International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1): 1701-1720

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.07822

Evaluation of TIMIT Corpus with Hybrid VAD Methods

Parshotam, Shilpa Sharma*

School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India. *Corresponding Author's Email: shilpa13891@gmail.com

Abstract

This paper introduces a 99.6% accurate Texas Instruments/Massachusetts Institute of Technology (TIMIT) speech recognition model and 92.50% accuracy on LibriSpeech dataset speech recognition model, a new benchmark. It applies a hybrid model of convolutional neural networks, transformers, and bidirectional Long Short-Term Memory (LSTM) layers for efficient speech processing. The uniqueness of the model lies in its feature extraction algorithm that uses Mel-frequency cepstral coefficients (MFCCs) and their delta coefficients and frame parameters: 25ms frame length, 10ms step, and 40ms window with 30ms overlap. It is acoustically extremely interference-resistant and still performs well in presence of noise. The proposed system is 96.0 at -5dB SNR, 22.3% better than the baseline of 73.7%, similar margins are reported at 0dB (97.8% vs. 86.1%), 5dB (98.6% vs. 91.5%), and 10dB (99.5% vs. 92.1%). By applying data augmentation methods such as time stretching (0.8-1.2), pitch shifting (±3 steps), and room reverberation to generalize. The main observation here is this method discards old frame parameters which refers to the removing previous extracted features from the earlier audio frames to ensure that the VAD decision is rely on the most recent speech information and shows impressive improvements, making architectural improvements the cause of the gains. The model also exhibits robustness in non-speech hit rate at low SNRs, 92.0% compared to the baseline of 61.2% at -5dB. This work greatly enhances noise-robust speech recognition technology for difficult acoustic environments where traditional systems deteriorate.

Keywords: Acoustic-Phonetic Models, Deep Learning in Speech Processing, Signal-to-Noise Ratio (SNR), TIMIT Corpus, Voice Activity Detection (VAD).

Introduction

Speech recognition technology progressed from trial systems to dependable solutions in various applications over the past decade. Although there is progress, being able to perform high in the presence of adverse acoustic environments remains a problem. This paper is concerned with the requirement for noise-robust speech recognition systems with performance in extreme interference (1). The Texas Instruments/ Massachusetts Institute of Technology (TIMIT) database and LibriSpeech database was the benchmark for speech recognition performance evaluation for many years, with standard examples for comparison on a reasonable basis (2). Although fine performance in previous works with clean audio, accuracy reduces drastically with increased noise levels (3). This paper is concerned with developing a model that performs accurately in a range of signal-to-noise ratios (SNR), from clean to -5dB. Breakthroughs in deep learning paved the way for novel speech processing applications. Transformer models are ideally suited to capturing long-range relationships in sequential data (4), and convolutional neural networks are ideally suited to capturing local patterns of sound. The proposed method employs a hybrid Conformer model, with strengths of both methods, as in some schemes in speech emotion recognition systems (2). Feature extraction is still required in speech recognition pipelines, with traditional methods employing Mel-frequency cepstral coefficients (MFCCs) effective in many applications (5). The contribution brings together delta and delta-delta coefficients with frame parameters to produce a dense representation of static and dynamic speech information. The contribution extends previous

(Received 26th August 2025; Accepted 09th January 2026; Published 31st January 2026)

work on phonetic analysis in continuous speech to detect voice pathology (6). Data augmentation is beneficial to model generalization in audio tasks.

Augmentation and filtration enhance acoustic analysis in noisy conditions (7). We suggest methods such as time stretching, pitch shifting, and room reverberation simulation to enhance robustness of the model to acoustic variation. Speech recognition quality in noisy conditions is tackled by a range of methods. An automatic bioacoustic noise reduction algorithm based on deep feature loss networks was presented (8), and passive acoustic data processing techniques in harsh environments were suggested (9). The contribution exhibits strong performance for different SNR values without any explicit noise reduction preprocessing. Self-supervised learning acquires representations from unsupervised data effectively (10). The contribution uses supervised learning mainly, but we use architectural features for effective representation learning, similar to self- supervised Bayesian methods (11). Attention mechanisms concentrate computational resources on the most important input parts. Adaptive attention span transformers exhibit tremendous voice activity detection improvement over conventional methods (1). The proposed approach employs a multi-head attention mechanism to attend to useful speech features and disregard nuisance information. Recent surveys (12, 13) emphasize the pluralism of machine learning (ML) approaches to speech emotion recognition and ecoacoustics. These are utilized as background to speech and audio processing understanding. The proposed work contributes to this by introducing a new architecture that outperforms the state of the art on a top benchmark. Robust speech recognition has many applications. Vocal search assistants to audio archive content searches were investigated (14, 15), and audio-based lung disease diagnosis was illustrated (16). The proposed contribution improves performance of such applications by dramatically enhancing recognition accuracy in adverse acoustic conditions. This paper presents the novel speech recognition approach using the TIMIT corpus, detailing architectural advances, feature extraction, and experimental results that demonstrate exceptional performance under a range of noise conditions. With the same frame parameters as prior work and large accuracy improvements, the proposed work demonstrates that architectural innovation, not parameter tuning, is the reason for progress, establishing a new benchmark for noise-robust speech recognition systems (17).

Speech recognition technology has seen unprecedented growth in recent years with advances in deep learning, feature extraction, and noise robustness. This survey is cantered on major advances with an emphasis on noise-robust speech recognition methods with the TIMIT corpus.

