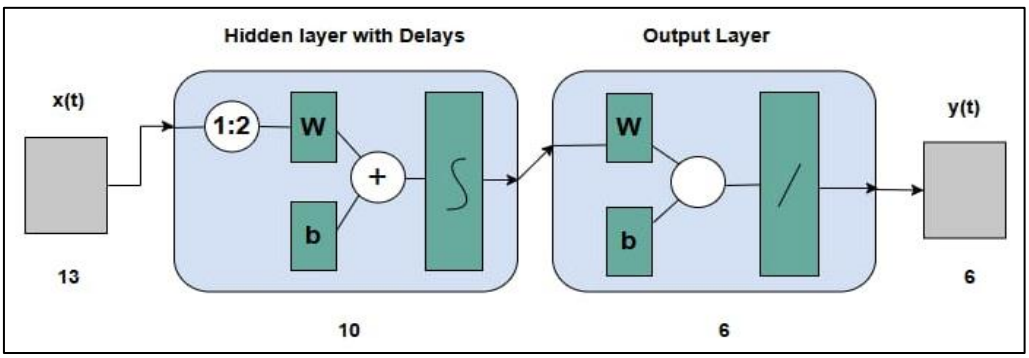

Figure 3. TDNN

The choice of window length plays a pivotal role in speech signal processing, particularly in stabilizing time-variant signals. By segmenting the signal into short frames, the system can assume quasi-stationarity within each frame, which is a prerequisite for accurate spectral analysis. Window length directly influences the trade-off between time and frequency resolution. Shorter windows, typically ranging from 5 to 25 milliseconds, are adept at capturing rapid transitions in speech, such as those found in plosive or fricative phonemes. However, their limited duration can lead to spectral smearing, reducing the precision of frequency-based features. On the other hand, longer windows spanning 25 to 64 milliseconds provide superior frequency resolution, making them suitable for analyzing steady-state vowel sounds and tonal variations. Yet, they may obscure transient features due to temporal averaging.

To address these competing demands, the experiment strategically employed both short and long window lengths. This dual-window approach enabled the capture of a broader spectrum of speech characteristics, from fast phonemic shifts to sustained harmonic structures. By leveraging the strengths of each window type, the analysis achieved a more holistic representation of the speech signal, thereby enhancing the robustness and accuracy of feature extraction for Bangla

speech recognition tasks [30].

High-quality datasets for Bangla speech recognition are notably scarce, making it challenging to conduct effective research in this domain. As a result, most researchers working on Bangla speech recognition tend to rely on their own datasets, typically primary data collected for specific experimental purposes.

A primary dataset comprising 1,500 samples (Table 1) was collected from male and female participants spanning various age groups. All participants were native Bangla speakers residing in Bangladesh. The dataset features eight Bangla phonemes - encompassing both vowels and consonants - namely: অ (/O/), আ (/A/), ই (/I/), উ (/OO/), এ (/EA/), ও (/O/), ঐ (/OI/), and ক (/KO/). Each phoneme sample had a time duration ranging from 1.018 to 1.201 seconds. For the phoneme recognition experiments, between 40 to 480 speech samples were utilized per trial. In the word recognition experiments, eight isolated Bangla words were used: অংক (Math), আমি (I), ইলিশ (Ilish), উট (Camel), কলা (Banana), খরগোশ (Rabbit), গরু (Cow), and ঘড়ি (Clock). Each word had a time duration of 1.201 seconds, with 40 to 400 speech samples employed for each experiment. For Bangla command recognition experiments, eight distinct commands were included in the dataset: এই কাজ কর (Do this job), দরজা খোলো (Open the door), টেবিল পরিস্কার 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), and জানালা বন্ধ কর (Close the window). Each command sample ranged from 1.802 to 2.716 seconds in duration, and 40 to 400 samples were used for each trial. In addition, six Bangla sentences were incorporated for speech recognition experiments: আমরা কলা খাই (We eat bananas), কলা ভালো ফল (Banana is a good fruit), ফল স্বাস্থ্যের জন্য ভালো (Fruit is good for health), তারা তিন বন্ধু (They are three friends), তিন বন্ধু খেলা করে (Three friends play), and তিন বন্ধু খায় (Three friends eat). The sentence durations ranged from 2.011 to 3.213 seconds. Each experiment involved between one to twelve speakers, with contributions from both male and female participants.

11. EXPERIMENTS AND RESULTS

The MATLAB code extracts speech features and partitions the dataset into 60% training, 20% validation, and 20% testing. The network learns by minimizing error during training, while validation monitors generalization and halts training when improvement stops. Testing uses independent data to evaluate final performance without affecting learning. Experiments involved Bangla phonemes, isolated words, commands, and sentences, using a diverse dataset of male and female speakers across various age groups. Results for each configuration are presented in detail. Feature extraction employed three parallel methods: FFT, LPC, and MFCC to capture complementary spectral and temporal characteristics. Framing used 20-ms and 64-ms windows with Hamming, Hanning, and Blackman functions, enabling robust time-frequency analysis and enhancing recognition accuracy.

11.1 Experiment using LMA, BRA, and SCGA in TDNN

The experiment utilized a diverse Bangla speech dataset comprising phonemes, isolated words, commands, and sentences (Table 1), ensuring broad linguistic coverage. Feature extraction was performed using FFT, LPC, and MFCC in parallel, leveraging their complementary strengths in capturing spectral and temporal speech characteristics. Separate experiments were conducted for each speech category to identify optimal feature sets and recognition strategies. Framing employed 20 ms and 64 ms windows with Hamming, Hanning, and Blackman functions to balance time-frequency resolution and reduce spectral leakage. Speech recognition was carried out using a TDNN trained with three different algorithms, enabling comparative analysis of training efficiency and model performance: LMA, BRA, and SCGA.

Table 2. Bangla phoneme recognition in TDNN

08 Unique Phonemes	Feature Extraction Methods	Window Frame Length (HM, HN, BL)	Utterances Recognized from 480 (TDNN& LMA)	Utterances Recognized from 480 (TDNN & BRA)	Utterances Recognized from 480 (TDNN & SCGA)	Percentage of Recognition (TDNN & LMA)	Percentage of Recognition (TDNN & BRA)	Percentage of Recognition (TDNN & SCGA)
Twelve male-female participants	FFT	20 Ms (HM)	319	282	282	67%	60%	60%
		20 Ms. (HN)	319	319	319	67%	67%	67%
		20 Ms. (BL)	312	312	312	65%	65%	65%
		64 Ms (HM)	285	285	285	60%	60%	60%
		64 Ms. (HN)	285	285	285	60%	60%	60%
		64 Ms. (BL)	285	285	285	60%	60%	60%
	LPC	20 Ms (HM)	297	297	297	62%	62%	62%
		20 Ms. (HN)	297	297	297	62%	62%	62%
		20 Ms. (BL)	297	297	297	62%	62%	62%
		64 Ms (HM)	341	341	341	71%	71%	71%
		64 Ms. (HN)	341	341	341	56%	56%	56%
		64 Ms. (BL)	341	341	341	56%	56%	56%
	MFCC	20 Ms (HM)	384	384	384	80%	80%	80%
		20 Ms. (HN)	384	384	384	80%	80%	80%
		20 Ms. (BL)	384	384	384	80%	80%	80%
		64 Ms (HM)	425	425	417	**89%**	**89%**	87%
		64 Ms. (HN)	425	417	425	**89%**	**89%**	**89%**
		64 Ms. (BL)	417	417	417	87%	87%	87%

The LMA, BRA, and SCGA with TDNN offer a powerful framework for enhancing speech recognition performance, particularly in complex linguistic contexts like Bangla. TDNNs are well-suited for capturing temporal dependencies and sequential patterns inherent in spoken language. LMA improves training efficiency by balancing gradient descent and Gauss-Newton methods, yielding faster convergence and improved accuracy. BRA introduces regularization during training to prevent over-fitting, ensuring better generalization across diverse speech data. SCGA further optimizes the training process by reducing computational load and enhancing scalability, making it ideal for large speech datasets. Collectively, these algorithms enable TDNN architectures to effectively model intricate acoustic features and linguistic variations, resulting in higher recognition accuracy and robustness in speech-based applications.

Table 2 focuses on feature extraction of Bangla phonemes using FFT, LPC, and MFCC, and their recognition in a TDNN using three algorithms: LMA, BRA, and SCGA.

Table 3 presents the results of Bangla word feature extraction using FFT, LPC, and MFCC, followed by recognition using a TDNN with three algorithms: LMA, BRA, and SCGA.

Table 4 presents the feature extraction results of Bangla commands using FFT, LPC, and MFCC, followed by recognition using a TDNN with three algorithms: LMA, BRA, and SCGA.

Table 3. Bangla word recognition in TDNN

08 Unique Words	Feature Extraction Methods	Window Frame Length (HM, HN, BL)	Utterances Recognized from 400 (TDNN & LMA)	Utterances Recognized from 400 (TDNN & BRA)	Utterances Recognized from 400 (TDNN & SCGA)	Percentage of Recognition (TDNN & LMA)	Percentage of Recognition (TDNN & BRA)	Percentage of Recognition (TDNN & SCGA)
Ten male-female participants	FFT	20 Ms. (HM)	204	204	204	51%	51%	51%
		20 Ms. (HN)	204	204	208	51%	51%	52%
		20 Ms. (BL)	200	204	200	50%	51%	50%
		64 Ms. (HM)	179	200	200	45%	50%	50%
		64 Ms. (HN)	240	240	240	60%	60%	60%
		64 Ms. (BL)	240	236	240	60%	59%	60%
	LPC	20 Ms. (HM)	212	208	212	53%	52%	53%
		20 Ms. (HN)	212	211	216	53%	53%	54%
		20 Ms. (BL)	212	212	212	53%	53%	53%
		64 Ms. (HM)	191	191	212	48%	48%	53%
		64 Ms. (HN)	212	212	212	53%	53%	53%
		64 Ms. (BL)	212	212	212	53%	53%	53%
	MFCC	20 Ms. (HM)	303	303	303	76%	76%	76%
		20 Ms. (HN)	375	375	375	**94%**	**94%**	**94%**
		20 Ms. (BL)	375	367	375	**94%**	92%	**94%**
		64 Ms. (HM)	303	303	375	76%	76%	**94%**
		64 Ms. (HN)	375	371	375	**94%**	93%	**94%**
		64 Ms. (BL)	375	375	375	**94%**	**94%**	**94%**