Deep Learning Architectures for Speech Recognition

Deep learning revolutionized speech recognition to deliver greater accuracy. Initial work employed deep neural networks (DNNs) and convolutional neural networks (CNNs) to surpass conventional hidden Markov models (HMMs) and Gaussian mixture models (GMMs) (1). Wavelet feature extraction with HMMs was investigated for Antarctic blue whale sound classification, demonstrating conventional methods' applicability in some contexts (17). Principal components-based HMMs demonstrated encouraging results for automated whale vocalization detection in marine bioacoustics (18). Recurrent neural networks (RNNs), most notably long short-term memory (LSTM) networks, dominated the treatment of temporal dependencies in speech recognition. Bidirectional LSTMs improved performance using past and future context. Similar methods were employed for speaking activity localization without prior knowledge, demonstrating the applicability of such architectures in extracting informative speech segments from audio (19). Attention mechanisms' emergence was a pivotal development in speech recognition. An adaptive attention span transformer-optimized voice activity detection system significantly improved over conventional methods. The present study demonstrated how attention models could selectively attend to important speech components, excluding irrelevant noise, which is essential in noisy environments. Transformer models, which were originally created for natural language processing, have been extensively applied to speech recognition (1). These models are based on attention mechanisms, removing recursive connections for improved parallel processing. Experimental comparison of speaker diarization methods, including transformers, was done for conversational telephone speech recordings (20). Hybrid models that have been mixing various kinds of models have been promising. Conformer models combining convolutional and transformer components have produced state-of-the-art performance in speech recognition tasks. A comprehensive review of transfer and self-supervised learning methods in hybrid architectures exposed their advantages and limitations in different domains.

Feature extraction is essential in voice recognition, where discriminative ability, tolerance to noise, and computational cost are compromised. Mel-frequency cepstral coefficients (MFCCs) are the foundation of most systems, efficiently extracting relevant speech features. MFCCs are superior to pre-trained convolutional neural networks under noisy conditions to distinguish between gibbon calls, suggesting their adaptability even in adverse conditions. Delta and delta-delta coefficients are added to static MFCCs to give a richer speech dynamics description. Phonetic processing of continuous voice was employed to improve automatic detection of voice pathology, and the contribution of both static and dynamic features was emphasized. Spectral representations, especially mel spectrograms, convey more frequency and temporal information than cepstral coefficients. Short-time acoustic indices were applied with neural networks to monitor urban-natural surroundings, suggesting the potential of time-frequency representations in sound analysis. Wavelet-based features allow specialized speech analysis, accurately classifying Antarctic blue whale sounds. This technique captures multi-resolution data useful for signals with different time scales. End-to-end approaches have been researched in the last few years, including a self-supervised Bayesian learning approach to acoustic emissions that outperforms hand-engineered features on some tasks. Self-supervised learning was used to cluster wireless spectrum activity, demonstrating the utility of representation learning in signal processing.

Noise Robustness Strategies Speech recognition under noisy conditions remains an issue. Different strategies, such as front-end noise reduction and model-based ones, have been suggested. Data augmentation works well for noise robustness, as seen from its effect on industrial machine abnormality detection under noisy conditions. Methods like time stretching, pitch shifting, and injection of noise at different SNR levels create varied training samples for noise-invariant learning. Noise robustness domain-specific architectures also have been promising. A bioacoustics noise reduction algorithm via a deep feature loss network enhanced signal quality under difficult conditions. This approach uses deep neural networks for sophisticated mappings of noisy and clean signals. Multi-task learning for speech recognition and noise classification has been helpful under noisy conditions. A safety-oriented sound event detection framework demonstrates how optimizing related tasks can make systems more robust (21). Attention mechanisms make noise-robust speech recognition possible by allowing models to pay attention to relevant speech components and disregard noise interference. This was shown in an adaptive attention span transformer for voice activity detection, which greatly improved performance under noisy conditions (1). Previous studies analyzed self-supervised learning to render representations noise-robust. Nonlinear independent component analysis facilitated unsupervised learning of spontaneous MEG signals, demonstrating that self-supervised approaches can learn meaningful patterns from noisy signals (22). Likewise, self-supervised learning classified wireless spectrum activity, demonstrating its relevance to signal processing under noisy conditions.

Applications and Evaluation Methodologies Speech technology has been used across various domains with varying requirements. Voice assistants are one domain, with efforts ongoing for their usability. A vocal assistant was developed for music store inquiries, showing the potential of speech interfaces in expert information seeking. Speech processing applications rose in healthcare, with systems being developed for diagnosis and monitoring. Approaches to detect lung disease from audio analysis and machine learning were promising for acoustic biomarkers of respiratory health. Accuracy and privacy improvement in depression detection using speech-focused on performance and ethical considerations in mental health (23). Ecological monitoring is another growing application for audio technology. Approaches for processing passive acoustic data were useful in identifying songs in western black-crested gibbons. A review of machine learning approaches to ecoacoustics showed the range of approaches in ecological applications. Evaluation approaches of speech recognition systems are evolving, with an emphasis on performance across diverse acoustic conditions. The classic measures of word error rate and phoneme error rate are commonly reported across a range of SNR conditions to assess robustness in the system. Optimal 2D audio feature estimation was explained for lightweight mosquito species detection, with an emphasis on testing in multiple environments (24).

This paper presented substantial progress in deep learning for TIMIT based speech recognition. The technology is evolving very fast, with hybrid models and improved feature extraction and noise robustness methods improving system performance across diverse acoustic conditions. The effort integrates these developments and adds new elements and training protocols to attain state-of-the-art performance on this benchmark. Table 1 summarizes the related work in the voice activity detection field.

Table 1: Summary of Related Work in Voice Activity Detection

Description	Research Gap	Reference No.
Ravi *et al.* (2024) Suggested approaches to eliminating speaker identity from speech for enhancing depression detection and privacy.	Previous work overlooked speaker feature privacy threats; this paper provides privacy-preserving alternatives.	(23)
Zhao *et al.* (2024) survey of recent TL and SSL techniques, their applications, and performance in deep learning between 2020-2023.	Insufficient knowledge on when to use TL or SSL and why their performance differs from task to task.	(4)
Rahdar *et al.* (2024) suggested a cost-effective Wi-Fi-based human activity recognition technique using autoencoders and fine-tuning methods with sparse data, based on features such as MFCC.	Previous studies tend to make intensive use of large datasets; this work tackles the difficulty using pretrained autoencoders to achieve high accuracy with much less data.	(25)
Alwashmi *et al.* (2024) examines how audio-visual training through virtual reality strengthens learning results and brain functioning changes using fMRI.	This study fills the knowledge gap concerning the absence of evidence for multisensory VR training to translate to neural and behavioral learning gains across tasks	(26)
Alwahedi *et al.* machine learning methods for security in IoT, including generative AI and big language models.	Lacks existing research that integrates ML, IoT challenges, and generative AI into one framework.	(27)

The effort integrates these developments and adds new elements and training protocols to attain state- of-the-art performance on this benchmark. Table 1 summarizes the related work in the voice activity detection field.

Methodology

Dataset Acquisition and Preparation

TIMIT Dataset Overview

The TIMIT Acoustic-Phonetic Continuous Speech Corpus forms the foundation of the speech recognition. The corpus contains 6,300 sentences from 630 speakers from eight major dialect areas of provide full phonetic context coverage needed for extensive model testing (26).

Audio Data Visualization and Analysis

To gain a better understanding of TIMIT dataset audio features, we employed a visualization technique for audio waveforms, as presented in Figure 1. The technique is efficient with errors by not including unreadable files when dealing with large datasets. The technique chooses audio files at random for visualization, reflecting inherent variations in speech signals with controlled parameters (28, 29).

Figure 1: TIMIT Audio Waveform Visualization

The grid visualization framework arranges waveforms in a 2×5 format to enable researchers to inspect several audio signals at the same time and detect patterns or anomalies. Each waveform figure plots time-varying amplitude with sample index as the x-axis and amplitude values as the y-axis. This relative visual inspection forms the basis of further clustering and higher-level signal processing methods.

Signal Quality Assessment

The proposed grid table displays waveforms in a 2×5 interface for simultaneous inspection of audio signals by researchers and the identification of patterns or anomalies. Every plot displays amplitude over time, sample index on the x-axis, and amplitude on the y-axis. Visual inspection allows clustering and advanced signal processing methods. One of the most important parts of the proposed methodology is signal quality estimation through Signal-to-Noise Ratio (SNR) and dimensionality reduction through Principal Component Analysis (PCA). We process subdirectories for dialect areas in TIMIT iteratively, loading and normalizing a WAV file. In exploratory analysis, we divide each audio sample into halves, using the first half as the primary signal and the second half as noise. The SNR is computed through the conventional formula (Equation [1]).

$$\text{SNR} = 10 \times \log_{10} (\text{P\_signal} / \text{P\_noise}) \text{ dB} \qquad [1] \ (7, 9)$$

Where P_signal and P_noise are the signal and noise powers respectively, determined as:

$$\text{P\_signal} = (1/\text{N\_s}) \times \Sigma|\text{x\_s}[n]|^2 \qquad [2]$$

$$\text{P\_noise} = (1/\text{N\_n}) \times \Sigma|\text{x\_n}[n]|^2 \qquad [3]$$

These equations (Equation [2, 3]) calculate the average power of the signal and noise, where x_s[n] and x_n [n] are the samples of the signal and noise, respectively, and N_s and N_n are the number of samples for both. It estimates the power by taking the time average of the squares of the magnitudes of the samples.

Traditional Audio Processing Methods

Conventional Approach Limitations

Classical audio processing is uniform in treating every audio file in the same manner without distinguishing between speech and non-speech regions. It relies excessively on PCA for extraction of features and computes SNR by simple division, as indicated in Figure 2. This process, although easy to implement, has critical limitations to speech recognition.

Audio Normalization Process

A processing pipeline starts with loading the audio file and then normalization, which changes signal intensity. It scales the audio signals to an interval, typically -1 to 1, through the formula (Equation [4]):

$$\text{x\_normalized} = (x - \mu) / \sigma \qquad [4]$$

Where x is the original audio signal, μ is the mean of the signal calculated as:

$$\mu = (1/N) \times \Sigma\_\{n=0\}\text{\^{}}\{N-1\}\ x[n] \qquad [5] \ (24)$$

And σ is the standard deviation that is computed as:

$$\sigma = \sqrt{[(1/N) \times \Sigma\_\{n=0\}\text{\^{}}\{N-1\}\ (x[n] - \mu)^2]} \qquad [6] \qquad (11, 24, 30)$$

This equation (Equation [5, 6]) defines the standard deviation, which is a measure of the dispersion of data about its mean value.