Table 4. Bangla command recognition in TDNN

08 Unique Commands	Feature Extraction Methods	Window Frame Length (HM, HN, BL)	Utterances Recognized from 400 (TDNN & LMA)	Utterances Recognized from 400 (TDNN & BRA)	Utterances Recognized from 400 (TDNN & SCGA)	Percentage of Recognition (TDNN & LMA)	Percentage of Recognition (TDNN & BRA)	Percentage of Recognition (TDNN & SCGA)
Ten male-female participants	FFT	20 Ms. (HM)	91	91	91	23%	23%	23%
		20 Ms. (HN)	91	84	91	23%	21%	23%
		20 Ms. (BL)	67	67	67	17%	17%	17%
		64 Ms. (HM)	100	100	99	25%	25%	25%
		64 Ms. (HN)	131	131	131	33%	33%	33%
		64 Ms. (BL)	131	127	131	33%	32%	33%
	LPC	20 Ms. (HM)	99	99	99	25%	25%	25%
		20 Ms. (HN)	99	100	99	25%	25%	25%
		20 Ms. (BL)	84	84	84	21%	21%	21%
		64 Ms. (HM)	180	180	180	45%	45%	45%
		64 Ms. (HN)	84	84	84	21%	21%	21%
		64 Ms. (BL)	84	84	84	21%	21%	21%
	MFCC	20 Ms. (HM)	228	228	228	57%	57%	57%
		20 Ms. (HN)	320	320	316	**80%**	**80%**	79%
		20 Ms. (BL)	320	316	320	**80%**	79%	**80%**
		64 Ms. (HM)	243	243	243	61%	61%	61%
		64 Ms. (HN)	291	291	291	73%	73%	73%
		64 Ms. (BL)	291	291	291	73%	73%	73%

Table 5. Bangla sentence recognition in TDNN

06 Unique Sentences	Feature Extraction Methods	Window Frame Length (HM, HN, BL)	Utterances Recognized from 300 (TDNN & LMA)	Utterances Recognized from 300 (TDNN & BRA)	Utterances Recognized from 300 (TDNN & SCGA)	Percentage of Recognition (TDNN & LMA)	Percentage of Recognition (TDNN & BRA)	Percentage of Recognition (TDNN & SCGA)
Ten male-female participants	FFT	20 Ms. (HM)	131	131	141	44%	44%	47%
		20 Ms. (HN)	167	167	175	57%	57%	60%
		20 Ms. (BL)	150	141	150	50%	47%	50%
		64 Ms. (HM)	141	141	150	47%	47%	50%
		64 Ms. (HN)	141	141	141	47%	47%	47%
		64 Ms. (BL)	141	141	150	47%	47%	50%
	LPC	20 Ms. (HM)	147	147	147	49%	49%	49%
		20 Ms. (HN)	147	150	147	49%	50%	49%
		20 Ms. (BL)	147	141	147	49%	47%	49%
		64 Ms. (HM)	132	133	147	44%	44%	49%
		64 Ms. (HN)	147	147	150	49%	49%	50%
		64 Ms. (BL)	147	147	147	49%	49%	49%
	MFCC	20 Ms. (HM)	231	231	240	77%	77%	80%
		20 Ms. (HN)	281	281	281	**94%**	**94%**	**94%**
		20 Ms. (BL)	270	270	281	90%	90%	**94%**
		64 Ms. (HM)	197	197	197	66%	66%	66%
		64 Ms. (HN)	197	197	197	66%	66%	66%
		64 Ms. (BL)	197	197	201	66%	66%	67%

Table 6. Comparison analysis in TDNN with LMA, BRA, and SCGA for Bangla phoneme recognition

Eight Unique Phonemes	Feature Extraction Methods	Percentage of Recognition (Range, Mean Accuracy) TDNN& LMA	Percentage of Recognition (Range, Mean Accuracy) TDNN& BRA	Percentage of Recognition (Range, Mean Accuracy) TDNN& SCGA
Twelve male-female participants	FFT	60%-67%, 63%	60%-67%, 62%	60%-67%, 62%
	LPC	56%-71%, 62%	56%-71%, 62%	56%-71%, 62%
	MFCC	80%-89%, 86%	80%-89%, 86%,	80%-89%, 86%

Table 7. Comparison analysis in TDNN with LMA, BRA, and SCGA for Bangla word recognition

Eight Unique Words	Feature Extraction Methods	Percentage of Recognition (Range, Mean Accuracy) TDNN& LMA	Percentage of Recognition (Range, Mean Accuracy) TDNN& BRA	Percentage of Recognition (Range, Mean Accuracy) TDNN& SCGA
Ten male-female participants	FFT	45%–60%, 53%	51%–60%, 54%	50%–60%, 54%
	LPC	48%–53%, 53%	48%–53%, 52%	53%–54%, 54%
	MFCC	76%–94%, 88%	76%–94%, 87%	76%–94%, 91%

Table 8. Comparison analysis in TDNN with LMA, BRA, and SCGA for Bangla command recognition

Eight Unique Commands	Feature Extraction Methods	Percentage of Recognition (Range, Mean Accuracy) TDNN & LMA	Percentage of Recognition (Range, Mean Accuracy) TDNN & BRA	Percentage of Recognition (range, Mean Accuracy) TDNN & SCGA
Ten male-female participants	FFT	17%–33%, 26%	17%–33%, 25%	17% –33%, 26%
	LPC	21%–45%, 27%	21%–45%, 27%	21%–45%, 27%
	MFCC	57%–80%, 71%	57%–80%, 71%	57%– 80%, 71%

Table 9. Comparison analysis in TDNN with LMA, BRA, and SCGA for Bangla sentence recognition

Six Unique Sentences	Feature Extraction Methods	Percentage of Recognition (Range, Mean Accuracy) TDNN & LMA	Percentage of Recognition (Range, Mean Accuracy) TDNN & BRA	Percentage of Recognition (Range, Mean Accuracy) TDNN & SCGA
Ten male-female participants	FFT	44%–57%, 49%	44%–57%, 48%	47%–60%, 51%
	LPC	44%–49%, 48%	45%–50%, 48%	49%–50%, 49%
	MFCC	66%–94%, 77%	66%–93%, 77%	66%–94%, 78%

Table 5 details the feature extraction of Bangla sentences using FFT, LPC, and MFCC, followed by their recognition using a TDNN with three algorithms: LMA, BRA, and SCGA.

Table 6 focuses on Bangla phoneme feature extraction using FFT, LPC, and MFCC, and recognition in TDNN with three algorithms- LMA, BRA, and SCGA.

Table 7 presents the results of Bangla word feature extraction using FFT, LPC, and MFCC, followed by recognition using a TDNN model with three algorithms- LMA, BRA, and SCGA.

Table 8 presents the results of Bangla command feature extraction using FFT, LPC, and MFCC, followed by recognition using a TDNN model with three algorithms - LMA, BRA, and SCGA.

Table 9 details the feature extraction of Bangla sentences using FFT, LPC, and MFCC, and their recognition using a TDNN model with three algorithms-LMA, BRA, and SCGA.

11.2 Summary (Speech recognition)

TDNN models are trained using three optimization algorithms: SCGA, LMA, and BRA. Results indicate that MFCC combined with TDNN optimized via LMA, BRA, or SCGA achieves the highest recognition accuracy across multiple tasks: phoneme recognition (89%), word recognition (94%), command recognition (80%), and sentence recognition (94%), as detailed in Tables 2 to 9. As a feature extraction method, MFCC outperforms LPC and FFT by effectively modeling human auditory perception through the Mel scale, which emphasizes low-frequency speech components. Unlike FFT's raw spectral output, MFCC applies a Discrete Cosine Transform to produce compact and decorrelated features, enhancing phoneme discrimination. LPC, while efficient for vocal tract modeling, is more sensitive to noise and less effective in capturing the dynamic characteristics of natural speech. Due to its noise robustness and perceptually relevant features, MFCC is considered ideal for automatic speech recognition.

12. SYSTEM'S PERFORMANCE EVALUATION

Bangla phonemes, isolated words, commands, and sentences were recognized using three parallel feature extraction methods: FFT, LPC, and MFCC. These techniques capture complementary spectral and temporal aspects of speech. Recognition was performed with a TDNN, trained using LMA, BRA, and SCGA algorithms. The dataset included up to 480 samples from 12 male and female speakers, ensuring vocal diversity. Framing used 20-ms and 64-ms windows with HM, HN, and BL functions to balance time-frequency resolution and reduce spectral leakage. Comprehensive testing across all speech categories enabled detailed evaluation of recognition accuracy and the effectiveness of different feature extraction and training configurations.

The system's performance for Bangla speech recognition was thoroughly evaluated using MATLAB, applying diverse metrics to assess phoneme, word, command, and sentence-level accuracy. Feature extraction preceded machine learning processes, with results summarized in Table 10 and visualized in Tables 11 to 22 and Figures 4 to 11. Evaluation metrics, including *Best Validation Performance*, *Error Histogram*, *Regression Analysis*, *Time-Series Response, Error Autocorrelation, and Input-Error Cross-Correlation,* ensured robustness, generalizability, and bias reduction across various prediction scenarios. These evaluation metrics also ensure the developed system model is truly potential.

Table 10. System's performance evaluation

06 to 08 Phonemes, Words, Commands, Sentences	FEM	WL in HM, HN, BL	*PE (E)	*TST (G, E)	*ER_H	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 to 12 Male-female (uttered 300 to 480 times)	FFT, LPC and MFCC	20 & 64 Ms. of HM, HN, BL	Values range from 0.000207 to 0.13876, Ranges from E6 to E171	Values range from 0.000101 to 0.023711, Ranges from E12 to E144	Values range from 0.00123 to 0.1992, B20	Values range from 0.2123 to 0.89089	Ranges from -0.0101 to -0.6908	Values range from 0.01098 to 0.909	Values range from -0.00032 to -0.2907

Table 11. Performance evaluation of 08 unique Bangla phonemes in LMA

08 Phonemes	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
12 Male-female (uttered 480 times)	FFT	20 Ms.	HM	0.070235, E70	0.000207, E76	0.03663	0.60252	-0.3534	0.07957	-0.00253
			HN	0.070245, E81	0.000208, E59	0.03653	0.60262	-0.3537	0.07962	-0.00254
			BL	0.070241, E68	0.000207, E81	0.03671	0.60248	-0.3529	0.07959	-0.00253
		64 Ms.	HM	0.073716, E60	0.000303, E66	0.109	0.64006	-0.4065	0.07251	-0.00098
			HN	0.074717, E44	0.000303, E68	0.101	0.65007	-0.4062	0.07262	-0.00098
			BL	0.074719, E71	0.000303, E87	0.111	0.63009	-0.4069	0.07259	-0.00098
	LPC	20 Ms.	HM	0.071633, E127	0.001363, E133	0.06614	0.59662	-0.3873	0.06529	-0.00728
			HN	0.071531, E171	0.001263, E109	0.06711	0.59697	-0.3870	0.06530	-0.00727
			BL	0.061732, E111	0.001369, E121	0.07612	0.59761	-0.3971	0.06532	-0.00741
		64 Ms.	HM	0.064152, E75	0.002206, E81	0.00986	0.63081	-0.3562	0.05182	-0.01795
			HN	0.064161, E57	0.002216, E99	0.00987	0.64077	-0.3563	0.05283	-0.01796
			BL	0.064148, E79	0.002217, E77	0.00987	0.63079	-0.3570	0.05179	-0.01797