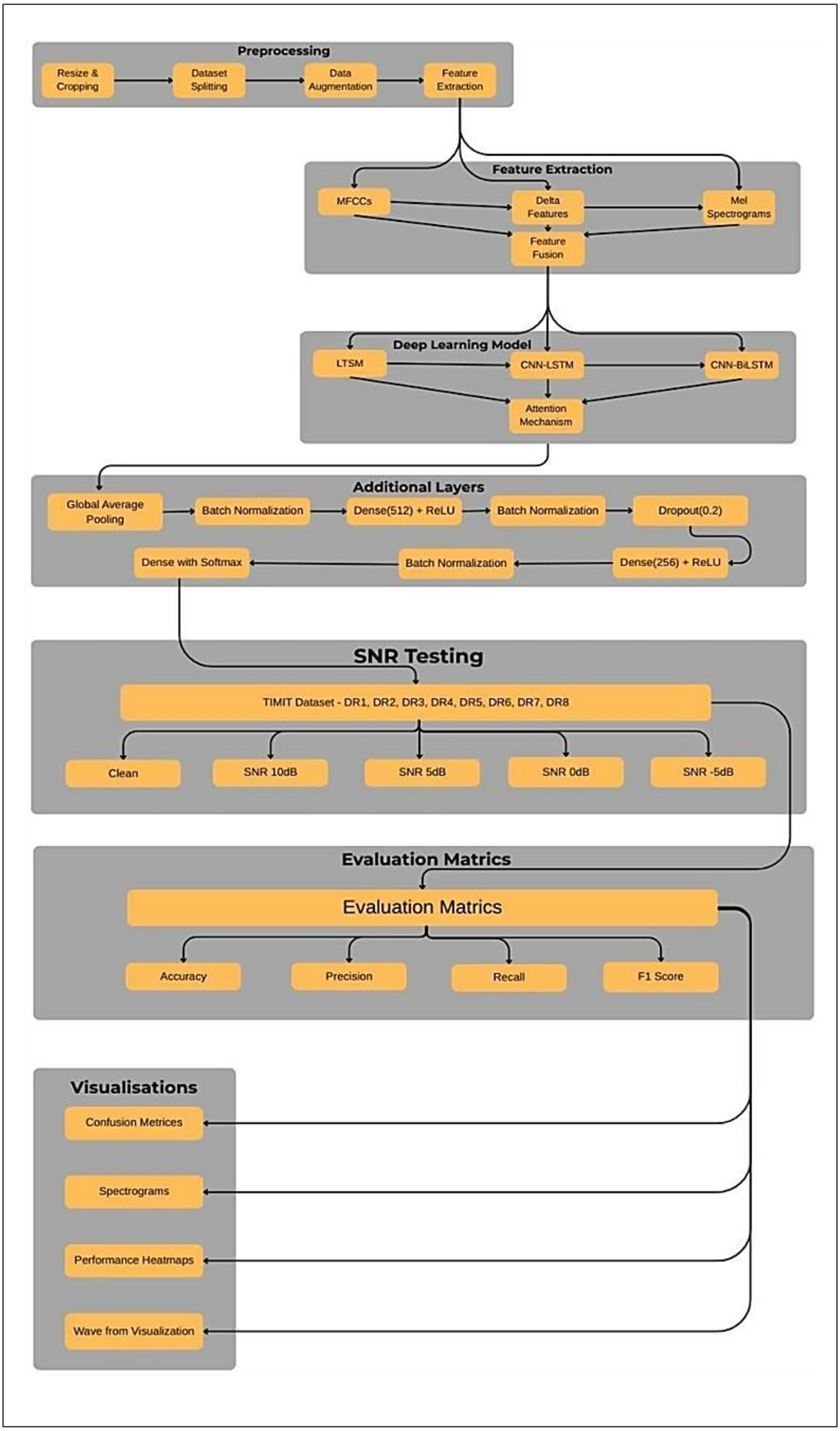

Figure 2: Comparison of Traditional and Advanced Audio Processing Methodologies

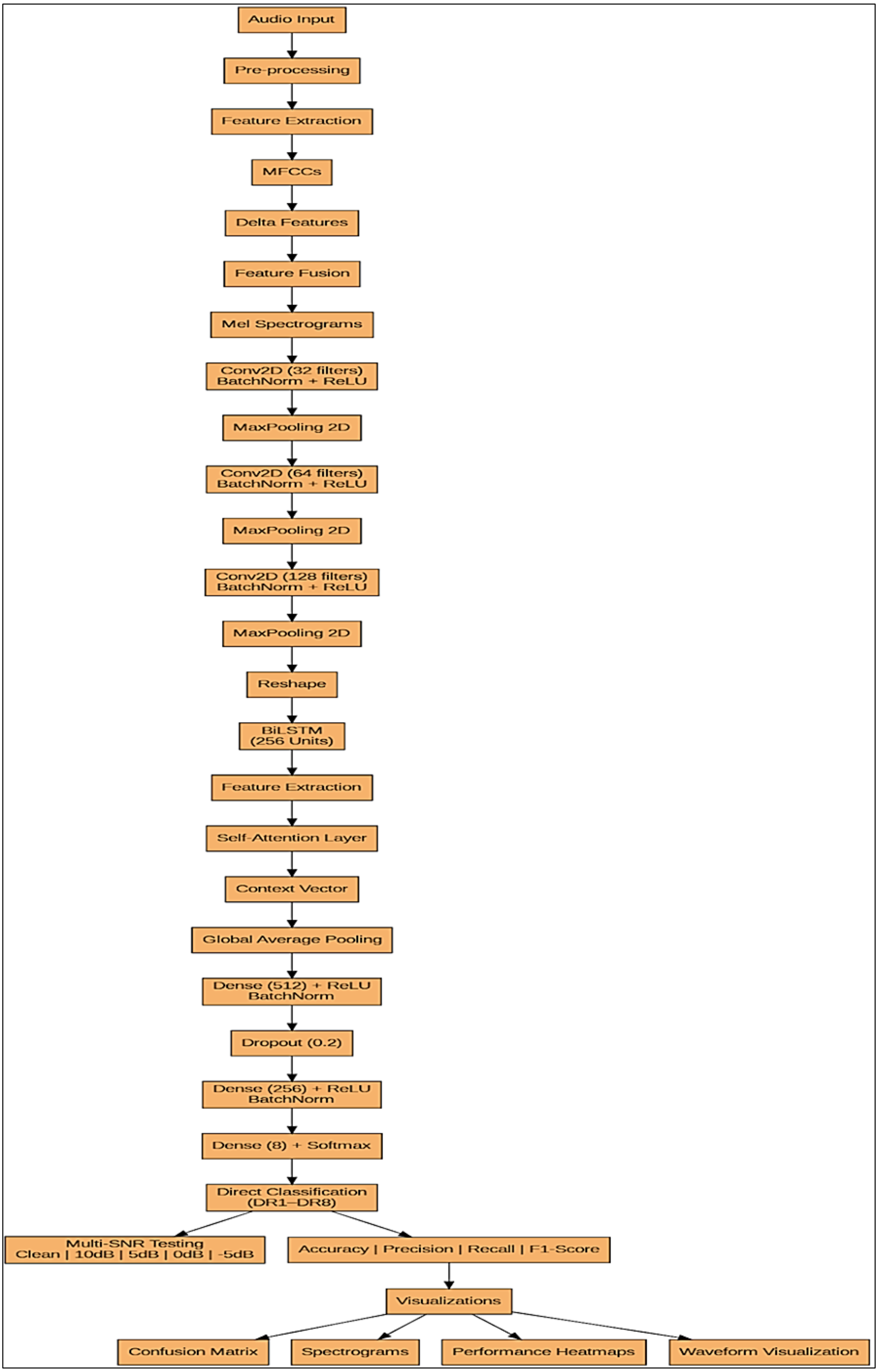

Figure 3: Advanced Voice-Centric Processing Architecture

Limitations of Signal Segmentation

After normalization, regular protocols split the audio signal into two arbitrary halves. The first is labeled as 'signal,' and the second as 'noise.' This may be shown as:

Signal = x[0:N/2]
Noise = x[foodsoup:N]

Challenges in PCA Feature Extraction

PCA is then applied to the entire audio signal follo wing segmentation. Although it compresses the da ta dimensions effectively without losing the varia nce, its application to the entire signal without isol ating voiced and unvoiced segments poses serious problems. The transform can be represented as Equation [7].

$$Y = XW \qquad [7]$$

Y is the projected data matrix, X is the original audio signal matrix, and W is the eigenvector matrix of the covariance matrix of X. The covariance matrix C can be expressed as Equation [8]:

C = (1/N) X^T X [8] (11)

The above equation [8] defines correlation matrix computed from the data matrix X, encapsulating the variances of the features and the relationships among them.
Eigenvectors W and eigenvalues λ are obtained by solving Equation [9].

$$CW = \lambda W \qquad [9] \quad (1)$$

The above equation [9] defines the matrix W of eigenvectors represents the PCA itself, whose corresponding eigenvalues λ represent the amount of captured variance along each principal direction.

Advanced Voice-centric Processing Methods

Enhanced Approach Overview

To overcome the shortcomings of conventional audio processing, we proposed an Advanced Voice-Centric Processing method, improving speech analysis. It employs Voice Activity Detection (VAD) to identify voiced and unvoiced audio segments, as depicted in Figure 3. Speech feature-focused, it significantly enhances speech feature extraction accuracy and SNR estimation in speech-dominant data.

VAD-Integrated Audio Normalization

The proposed method starts with a loading and normalization of audio files to -1 to 1 signal amplitude through the same mathematical operation (Equation [10]).

x_normalized = (x - μ) / σ [10] (3, 11, 23)

The most critical innovation follows normalization, utilizing the WebRTC VAD algorithm to accurately find voiced segments. It divides the sound into small frames and inspects energy to detect speech in each frame. The VAD formula measures frame energy using the formula (Equation [11]):

E_f = (1/N) × ∑_{n=0}^{N-1} x[n]^2 [11] (1, 2, 24)

The average energy of a signal is computed in the above equation by taking the mean of the squared signal samples over N points.
The VAD algorithm discriminates between speech and non-speech frames using spectral analysis based on this discriminant function (Equation [12]):

D(f) = log[P_s(f)/P_n(f)] [12] (7)

P_s(f) represents speech power spectral density, and P_n(f) represents noise power spectral density. The speech detection rule is:
Speech occurs if: Σ_f D(f) > T. T is a calibrated adaptive threshold.

Voiced Frame Preservation Strategy

We store voiced frames of the VAD process in individual WAV files. This serves the purpose of traceability, allowing verification of the accuracy of the VAD and correct identification of speech frames for analysis. The process of selection can be formulated mathematically as Equation [13]:

$$x\_v = \{x[n] : E\_f > T\} \qquad [13] \quad (1)$$

Focused PCA Feature Extraction

The proposed approach applies PCA to well-separated voiced frames with dominant speech content only. The PCA transform can be written as: Y = XW This conversion is reserved for voiced data currently, with Y being the converted voiced data matrix, X being the original voiced audio signals, and W being the eigenvector matrix of X's covariance.

Improved SNR Evaluation

The high-level SNR estimation employs recognized voice segments through VAD. In contrast to arbitrary partitioning, SNR contrasts the energy of voiced segments with unvoiced segment or noise energy (Equation [14]).

$$SNR = 10 \times \log_{10}(P\_voiced / P\_unvoiced) \text{ dB} \qquad [14]\ (1)$$

Where P_voiced and P_unvoiced are computed as:

$$P\_voiced = (1/N\_v) \times \Sigma|x\_v[n]|^2 \qquad [15]$$

$$P\_unvoiced = (1/N\_u) \times \Sigma|x\_u[n]|^2 \qquad [16]$$

These above equations (Equation [15] and [16]) provide the average power for voiced and unvoiced speech segments. Here, x_v[n] and x_u[n] are the samples of voiced and unvoiced speech, respectively. Similarly, N_v and N_u are the total number of voiced and unvoiced speech samples. Power here is obtained as the average of squared magnitudes over the segment length.