	MFCC	20 Ms.	HM	0.050343, E22	0.002201, E28	0.03174	0.73467	-0.4075	0.03772	-0.03812
			HN	0.050339, E31	0.002200, E19	0.03149	0.73479	-0.4059	0.03769	-0.03821
			BL	0.050341, E19	0.002201, E24	0.03181	0.73471	-0.4081	0.03770	-0.03809
		64 Ms.	HM	0.044984, E20	0.008354, E26	0.03309	0.79505	-0.4295	0.0178	-0.1737
			HN	0.044881, E31	0.008362, E33	0.03401	0.79499	-0.4287	0.0179	-0.1741
			BL	0.044901, E22	0.008370, E24	0.03299	0.80103	-0.4301	0.0181	-0.1743

Table 12. Performance evaluation of 08 unique Bangla phonemes in BRA

08 Phonemes	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
12 Male-female (uttered 480 times)	FFT	20 Ms.	HM	0.060333, E90	0.000199, E66	0.03336	0.50434	-0.3636	0.07666	-0.00443
			HN	0.060222, E72	0.000146, E87	0.03767	0.60545	-0.3838	0.07768	-0.00565
			BL	0.060111, E66	0.000125, E56	0.03773	0.60545	-0.3926	0.07879	-0.00675
		64 Ms.	HM	0.063212, E69	0.000326, E76	0.101	0.54098	-0.3164	0.07675	-0.00199
			HN	0.064432, E88	0.000235, E45	0.106	0.65989	-0.4161	0.07213	-0.00897
			BL	0.064123, E71	0.000333, E78	0.109	0.63787	-0.4169	0.06768	-0.00565
	LPC	20 Ms.	HM	0.061543, E99	0.001764, E77	0.06554	0.69565	-0.3774	0.06815	-0.00444
			HN	0.061876, E32	0.001557, E99	0.05743	0.59232	-0.4771	0.06806	-0.00878
			BL	0.051767, E88	0.001448, E109	0.07987	0.59343	-0.3872	0.07801	-0.00568
		64 Ms.	HM	0.054343, E66	0.001223, E91	0.00123	0.63676	-0.3764	0.05901	-0.01908
			HN	0.054232, E55	0.002551, E99	0.00545	0.74066	-0.3665	0.05794	-0.01658
			BL	0.054555, E77	0.001333, E55	0.00765	0.63034	-0.3771	0.05198	-0.01272
	MFCC	20 Ms.	HM	0.040232, E32	0.002569, E66	0.03742	0.73323	-0.4276	0.03676	-0.03292
			HN	0.040878, E32	0.002889, E55	0.02135	0.73878	-0.4158	0.03908	-0.03303
			BL	0.040454, E77	0.002657, E22	0.03647	0.83232	-0.4189	0.03765	-0.03594
		64 Ms.	HM	0.034878, E69	0.008656, E43	0.03555	0.79989	-0.4497	0.0155	-0.1755
			HN	0.034090, E23	0.008451, E34	0.03912	0.89089	-0.4388	0.0198	-0.1722
			BL	0.034098, E34	0.007331, E29	0.02242	0.80087	-0.4606	0.0133	-0.1755

Table 13. Performance evaluation of 08 unique phonemes in SCGA

08 Phonemes	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
12 Male-female (uttered 480 times)	FFT	20 Ms.	HM	0.070666, E55	0.000101, E55	0.03545	0.60464	-0.3211	0.07546	-0.00232
			HN	0.070373, E65	0.000109, E76	0.03656	0.60232	-0.3232	0.07876	-0.00255
			BL	0.070242, E67	0.000301, E44	0.03333	0.60876	-0.3432	0.07897	-0.00266
		64 Ms.	HM	0.073323, E77	0.000299, E89	0.121	0.64909	-0.4123	0.07134	-0.00099
			HN	0.074345, E88	0.000297, E45	0.109	0.65101	-0.4656	0.07135	-0.00099
			BL	0.074898, E90	0.000264, E66	0.131	0.63102	-0.4414	0.07123	-0.00077
	LPC	20 Ms.	HM	0.071565, E111	0.001321, E99	0.06234	0.59332	-0.3242	0.06432	-0.00708
			HN	0.071383, E132	0.001301, E101	0.06432	0.59786	-0.3363	0.06231	-0.00766
			BL	0.061898, E109	0.001299, E144	0.07876	0.59098	-0.3353	0.06909	-0.00032
		64 Ms.	HM	0.064223, E55	0.002198, E98	0.00908	0.63908	-0.3765	0.05126	-0.01755
			HN	0.064665, E76	0.002251, E41	0.00864	0.64801	-0.3876	0.05808	-0.01776
			BL	0.064998, E99	0.002199, E55	0.00807	0.63011	-0.3131	0.05292	-0.01087
	MFCC	20 Ms.	HM	0.050123, E44	0.002176, E37	0.03202	0.73356	-0.4011	0.03545	-0.03812
			HN	0.050323, E47	0.002170, E33	0.03305	0.73786	-0.4212	0.03981	-0.03865
			BL	0.050111, E32	0.002302, E21	0.03111	0.73242	-0.4111	0.03887	-0.03078
		64 Ms.	HM	0.044343, E33	0.008222, E20	0.03209	0.79575	-0.4212	0.0179	-0.1722
			HN	0.044657, E54	0.008234, E36	0.03301	0.79897	-0.4232	0.0166	-0.1754
			BL	0.044876, E23	0.008432, E41	0.03232	0.80099	-0.4122	0.0199	-0.1722

Table 14. Performance evaluation of 08 unique Bangla words in LMA

08 Words	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 times)	FFT	20 Ms.	HM	0.086621, E18	0.010268, E24	0.01845	0.47285	-0.3557	0.08088	-0.00562
			HN	0.086599, E22	0.010291, E33	0.01891	0.47333	-0.3498	0.08099	-0.00511
			BL	0.086722, E31	0.010302, E19	0.01901	0.47307	-0.3571	0.08102	-0.00498
		64 Ms.	HM	0.080386, E30	0.0048649, E36	0.09273	0.5755	-0.1686	0.05305	-0.00308
			HN	0.080401, E29	0.0048878, E41	0.09301	0.5777	-0.1866	0.05298	-0.00341
			BL	0.080368, E38	0.0048964, E28	0.09298	0.5801	-0.1801	0.05503	-0.00401
	LPC	20 Ms.	HM	0.08664, E62	0.0036627, E68	0.0663	0.46911	-0.07897	0.08269	-0.00591
			HN	0.08709, E87	0.0036762, E86	0.0697	0.46899	-0.07799	0.08302	-0.00583
			BL	0.08699, E71	0.0036596, E74	0.0701	0.46972	-0.07840	0.08298	-0.00601
		64 Ms.	HM	0.086909, E31	0.0011545, E37	0.09459	0.45675	-0.1054	0.09343	-0.01109
			HN	0.087001, E42	0.0011545, E55	0.09503	0.45713	-0.1076	0.09376	-0.01207

			BL	0.087040, E39	0.0011545, E49	0.09498	0.45702	-0.1081	0.09401	-0.01188
	MFCC	20 Ms.	HM	0.040485, E41	0.0015751, E47	0.01708	0.32425	-0.176	0.02388	-0.0131
			HN	0.040511, E61	0.0015788, E55	0.01801	0.32499	-0.189	0.02416	-0.0155
			BL	0.040522, E55	0.0015810, E41	0.01798	0.32476	-0.191	0.02404	-0.0161
		64 Ms.	HM	0.069587, E14	0.023657, E20	0.03791	0.68829	-0.5727	0.01505	-0.1259
			HN	0.069607, E21	0.023677, E33	0.03809	0.68888	-0.5802	0.01599	-0.1307
			BL	0.069689, E34	0.023711, E27	0.03833	0.68912	-0.5843	0.01609	-0.1345

Table 15. Performance evaluation of 08 unique Bangla words in BRA

08 Words	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 times)	FFT	20 Ms.	HM	0.086435, E32	0.010245, E34	0.01987	0.47765	-0.3445	0.08121	-0.00321
			HN	0.086876, E43	0.010321, E65	0.01786	0.47565	-0.3876	0.08334	-0.00675
			BL	0.086908, E44	0.010343, E87	0.01242	0.47909	-0.3796	0.08654	-0.00987
		64 Ms.	HM	0.080786, E33	0.0048675, E24	0.09987	0.5787	-0.1808	0.05678	-0.00654
			HN	0.080654, E45	0.0048675, E76	0.09898	0.5776	-0.1704	0.05909	-0.00234
			BL	0.080343, E55	0.0048923, E34	0.09909	0.5987	-0.1342	0.05464	-0.00876
	LPC	20 Ms.	HM	0.08676, E71	0.0036232, E98	0.0747	0.46898	-0.07346	0.08876	-0.00909
			HN	0.08897, E45	0.0036454, E33	0.0565	0.46242	-0.07765	0.08098	-0.00554
			BL	0.08709, E65	0.0036575, E65	0.0801	0.46786	-0.07876	0.08786	-0.00786
		64 Ms.	HM	0.086879, E22	0.0011987, E77	0.09987	0.45908	-0.1088	0.09033	-0.01121
			HN	0.087897, E55	0.0011345, E89	0.09565	0.45435	-0.1577	0.09786	-0.01199
			BL	0.087909, E37	0.0011987, E23	0.09231	0.45454	-0.1199	0.09091	-0.01211
	MFCC	20 Ms.	HM	0.040675, E61	0.0015123, E56	0.01987	0.32876	-0.174	0.02546	-0.0198
			HN	0.040453, E34	0.0015543, E87	0.01987	0.32234	-0.187	0.02432	-0.0177
			BL	0.040654, E67	0.0015383, E33	0.01897	0.32685	-0.195	0.02342	-0.0123
		64 Ms.	HM	0.069432, E19	0.023564, E23	0.03876	0.68123	-0.5833	0.01876	-0.1291
			HN	0.069897, E22	0.023987, E87	0.03897	0.68876	-0.5711	0.01987	-0.1308
			BL	0.069435, E29	0.023343, E33	0.03998	0.68843	-0.5734	0.01922	-0.1312