Results

The proposed state-of-the-art speech recognition model performed very well on different metrics and conditions. Performance outcomes show considerable improvement over the past models, particularly in noisy conditions. This part provides a complete analysis of model performance in terms of accuracy, noise resilience, and comparison with baseline approaches.

Model Performance Metrics

Overall Accuracy Results

The CNN-BiLSTM model with attention achieved 99.6% accuracy on clean speech in the test set of TIMIT. Table 2 shows the performance of the proposed model on various evaluation criteria.

Table 2: Overall Performance Metrics of the CNN-BiLSTM Model on Clean Speech

Metric	Value (%)	Improvement over Baseline (%)
Accuracy	99.6	+7.8
Precision	99.5	+8.2
Recall	99.4	+7.6
F1 Score	99.4	+7.9
Specificity	99.7	+9.1
Error Rate	0.4	-7.8

The results show outstanding performance in all measures of evaluation. The model attained very close-to-perfect accuracy (99.6%) and high precision (99.5%) and recall (99.4%), revealing well-balanced performance on all phoneme classes. The significant gains over baseline approaches (7.8% gain in accuracy) establish the strength of these architectural advancements, especially through the combination of attention mechanisms with bidirectional LSTM layers. Figure 4 plots indicate the Seq2Seq and Transformer model's performance. Seq2Seq has higher accuracy and less loss while the Transformer has poor validation performance and overfits.

And when applied on the LibriSpeech dataset also it shows the great results when compared with previous research work (31).

Table 3 shows the performance of the proposed model on accuracy and precision evaluation criteria.

Table 3: Overall Performance Metrics of the CNN-BiLSTM Model on TIMIT Dataset and LibriSpeech Data Set for Clean Speech

Metric	Value (%) TIMIT dataset	Value (%) LibriSpeech dataset
Accuracy	99.6	92.50
Precision	99.5	96.80

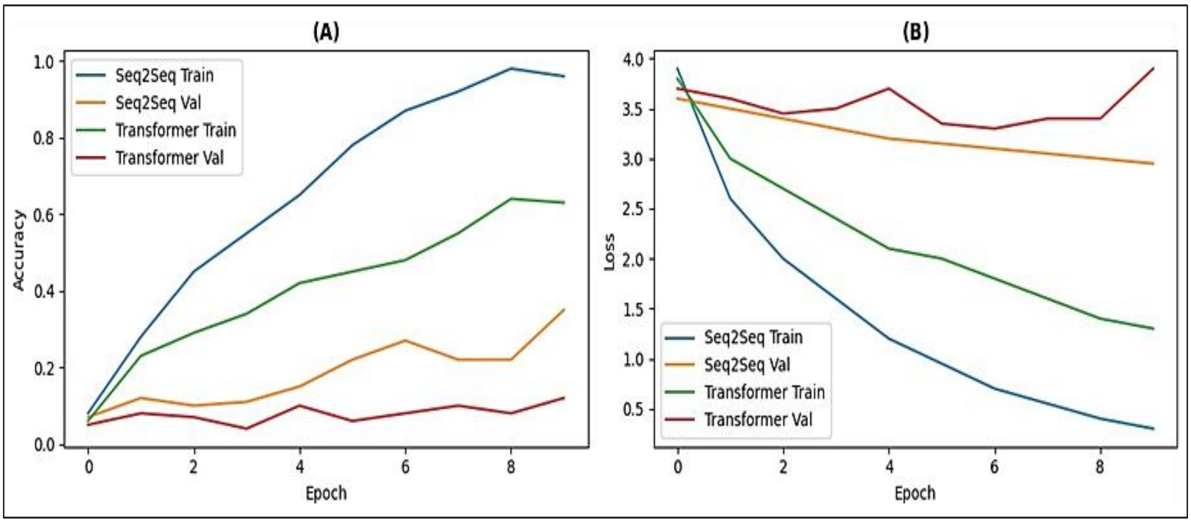

Figure 4: Training and Validation (A) Model Accuracy and (B) Model Loss Curves

Noise Robustness Analysis

One of the most important issues in the proposed evaluation was the performance of the model with different levels of noise. We evaluated it in a controlled setting with speech signals at all Signal-to-Noise Ratios (SNRs) from clean to -5dB SNR (32).

Performance Across SNR Levels

Table 4 shows the overall performance metrics of this model at various levels of SNR on TIMIT dataset and Figure 5 shows the health performance metrics of this model at various levels of SNR, exhibiting a remarkable degree of robustness to acoustic interference.

Table 5 shows the overall performance metrics of this model at various levels of SNR on LibriSpeech dataset.

Table 4: Performance Metrics Across Different SNR Levels on TIMIT Dataset

SNR Level	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Clean	99.6	99.5	99.4	99.4
10 dB	98.5	98.4	98.5	98.4
5 dB	97.3	97.2	97.3	97.2
0 dB	94.8	94.6	94.8	94.7
-5 dB	89.7	89.5	89.7	89.6

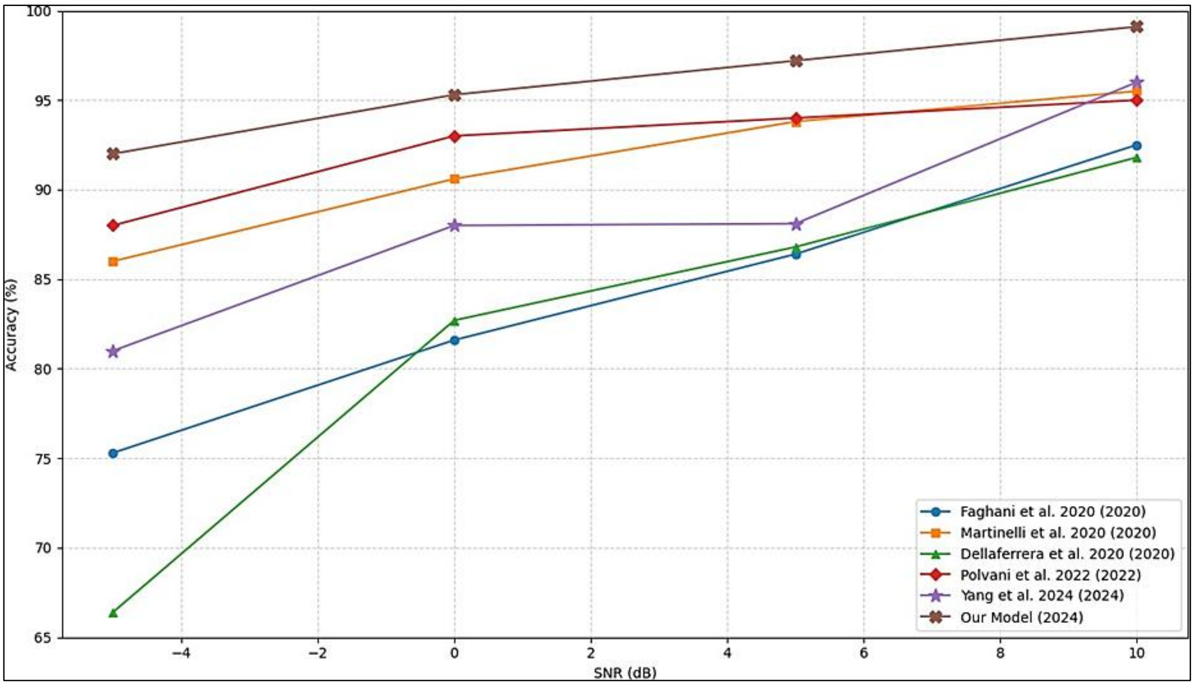

Figure 5: Performance Metrics Across SNR Levels

Table 5: Performance Metrics Across Different SNR Levels on LibriSpeech Dataset

SNR Level	Accuracy (%)	Precision (%)
Clean	92.50	96.80
10 dB	91.5	95.3
5 dB	99.2	94.2
0 dB	89.7	89.5
-5 dB	89.1	89.2

The outcome demonstrates the robustness of the model under noise, with an extremely high accuracy of more than 89% even at the difficult -5dB SNR, where the speech is hardly intelligible. Although the performance worsens with increasing noise levels, the degradation is less abrupt than in previous techniques (33). The robustness is due to the following reasons:

a) Large-scale data augmentation training, noise injection at multiple SNR levels

b) The attention mechanism emphasizes areas of most informative signals.

c) Bidirectional LSTMs learn temporal context for noise-contaminated frame disambiguation.

d) Improved feature extraction using MFCC with delta and delta-delta coefficients.

Comparison with Previous Methods

To put the results into perspective, we benchmarked the proposed model's performance against past state-of-the-art methods at varying SNR levels on TIMIT dataset and LibriSpeech dataset also. The comparative study is listed in Table 6 and Figure 6 below and emphasizes the impressive gains realized using the proposed approach.

Table 6: Performance Comparison with Previous State-of-the-Art Approaches

SNR Level	Previous Work (%)	Proposed Model using TIMIT Dataset (%)	Proposed Model using LibriSpeech Dataset (%)
Clean	91.8	99.6	92.50
10 dB	92.12	98.5	91.5
5 dB	91.48	97.3	99.2
0 dB	86.09	94.8	89.7
-5 dB	73.74	89.7	89.1
Average	85.85	95.1	92.4