Table 16. Performance evaluation of 08 unique Bangla words in SCGA

08 Words	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 times)	FFT	20 Ms.	HM	0.086543, E22	0.010268, E27	0.01091	0.47199	-0.3443	0.08599	-0.00432
			HN	0.086987, E45	0.010291, E40	0.01664	0.47231	-0.3123	0.08721	-0.00511
			BL	0.086654, E76	0.010302, E33	0.01665	0.47421	-0.3765	0.08345	-0.00765
		64 Ms.	HM	0.080098, E89	0.0048649, E53	0.09897	0.5821	-0.1876	0.05654	-0.00876
			HN	0.080675, E23	0.0048878, E67	0.09901	0.5342	-0.1098	0.05321	-0.00987
			BL	0.080091, E76	0.0048964, E33	0.09665	0.5543	-0.1701	0.05099	-0.00432
	LPC	20 Ms.	HM	0.08876, E44	0.0036627, E88	0.0711	0.46871	-0.07554	0.08543	-0.00865
			HN	0.08321, E67	0.0036762, E90	0.0737	0.46098	-0.07443	0.08984	-0.00123
			BL	0.08785, E87	0.0036596, E65	0.0799	0.46803	-0.07905	0.08569	-0.00432
		64 Ms.	HM	0.086443, E44	0.0011545, E53	0.09765	0.45765	-0.1044	0.09776	-0.01876
			HN	0.087341, E67	0.0011545, E41	0.09561	0.45453	-0.1011	0.09987	-0.01987
			BL	0.087908, E34	0.0011545, E78	0.09098	0.45771	-0.1034	0.09388	-0.01122
	MFCC	20 Ms.	HM	0.040342, E76	0.0015751, E90	0.01903	0.32061	-0.183	0.02098	-0.0234
			HN	0.040761, E25	0.0015788, E71	0.01788	0.32841	-0.191	0.02765	-0.0321
			BL	0.040896, E55	0.0015810, E40	0.01544	0.32931	-0.184	0.02321	-0.0291
		64 Ms.	HM	0.069651, E21	0.023657, E39	0.03821	0.68061	-0.5345	0.01098	-0.1321
			HN	0.069906, E37	0.023677, E77	0.03554	0.68906	-0.5765	0.01678	-0.1431
			BL	0.069333, E43	0.023711, E64	0.03519	0.68456	-0.5908	0.01509	-0.1213

Table 17. Performance evaluation of 08 unique Bangla commands in LMA

08 Commands	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 Times)	FFT	20 Ms.	HM	0.10324, E99	0.000211, E105	0.03812	0.25437	-0.6741	0.09594	-0.00238
			HN	0.10433, E101	0.000212, E117	0.03910	0.25501	-0.6801	0.09587	-0.00240
			BL	0.10401, E89	0.000211, E98	0.03799	0.25510	-0.6811	0.09641	-0.00243
		64 Ms.	HM	0.10094, E17	0.012434, E23	0.06912	0.30106	-0.3116	0.08114	-0.00568
			HN	0.10097, E22	0.012434, E31	0.06901	0.30299	-0.3210	0.08188	-0.00575
			BL	0.10099, E29	0.012434, E38	0.07003	0.30303	-0.3302	0.08299	-0.00581
	LPC	20 Ms.	HM	0.10336, E62	0.000435, E68	0.08516	0.27942	-0.2578	0.09031	-0.02758
			HN	0.10512, E49	0.000436, E77	0.08613	0.27998	-0.2581	0.09109	-0.02864
			BL	0.103444, E54	0.000437, E81	0.08598	0.27897	-0.2531	0.09210	-0.02821
		64 Ms.	HM	**0.10527, E6**	0.001789, E12	0.03975	0.23414	-0.09793	0.0919	-0.00039
			HN	0.10577, E9	0.001789, E19	0.04110	0.23499	-0.09854	0.0997	-0.00041

			BL	0.10601, E11	0.001789, E22	0.03999	0.23501	-0.09833	0.0981	-0.00038
	MFCC	20 Ms.	HM	0.093885, E29	0.004014, E35	0.1072	0.39192	-0.1656	0.08553	-0.1531
			HN	0.093899, E21	0.004019, E48	0.1219	0.39321	-0.1665	0.08599	-0.1569
			BL	0.093902, E44	0.004021, E51	0.1171	0.39210	-0.1671	0.08609	-0.1610
		64 Ms.	HM	0.094032, E18	0.002580, E24	0.00223	0.4632	-0.182	0.07838	-0.2026
			HN	0.094106, E27	0.002580, E29	0.00233	0.4710	-0.199	0.07919	-0.2222
			BL	0.094210, E24	0.002580, E33	0.00231	0.4555	-0.189	0.07899	-0.2323

Table 18. Performance evaluation of 08 unique Bangla commands in BRA

08 Commands	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 times)	FFT	20 Ms.	HM	0.10554, E23	0.000989, E99	0.03123	0.25223	-0.6876	0.09543	-0.0432
			HN	0.10876, E87	0.000199, E88	0.03324	0.25432	-0.6876	0.09123	-0.0123
			BL	0.10887, E65	0.000234, E55	0.03543	0.25654	-0.6901	0.09654	-0.0098
		64 Ms.	HM	0.10098, E22	0.012654, E66	0.06654	0.30765	-0.3321	0.08765	-0.0611
			HN	0.10123, E65	0.012234, E41	0.06765	0.30876	-0.3543	0.08876	-0.0676
			BL	0.10765, E67	0.012654, E29	0.07987	0.30123	-0.3343	0.08098	-0.0721
	LPC	20 Ms.	HM	0.10098, E87	0.000808, E55	0.08098	0.27765	-0.2765	0.09789	-0.2821
			HN	0.10765, E44	0.000776, E83	0.08231	0.27876	-0.2876	0.09098	-0.2799
			BL	0.10098, E87	0.000543, E76	0.08543	0.27098	-0.2564	0.09368	-0.0291
		64 Ms.	HM	0.10001, E64	0.001876, E33	0.03654	0.23765	-0.0678	0.0976	-0.0099
			HN	0.10987, E21	0.001098, E41	0.04432	0.23123	-0.0578	0.0966	-0.0055
			BL	0.10554, E32	0.001801, E53	0.03211	0.23654	-0.0876	0.0992	-0.0054
	MFCC	20 Ms.	HM	0.09098, E76	0.004135, E76	0.1083	0.39876	-0.1445	0.08543	-0.1678
			HN	0.09512, E99	0.004432, E33	0.1244	0.39986	-0.1872	0.08765	-0.1987
			BL	0.09704, E45	0.004861, E80	0.1181	0.39776	-0.1112	0.08123	-0.1567
		64 Ms.	HM	0.09665, E34	0.002071, E43	0.00234	0.4665	-0.191	0.07543	-0.2111
			HN	0.09234, E21	0.002082, E44	0.00951	0.4876	-0.181	0.07654	-0.2231
			BL	0.09876, E33	0.002022, E45	0.00781	0.4532	-0.199	0.07123	-0.2251

Table 19. Performance evaluation of 08 unique Bangla commands in SCGA

08 Commands	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 400 times)	FFT	20 Ms.	HM	0.10321, E87	0.000199, E99	0.0391	0.25543	-0.6666	0.09711	-0.0021
			HN	0.10123, E99	0.000404, E101	0.0379	0.25654	-0.6786	0.09765	-0.0199
			BL	0.10401, E77	0.000389, E76	0.0368	0.25987	-0.6908	0.09908	-0.0311
		64 Ms.	HM	0.10087, E43	0.012453, E33	0.0631	0.30123	-0.3131	0.08231	-0.0642
			HN	0.10123, E34	0.012681, E40	0.0579	0.30909	-0.3654	0.08388	-0.0755
			BL	0.10103, E44	0.012539, E37	0.0711	0.30432	-0.3101	0.08397	-0.0801
	LPC	20 Ms.	HM	0.10191, E55	0.000297, E47	0.0841	0.27876	-0.2675	0.09123	-0.2677
			HN	0.10312, E33	0.000651, E60	0.0759	0.27907	-0.2987	0.09244	-0.2791
			BL	0.10432, E66	0.000643, E79	0.0847	0.27309	-0.2543	0.09432	-0.2907
		64 Ms.	HM	0.10101, E11	0.001695, E19	0.0375	0.23579	-0.0101	0.0811	-0.0101
			HN	0.10721, E13	0.001839, E21	0.0411	0.23701	-0.0811	0.0877	-0.0109
			BL	0.10333, E33	0.001794, E31	0.0338	0.23404	-0.0809	0.0845	-0.0099
	MFCC	20 Ms.	HM	0.09432, E12	0.004052, E41	0.1099	0.39169	-0.1755	0.0788	-0.1601
			HN	0.09123, E22	0.004901, E55	0.1301	0.39654	-0.1566	0.01234	-0.1499
			BL	0.09751, E39	0.004001, E61	0.1233	0.39654	-0.1754	0.07888	-0.1597
		64 Ms.	HM	0.09581, E22	0.002391, E39	0.0909	0.4579	-0.191	0.06779	-0.2078
			HN	0.09329, E29	0.002431, E20	0.0676	0.4681	-0.189	0.07876	-0.2255
			BL	0.09641, E31	0.002402, E21	0.0101	0.4474	-0.198	0.07567	-0.2299

Table 20. Performance evaluation of 06 unique Bangla sentences in LMA

06 Sentences	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 300 times)	FFT	20 Ms.	HM	0.13097, E19	0.003829, E25	0.04715	0.24434	-0.1867	0.107	-0.00551
			HN	0.13210, E22	0.003833, E33	0.04811	0.24501	-0.1888	0.111	-0.00777
			BL	0.13110, E32	0.003841, E12	0.04810	0.24522	-0.1967	0.119	-0.04610
		64 Ms.	HM	0.12807, E11	0.009811, E17	0.02406	0.29585	-0.3211	0.08873	-0.02248
			HN	0.12708, E17	0.009821, E24	0.02532	0.29610	-0.3279	0.08699	-0.02332
			BL	0.12699, E19	0.009818, E31	0.02499	0.29609	-0.3301	0.08783	-0.02222
	LPC	20 Ms.	HM	0.13048, E57	0.000991, E63	0.02345	0.29162	-0.2259	0.1117	-0.00172
			HN	0.13109, E66	0.000989, E44	0.02434	0.29244	-0.2121	0.1201	-0.00179
			BL	0.13101, E75	0.000999, E51	0.02343	0.30009	-0.2212	0.1199	-0.00180
		64	HM	0.13153, E18	0.008669, E24	0.143	0.30985	-0.1951	0.09529	-0.00193

		Ms.	HN	0.13333, E21	0.008671, E41	0.166	0.31011	-0.1999	0.09611	-0.00199
			BL	0.13210, E33	0.008677, E33	0.159	0.31089	-0.2001	0.09677	-0.00201
	MFCC	20 Ms.	HM	0.11883, E32	0.014634, E38	0.008258	0.39277	-0.1716	0.05323	-0.1216
			HN	0.11901, E44	0.014796, E52	0.008302	0.39298	-0.1787	0.05555	-0.1287
			BL	0.11934, E39	0.014899, E45	0.008333	0.39300	-0.1809	0.05433	-0.1333
		64 Ms.	HM	0.12071, E12	0.007361, E18	0.08902	0.42735	-0.353	0.07355	-0.1135
			HN	0.12112, E19	0.007370, E22	0.08911	0.42811	-0.360	0.07401	-0.1231
			BL	0.12211, E32	0.007377, E31	0.08999	0.42833	-0.369	0.07414	-0.1210

Table 21. Performance evaluation of 06 unique Bangla sentences in BRA

06 Sentences	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 300 times)	FFT	20 Ms.	HM	0.13871, E21	0.003654, E23	0.04654	0.24321	-0.1871	0.134	-0.0577
			HN	0.13432, E31	0.003123, E41	0.04159	0.24123	-0.1983	0.321	-0.0907
			BL	0.13198, E44	0.003987, E18	0.04953	0.24345	-0.1786	0.432	-0.0491
		64 Ms.	HM	0.12158, E19	0.009099, E41	0.02598	0.29543	-0.3321	0.101	-0.0096
			HN	0.12321, E34	0.009631, E32	0.02543	0.29567	-0.3909	0.108	-0.0505
			BL	0.12571, E31	0.009891, E66	0.02567	0.29654	-0.3542	0.091	-0.0065
	LPC	20 Ms.	HM	0.13129, E44	0.000077, E78	0.02879	0.29567	-0.2123	0.909	-0.0981
			HN	0.13941, E77	0.000546, E90	0.02231	0.29765	-0.2321	0.111	-0.0099
			BL	0.13876, E71	0.000564, E49	0.02672	0.30086	-0.2432	0.121	-0.0011
		64 Ms.	HM	0.13123, E29	0.008733, E33	0.1543	0.30895	-0.1876	0.229	-0.0192
			HN	0.13231, E37	0.008598, E43	0.1673	0.31243	-0.1955	0.078	-0.0019
			BL	0.13321, E54	0.008436, E29	0.1987	0.31341	-0.2565	0.076	-0.0021
	MFCC	20 Ms.	HM	0.11902, E49	0.014541, E78	0.00654	0.39987	-0.1654	0.088	-0.1216
			HN	0.11877, E30	0.014981, E66	0.00765	0.39654	-0.1899	0.077	-0.1287
			BL	0.11899, E61	0.014596, E54	0.00652	0.39982	-0.165	0.012	-0.1333
		64 Ms.	HM	0.12099, E17	0.007591, E19	0.08879	0.42908	-0.766	0.099	-0.1135
			HN	0.12101, E23	0.007876, E17	0.08543	0.42432	-0.299	0.087	-0.1231
			BL	0.12109, E20	0.007763, E27	0.08549	0.42234	-0.298	0.044	-0.1210

Table 22. Performance evaluation of 06 unique Bangla sentences in SCGA

06 Sentences	FEM	WL in HM, HN, BL		*PE (E)	*TST (G, E)	*ER_H (Max Bins = 20)	**R_A (R)	**TSR (Er)	*E_AC	*IE_CC (Er)
10 Male-female (uttered 300 times)	FFT	20 Ms.	HM	0.13921, E91	0.00665, E67	0.0543	0.2544	-0.1911	0.143	-0.0876
			HN	0.13321, E34	0.00267, E34	0.0356	0.2123	-0.1777	0.145	-0.0543
			BL	0.13123, E77	0.00776, E18	0.0432	0.2987	-0.1763	0.123	-0.0559
		64 Ms.	HM	0.12234, E61	0.00877, E65	0.0321	0.2456	-0.3741	0.0954	-0.0909
			HN	0.12432, E34	0.00866, E44	0.0533	0.2532	-0.3123	0.0853	-0.0776
			BL	0.12345, E35	0.00799, E76	0.0301	0.2876	-0.3234	0.0966	-0.0432
	LPC	20 Ms.	HM	0.13543, E76	0.00123, E98	0.0401	0.2766	-0.2432	0.1255	-0.0123
			HN	0.13456, E99	0.00321, E99	0.0499	0.2455	-0.2543	0.1987	-0.0322
			BL	0.13654, E55	0.00145, E44	0.0232	0.3089	-0.2345	0.1234	-0.0123
		64 Ms.	HM	0.13567, E39	0.00766, E65	0.1992	0.3977	-0.1654	0.0765	-0.0231
			HN	0.13654, E54	0.00909, E88	0.177	0.3433	-0.1987	0.0856	-0.1432
			BL	0.13765, E66	0.00689, E54	0.1671	0.3544	-0.2709	0.0999	-0.2329
	MFCC	20 Ms.	HM	0.11567, E64	0.01566, E19	0.0988	0.3766	-0.1368	0.0597	-0.1432
			HN	0.11987, E22	0.01654, E39	0.00345	0.3833	-0.1962	0.0587	-0.1364
			BL	0.11876, E18	0.01721, E54	0.0907	0.3799	-0.1908	0.0654	-0.1973
		64 Ms.	HM	0.12088, E24	0.00688, E21	0.0766	0.4453	-0.299	0.0543	-0.1176
			HN	0.12123, E11	0.00861, E28	0.08743	0.4432	-0.101	0.0432	-0.1188
			BL	0.12432, E24	0.00639, E18	0.08409	0.4322	-0.011	0.7234	-0.1199

Best Validation Performance tracks validation error to prevent overfitting and stops training optimally, while ***Error Histogram*** visualizes prediction error distribution to detect biases and ensure balanced generalization. ***Validation Checks*** stop training when validation error increases to prevent overfitting and confirm model performance, while ***Regression Analysis (R)*** evaluates correlation between predicted and actual values, ensuring strong predictive ability and generalization. ***Time-Series Response*** evaluates the model's adaptability to sequential data trends, **while** ***Error Autocorrelation*** ensures error independence to prevent bias and enhance robustness in forecasting tasks. ***Input-Error Cross-Correlation*** evaluates the extent to which input variables influence errors, with high correlation signaling potential bias and minimal correlation ensuring fair, generalizable predictions across diverse input conditions.

Tables 10-22, where '*' ***One asterisk*** denotes the ***Mean Squared Error*** (**MSE**), which is the average squared difference between outputs and targets. Lower values are better, with ***zero*** indicating no error. '**' ***Two asterisks*** denote Regression ***R*** Values, which measure the correlation between outputs and targets. An ***R*** value of 1 indicates a close relationship, while an ***R*** value of 0 indicates a random relationship.

The factors (Table 10) considered for evaluating the performance of the developed system are as follows:

- ❖ Feature Extraction Methods (FEM): FFT, LPC and MFCC
- ❖ Window length or WL (in milliseconds) Hamming = HM, Hanning = HN, Blackman = BL
- ❖ *Performance Evaluation with Epoch (E) = PE (E)
- ❖ *Training State (Gradient, Epoch = E) = TST (G, E)
- ❖ *Error Histogram (Max Bins = 20) = ER_H
- ❖ **Regression Analysis (R) = R_A (R)
- ❖ **Time Series Response Error (R) = TSR(Er)
- ❖ *Error Autocorrelation (Correlation) EA = E_AC
- ❖ *Input-Error Cross-correlation (Error) = IE_CC (Er)

Table 10 is essentially a summary of Tables 11 to 22, which present performance evaluations across various Bangla linguistic units using three different methods: LMA, BRA, and SCGA. Specifically, Tables 11-13 evaluate eight distinct Bangla phonemes, Tables 14-16 assess eight distinct Bangla words, Tables 17-19 focus on eight distinct Bangla commands, and Tables 20-22 examine six distinct Bangla sentences- all using LMA, BRA, and SCGA, respectively.

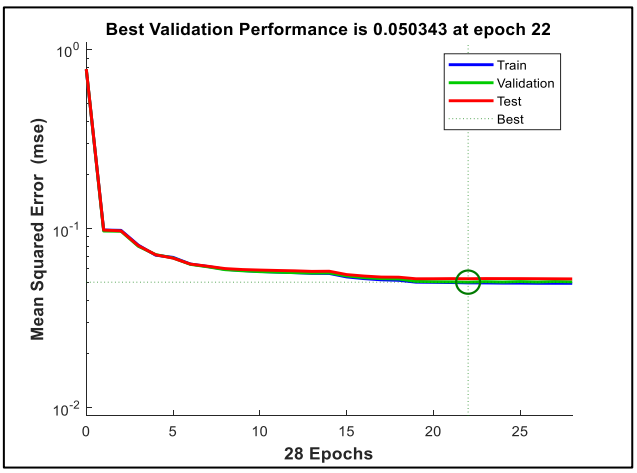

Figure 4. Best validation performance (MFCC & LMA algorithm)

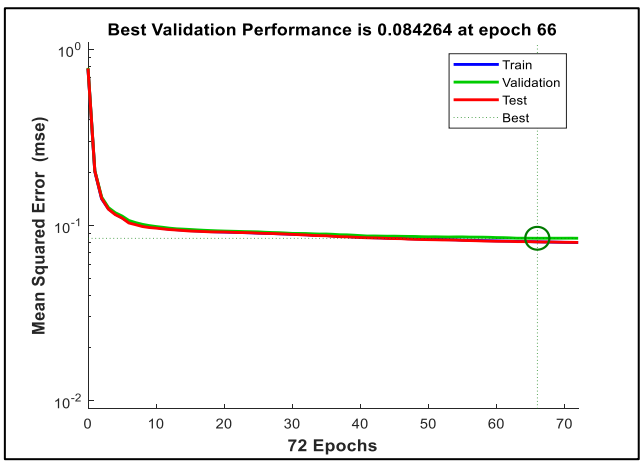

Figure 5. Best validation performance (MFCC & SCGA algorithm)

12.1 Best validation performance

Figures 4 and 5 graphically present training and validation results. Figure 4 represents the best validation performance, TDNN training model for Bangla Phoneme with MFCC & LMA algorithm. Figure 5 represents the best validation performance, TDNN training model for Bangla Word with MFCC & SCGA algorithm. Achieving best validation performances with near-zero mean squared error rates demonstrates the system's accuracy, efficiency, and robustness, achieved within just a few epochs. In Table 10, the **Performance Evaluation with Epoch (E) or PE (E)* values range from 0.000207 to 0.13876, while the corresponding epochs span from *E6 to E171,* as observed across all experiments detailed in Tables 11 to 22.

12.2 TDNN network model training (Validation checks)

Achieving gradient points close to zero with only a few epochs during "validation checks" indicates that Time Delay neural network (TDNN) is highly effective and well-optimized and that is observed in Figure 6. In Table 10, the Training State (Gradient, Epoch = E) or TST (G, E) values range from 0.000101 to 0.023711, with epochs spanning from E12 to E144, based on all experiments presented in Tables 11 to 22.

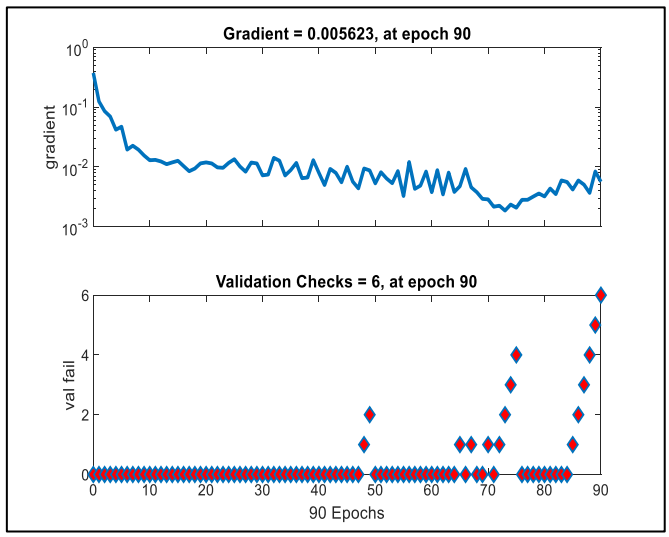

Figure 6. TDNN model training (Bangla Sentence in MFCC & SCGA algorithm)

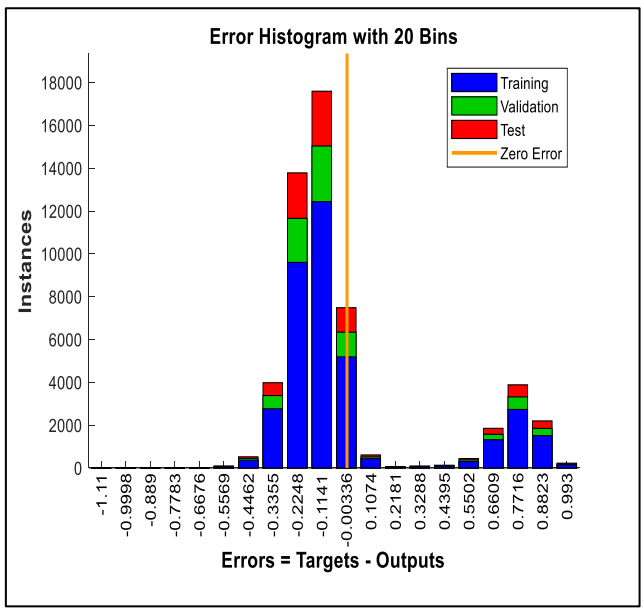

Figure 7. TDNN model training for Error Histogram (Bangla Sentence in MFCC& SCGA algorithm)

12.3 Error Histogram

Figure 7 is a 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. In Table 10, the *Error Histogram (ER_H) with a maximum of 20 bins shows values ranging from 0.00123 to 0.1992, based on all experiments detailed in Tables 11 to 22.

12.4 Regression analysis

Figure 8 illustrates the system's speech recognition performance through ***Regression Analysis***. The ***R*** value measures ***correlation*** between outputs and targets, with 1 indicating a strong relationship and 0 signifying randomness. A value close to 1 highlights the model's accuracy and reliability in predicting true values. In Table 10, the **Regression Analysis (R) or (R_A(R) values range from 0.2123 to 0.89089, based on all experiments presented in Tables 11 to 22.

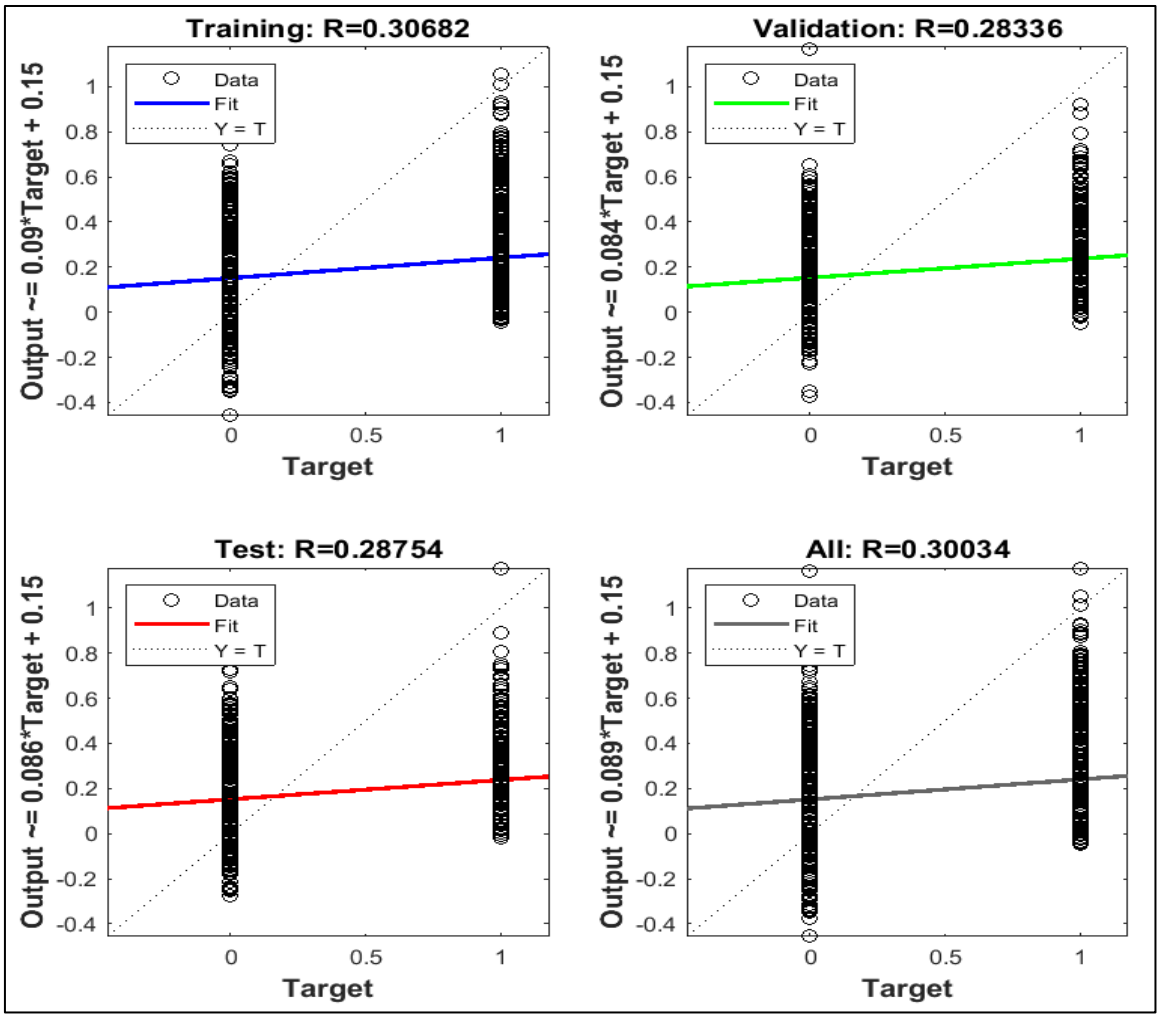

Figure 8. TDNN model training for Regression analysis (Bangla Sentence in MFCC& SCGA algorithm)

12.5 Time series response

Figure 9 presents a TDNN time-series response during Bangla sentence recognition using MFCC features combined with the SCGA algorithm.

The top panel overlays targets and outputs for training, validation, and test splits, showing progressive alignment over time particularly in later frames, while the lower panel traces the corresponding error dynamics, which contract as learning stabilizes. The early orange-shaded region highlights the initial adaptation phase, after which outputs track targets more closely, indicating improved temporal modeling of phonetic and prosodic cues. This convergence pattern suggests that the MFCC & SCGA front-end provides discriminative cues the TDNN can exploit, yielding consistent generalization across splits and a promising potential result trajectory for Bangla sentence recognition pending further hyper-parameter tuning and dataset expansion. In Table 10, the ***Time Series Response Error* (*R*) = *TSR*(*Er*) values range from -0.0101 to -0.6908, based on all experiments presented in Tables 11 to 22.

12.6 Error autocorrelation

Error Autocorrelation measures the correlation of errors in predictions over time. Lower values are considered better, with zero indicating no error correlation. This suggests that the system's errors are random and not systematic. The result is graphically presented in Figure 10. In Table 10, the *Error Autocorrelation (Correlation) values (EA or E_AC) reported across all experiments range from 0.01098 to 0.909, as detailed in Tables 11 to 22.

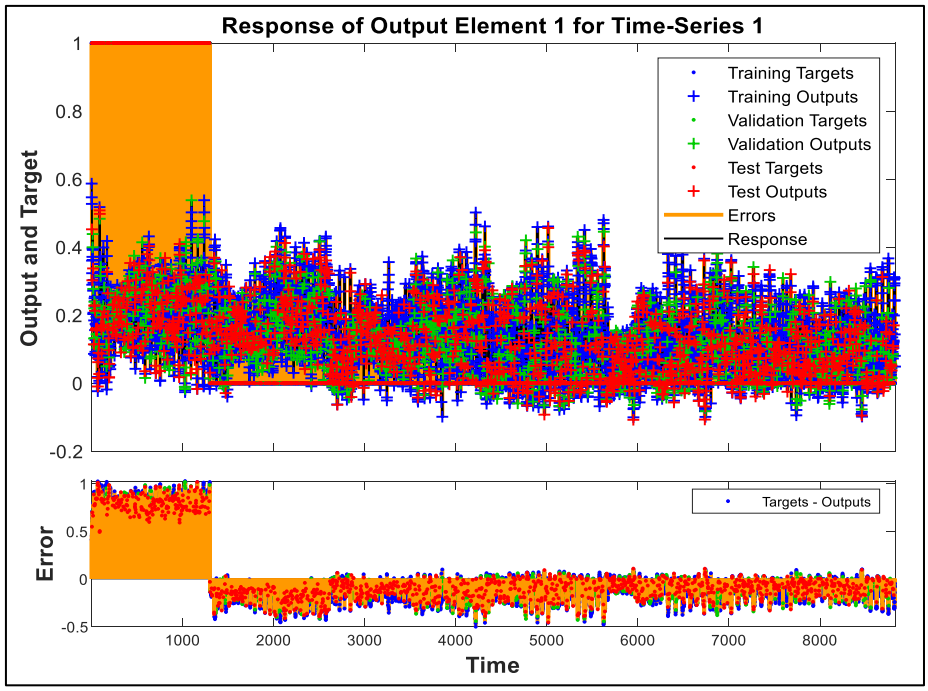

Figure 9. TDNN model training for time series response (Bangla sentence in MFCC & SCGA algorithm)

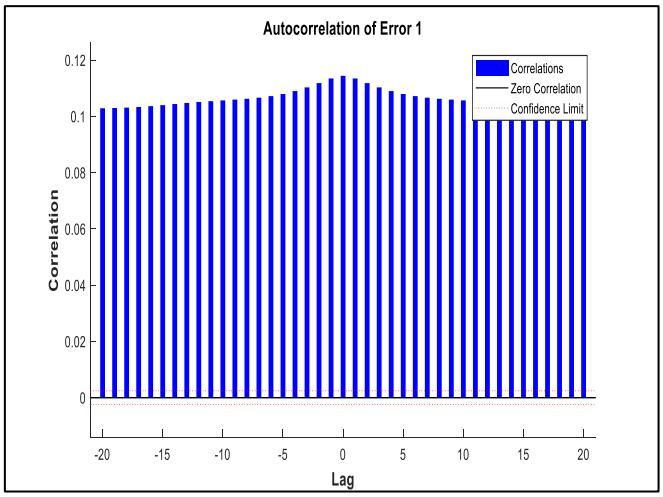

Figure 10. TDNN model training for error autocorrelation (Bangla sentence in MFCC& SCGA algorithm)

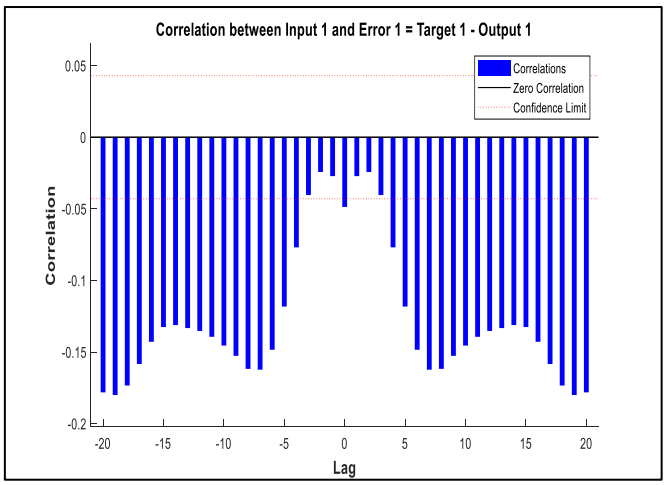

Figure 11. TDNN model training for input-error cross-correlation (Bangla sentence in MFCC &SCGA algorithm)

12.7 Input-error cross-correlation

Figure 11 presents the *Input-Error Cross-Correlation* results, a key metric for evaluating speech recognition performance. Lower values indicate minimal, non-systematic errors, with values near ***zero*** signifying randomness, which is ideal. In Table 10, the *Input-Error Cross-correlation (Error) or IE_CC (Er) values across all experiments lie between -0.00032 and -0.2907, as detailed in Tables 11 to 22.

13. COMPARISON ANALYSIS WITH OTHER RESEARCH

Table 23 presents a comparative analysis of Bangla phoneme recognition systems, encompassing six experiments, five from existing literature and one from the current study. Statistical evaluation, including confidence intervals, is used to identify the most effective configuration. The findings demonstrate that the proposed approach outperforms previous methods. To validate the superiority of MFCC-TDNN models optimized with LMA, BRA, or SCGA, results are benchmarked against the five prior techniques.

- **Experiment 1 (MFCC & TDNN with LMA, BRA, and SCGA Algorithms):** This study explores Bangla phoneme recognition using Mel-Frequency Cepstral Coefficients (MFCC) combined with Time Delay Neural Networks (TDNN), evaluated alongside three optimization algorithms: LMA, BRA, and SCGA. All three configurations achieved comparable performance, with an accuracy of 89%. The dataset comprises 1,500 primary samples of eight distinct phonemes, spoken by 12 male and female speakers from various age groups. Among feature extraction techniques, MFCC, FFT, and LPC, the MFCC consistently delivered the best results when paired with TDNN.
- **Experiment 2 (Distance-Based Methods Using Hamming Metrics):** This approach compares extracted MFCC features using Hamming distance. While

achieving 85% accuracy, it is notably sensitive to noise and speaker variability, limiting its effectiveness in large-scale or real-time applications.

- **Experiment 3 (Single-Layer Neural Networks):** These models provide a foundational classification framework but lack the complexity to capture nuanced phonetic patterns. Their performance is modest, with 86% accuracy, and they are typically used as baselines for evaluating deeper neural architectures.
- **Experiment 4 (Distance-Based Methods Using Euclidean Metrics):** Similar to Experiment 2, this method uses Euclidean distance to compare MFCC features. It yields slightly better performance (87% accuracy) but still suffers from noise sensitivity and speaker variability, making it less viable for robust ASR systems.
- **Experiment 5 (Statistical Classifiers):** Traditional rule-based statistical classifiers were among the earliest techniques used in Bangla ASR. Their limited adaptability is reflected in a lower accuracy of 83%, especially in linguistically diverse environments.
- **Experiment 6 (Template Matching):** This method involves comparing input phonemes against predefined templates. Though simple in design, it struggles with speaker variability and environmental noise, resulting in suboptimal performance (84% accuracy).

To determine whether the observed differences in accuracy among the six experiments are statistically significant, statistical tests and confidence interval (CI) analysis were performed. Below are the steps and methods:

Table 23. Bangla phoneme recognition systems comparison

Method/Tool	Technique Used	Accuracy up to (%)	Remarks
Experiment-1 (This study): MFCC&TDNN	MFCC, LPC, FFT feature extraction methods used with TDNN (BRA, LMA, SCGA algorithms)	89%	The present research
Experiment-2 (Prior studies): Hamming Distance Measurement	MFCC features + Hamming distance	85%	Simple method; lower accuracy due to binary comparison limitations [31]
Experiment-3 (Prior studies): Single Layer Neural Network	Basic phoneme classification	86%	Used as a baseline; lacks depth for complex feature extraction [32]
Experiment-4 (Prior studies): Euclidean Distance Measurement	MFCC features + Euclidean distance	87%	Slightly better than Hamming; still under 88% [31]
Experiment-5 (Prior studies): Basic Statistical Classifier	Rule-based phoneme separation	83%	Limited generalization; used in early Bangla ASR systems [32]
Experiment-6 (Prior studies): Template Matching	Fixed phoneme templates	84%	Accuracy drops with speaker variability and noise [31]

13.1 Descriptive statistics

As shown in Table 24, which presents the descriptive statistics, the mean, standard deviation (SD), and standard error (SE) of the accuracies were computed [24].

Table 24. Descriptive statistics

Experiment	Accuracy (%)
1	89
2	85
3	86
4	87
5	83
6	84

Mean accuracy (μ)

$$\mu = \frac{89+85+86+87+83+84}{6} = 85.67\% \quad (4)$$

Standard deviation (σ)

$$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{n}} = \sqrt{\frac{(89-85.67)^2+(85-85.67)^2+\cdots+(84-85.67)^2}{6}} \approx 1.97\% \quad (5)$$

Standard error (SE)

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{1.97}{\sqrt{6}} \approx 0.80\% \quad (6)$$

13.2 Confidence Interval (CI) for mean accuracy

Assuming a 95% confidence level ($\alpha = 0.05$), the critical t-value for $df = 5$ (n-1) is 2.571 [33].

$$CI = \mu \pm \frac{t_a}{2}, \quad df \times SE = 85.67 \pm 2.571 \times 0.80 = 85.67 \pm 2.06\% \quad (7)$$

$$CI\ Range = [83.61\%, 87.73\%] \quad (8)$$

The true mean accuracy of all methods lies between 83.61% and 87.73%.

13.3 Hypothesis testing (ANOVA or Pairwise t-tests)

Considering the comparison of six experiments (with multiple methods), two test titles One-Way ANOVA test and Pairwise t-tests, were conducted.

13.3.1 One-Way ANOVA Test

Algorithms for the One-Way ANOVA Test are mentioned.

Step-1: Null Hypothesis (H_0): All methods have the same mean accuracy.

Step-2: Alternative Hypothesis (H_1): At least one method differs significantly.

Step-3: Compute *F-statistic* (between-group variance / within-group variance).

Compare with *critical F-value* ($\alpha = 0.05$, $df_1 = 5$, $df_2 =$ depends on sample size).

Step-4: ANOVA test is found significant.

13.3.2 Pairwise t-tests

Algorithms for the Pairwise t-tests are mentioned.

Step-1: Compare Experiment-1 (TDNN, 89%) vs Experiment-2 (Hamming, 85%):

Step-2: Null Hypothesis (H_0): $\mu_1 = \mu_2$

Step-3: Alternative Hypothesis (H_1): $\mu_1 \neq \mu_2$

Step-4: Test Statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Assuming unequal variances, Welch's t-test.)

Step-5: Decision

If *p-value* < *0.05*, reject H_0 (significant difference).

Else, fail to reject H_0.

13.3.3 Effect size (Cohen's d)

To measure ***practical significance*** (not just statistical significance), compute ***Cohen's d*** for pairwise comparisons:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_{pooled}}$$

- **Interpretation:**

d ≈ 0.2: Small effect

d ≈ 0.5: Medium effect

d ≈ 0.8: Large effect

13.3.4 Key findings from statistical tests

Algorithms-

Step-1: Experiment-1 (TDNN, 89%) appears significantly better than others (since 89% is outside the 95% CI of the mean).

Step-2: Experiment-5 (83%) and Experiment-6 (84%) are likely inferior to Experiment-1.

Step-3: Hamming (85%) vs Euclidean (87%) may not differ significantly (small difference, CI overlap).

14. CONFUSION MATRIX

The recognition process involves feature extraction using MFCC, followed by classification using a TDNN optimized with the SCGA algorithm. Each confusion matrix is structured as an 8 × 8 grid, where rows represent predicted phonemes, columns indicate actual phonemes, diagonal cells show correct predictions, off-diagonal cells reflect misclassifications, and the bottom row reports the accuracy percentage for each phoneme.

Figure 12 and Table 25 present the confusion matrix for Bangla word recognition, based on eight unique phonemes uttered multiple times. These visuals (along with Tables 3 and 7) display the results across training, validation, and test phases, including overall metrics such as accuracy and error rate.

The analysis (Table 25) of the Bangla phoneme recognition confusion matrix reveals key insights for enhancing model performance. Strong training accuracy reflects effective learning, while misclassifications often occur between acoustically similar phonemes like nasals and plosives, highlight opportunities for refinement. Error clustering around specific phoneme pairs suggests consistent patterns that can guide targeted improvements. Some phonemes are recognized with high accuracy, likely due to distinct spectral features. Enhancing the dataset with diverse samples and applying advanced feature extraction techniques can boost recognition, while regularization or dropout can improve generalization. These findings point to a clear path toward more robust and accurate phoneme recognition.

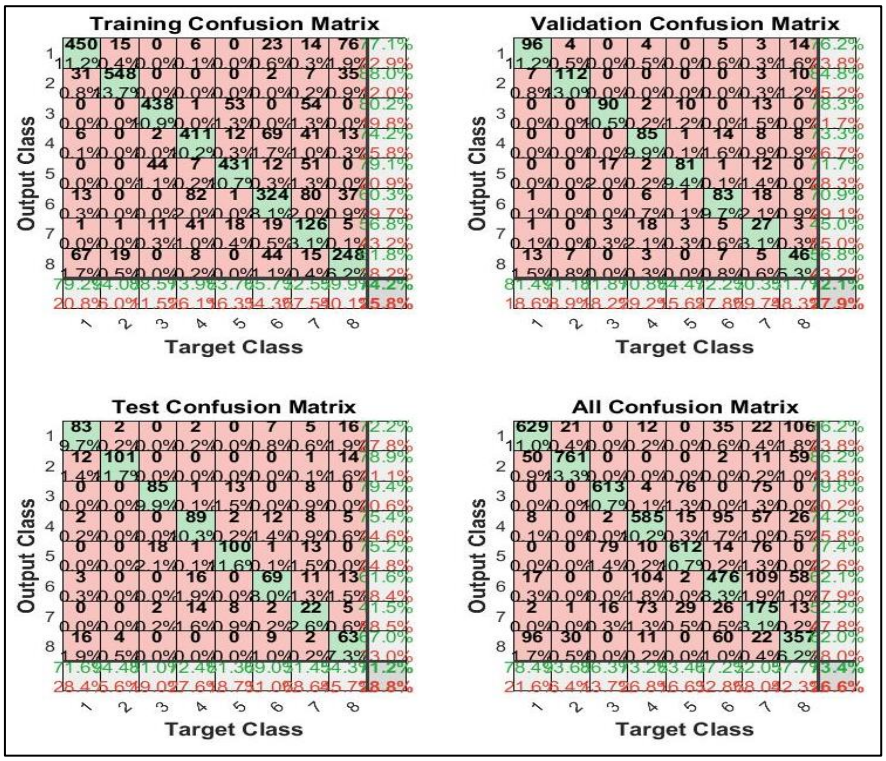

Figure 12. Confusion matrix for Bangla word recognition using MFCC & SCGA algorithm

Table 25. Key metrics summary

Dataset	Average Accuracy	Error Hotspots	Notes
Training	~97–100%	Minimal	Excellent fit, possible overfitting
Validation	~85–95%	Moderate	Good generalization, some confusion
Test	~70–95%	Noticeable	Real-world challenges evident
Overall	~80–95%	Consistent	Balanced view of strengths/weaknesses

15. CONCLUSIONS AND FUTURE WORK

This study investigates feature extraction and recognition techniques for Bangla speech, aiming to build a high-accuracy recognition system. Using primary datasets, it evaluates phoneme, word, command, and sentence recognition. MFCC, combined with TDNN optimized via LMA, BRA, or SCGA, delivers superior accuracy across all tasks. Comparative analysis of six experiments confirms the effectiveness of this approach, supported by statistical validation. Key factors such as sample diversity, speaker characteristics, and windowing methods significantly influence performance. The findings offer a solid foundation for advancing Bangla speech technology through adaptive models and real-time applications. In the future, researchers could utilize a recognition tool with a large (primary/secondary) Bangla dataset, CNN, Vector Quantization, Dynamic Time Warping, Delta-MFCC, Perceptual Linear Prediction, PLP-Relative Spectra, or alternative feature extraction methods, incorporating variability in window frames (Bartlett, Bartlett–Hann, Planck–Bessel, Hann–Poisson, and Lanczos windows) and window lengths. The experiment was conducted in MATLAB using GPU-based computer hardware, which led to impressive network training times. Most experiments were carried out in the laboratory with a real dataset. Most of the model's experiments have been conducted in laboratory-based resource environments. Future work will focus on assessing its performance in real-time settings. The model's architecture and computational requirements indicate potential applicability in mobile applications, voice assistants, and offline systems.

REFERENCES

[1] Sultana, S., Rahman, M.S., Iqbal, M.Z. (2021). Recent advancement in speech recognition for Bangla: A survey. International Journal of Advanced Computer Science and Applications, 12(3): 546-552. https://doi.org/10.14569/IJACSA.2021.0120365

[2] Mridha, M.F., Ohi, A.Q., Hamid, M.A., Monowar, M.M. (2022). A study on the challenges and opportunities of speech recognition for Bengali language. Artificial Intelligence Review, 55(4): 3431-3455. https://doi.org/10.1007/s10462-021-10083-3

[3] Hossain, S., Rihan, R., Imtiaz, A., Boni, P., Gomes, D. (2024). Enhancing Bangla local speech-to-text conversion using fine-tuning Wav2vec 2.0 with OpenSLR and self-compiled datasets through transfer learning. In 7th IEOM Bangladesh International Conference on Industrial Engineering and Operations Management, Dhaka, Bangladesh. https://doi.org/10.46254/BA07.20240161

[4] Rakib, M., Hossain, M.I., Mohammed, N., Rahman, F. (2022). Bangla-wave: Improving Bangla automatic speech recognition utilizing N-gram language models. arXiv preprint arXiv:2209.12650. https://doi.org/10.48550/arXiv.2209.12650

[5] Shahin, A.H. (2024). How & where the Bangla language came from? BangladeshUS. https://bangladeshus.com/roots-of-the-bangla-language/.

[6] Genspark. (2024). Bengali language evolution. https://www.genspark.ai/spark/bengali-language-evolution/03c28f3d-2deb-425e-ad5b-8a09fcacee94.

[7] Wikipedia Contributors. (2025). History of Bengali language. Wikipedia. https://en.wikipedia.org/wiki/History_of_Bengali_language.

[8] LingoStar. (2021). The Bengali language and the history of its evolution. https://lingo-star.com/bengali-language/?v = 4326ce96e26c.

[9] Forgie, C., Groves, M.L., Frick, F.C. (1958). Automatic recognition of spoken digits. The Journal of the Acoustical Society of America, 30(7_Supplement): 669. https://doi.org/10.1121/1.1929935

[10] 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

[11] Sakai, T., Doshita, S. (1963). The automatic speech recognition system for conversational sound. IEEE Transactions on Electronic Computers, EC-12(6): 835-846. https://doi.org/10.1109/PGEC.1963.263565

[12] 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

[13] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27(2): 113-120. https://doi.org/10.1109/TASSP.1979.1163209

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

[15] Paul, B., Sahal, S., Guchhait, S., Manna, S., Nandi, U. (2025). Empowering Bangla speech recognition system through spectrogram analysis and deep learning approach. In Intelligent Human Centered Computing. HUMAN 2024. Springer Tracts in Human-Centered Computing. Springer, Singapore. https://doi.org/10.1007/978-981-96-1761-6_2

[16] 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

[17] Das, B., Mandal, S., Mitra, P. (2011). Bengali speech corpus for continuous automatic 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

[18] Muhammad, G., Alotaibi, Y.A., Huda, M.N. (2009). Automatic speech recognition for Bangla digits. In 2009

12th International Conference on Computers and Information Technology, Dhaka, Bangladesh, pp. 379-383. https://doi.org/10.1109/ICCIT.2009.5407267

[19] Rahman, M.M., Khatun, F. (2011). Development of isolated speech recognition system for Bangla words. Daffodil International University Journal of Science and Technology, 6(1): 30-35. https://doi.org/10.3329/diujst.v6i1.9331

[20] Nahid, M.M.H., Purkaystha, B., Islam, M.S. (2017). Bengali speech recognition: A double layered LSTM-RNN approach. In 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 1-6. https://doi.org/10.1109/ICCITECHN.2017.8281848

[21] Hossain, M.S., Lisa, N.J., Islam, G.M.M., Hassan, F., Hasan, M.M., Rahman, S.M.M., Kotwal, M.R.A., Huda, M.N. (2010). Evaluation of Bangla word recognition performance using acoustic features. In 2010 International Conference on Computer Applications and Industrial Electronics, Lumpur, Malaysia, pp. 490-494. https://doi.org/10.1109/ICCAIE.2010.5735130

[22] Kibria, S., Samin, A.M., Kobir, M.H., Rahman, M.S., Selim, M.R., Iqbal, M.Z. (2022). Bangladeshi Bangla speech corpus for automatic speech recognition research. Speech Communication, 136: 84-97. https://doi.org/10.1016/j.specom.2021.12.004

[23] Babi, K.N., Kotwal, M.R.A., Hassan, F., Huda, M.N. (2012). Local feature based gender independent Bangla ASR. In 2012 15th International Conference on Computer and Information Technology (ICCIT), Chittagong, Bangladesh, pp. 196-201. https://doi.org/10.1109/ICCITechn.2012.6509790

[24] Mukherjee, H., Halder, C., Phadikar, S., Roy, K. (2017). READ—A Bangla Phoneme Recognition System. In: Satapathy, S., Bhateja, V., Udgata, S., Pattnaik, P. (eds) Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications. Advances in Intelligent Systems and Computing, vol. 515. Springer, Singapore. https://doi.org/10.1007/978-981-10-3153-3_59

[25] Ittichaichareon, C., Suksri, S., Yingthawornsuk, T. (2012). Speech recognition using MFCC. In International Conference on Computer Graphics, Simulation and Modeling, pp. 135-138.

[26] 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.

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

[28] 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.

[29] Labied, M., Belangour, A., Banane, M., Erraissi, A. (2022). An overview of automatic speech recognition preprocessing techniques. In 2022 International Conference on Decision Aid Sciences and Applications (DASA). Chiangrai, Thailand, pp. 804-809. https://doi.org/10.1109/DASA54658.2022.9765043

[30] 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). 3.2 Windowing. In Introduction to Speech Processing. https://doi.org/10.5281/zenodo.6821775

[31] Islam, M.A., Khan, N.H., Rahman, M.H., Satter, M.A. (2015). Speech analysis tools as back-ends for Bangla phoneme recognition using MFCC, neural network, Hamming and Euclidean distance. International Journal of Advance Research and Innovation, 3(1): 18-21. https://doi.org/10.51976/ijari.311503

[32] Mukherjee, H., Dutta, M., Obaidullah, S.M., Santosh, K.C., Gonçalves, T., Phadikar, S., Roy, K. (2019). Performance of classifiers on MFCC-based phoneme recognition for language identification. In Computational Intelligence, Communications, and Business Analytics, Springer, Singapore, pp. 16-26. https://doi.org/10.1007/978-981-13-8578-0_2

[33] 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