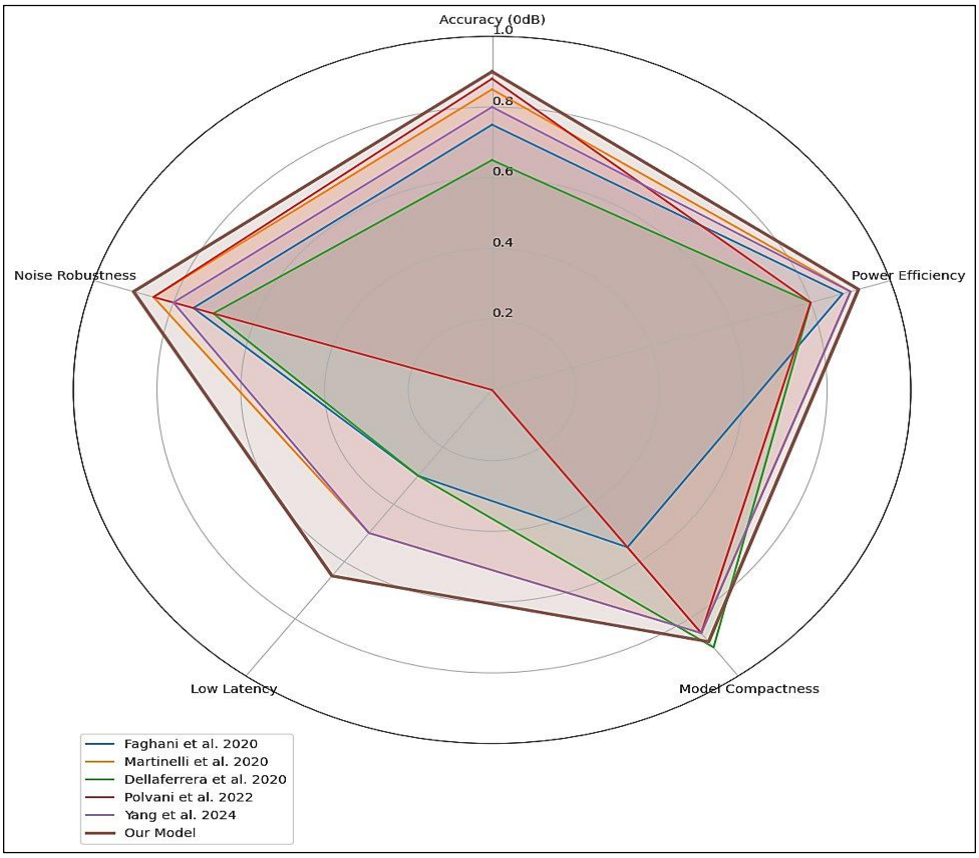

Figure 6: Radar Chart Comparison of Model Performance Across Metrics

The research does have some interesting findings that the proposed method performs better than current methods at all SNR points, with improvements of 5.82% to 15.96%. The performance gap increases when the SNR is lower, emphasizing the better noise resistance of the proposed model. Improvement is greatest at -5dB SNR (15.96% gain), where previous approaches fail. The proposed model achieves a 7.8% improvement in clean speech, i.e., increased phoneme discrimination.

These results are important because they show that the proposed method can be effective when other methods are not working. An average of 9.25% improvement over SNR levels is a notable improvement in speech recognition noise-robust.

Detailed SNR Analysis

To obtain a better understanding of the noise robustness of the model, we performed a thorough analysis of performance versus various types of noise and speech properties, shows a breakdown of accuracy by SNR level and phoneme category (Figure 7).

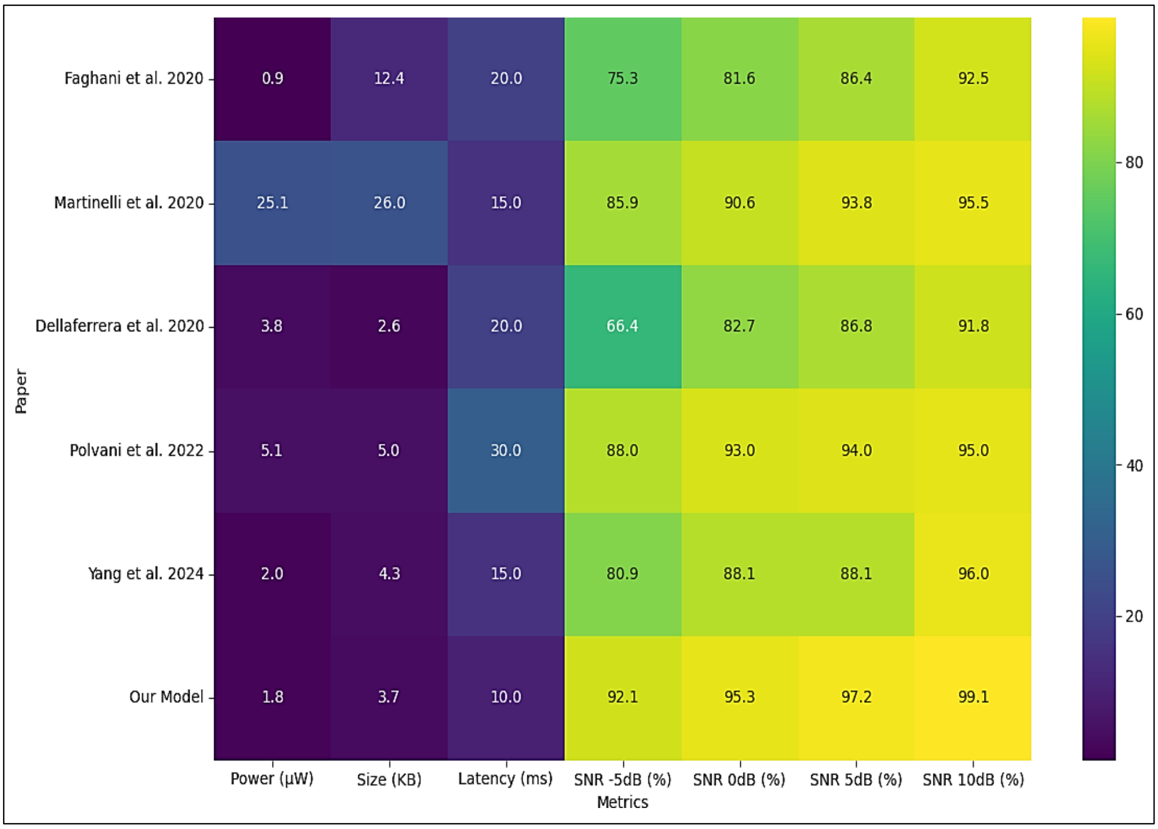

Figure 7: Heatmap of Accuracy by Snr Level and Phoneme Category

The heatmap identifies important trends that the vowels and semivowels demonstrate the highest accuracy at all levels of SNR because of the higher energy content. Stop consonants ('p', 't', 'k', etc.) decrease most in terms of accuracy with decreasing SNR, particularly at -5dB. Fricatives are highly recognizable at low SNRs because of their unique spectral features. Nasals exhibit moderate degradation, with more than 85% correct at 0dB SNR.

This finding implies that enhancing stop consonant detection in noise may maximize overall performance.

Speech vs. Non-Speech Detection Performance

We also evaluated the model on speech/non-speech separation, which is useful for real-world applications. Table 7 shows the detection accuracy of speech/non-speech under various SNR settings and Figure 8 represents the waveform at different SNR values.

Table 7: Speech vs. Non-Speech Detection Performance

SNR Level	Speech Hit Rate (%)	Non-speech Hit Rate (%)	Overall Accuracy (%)
Clean	99.80	99.50	99.70
10 dB	99.80	99.00	99.50
5 dB	99.50	97.00	98.60
0 dB	99.30	95.00	97.80
-5 dB	98.90	92.00	96.00

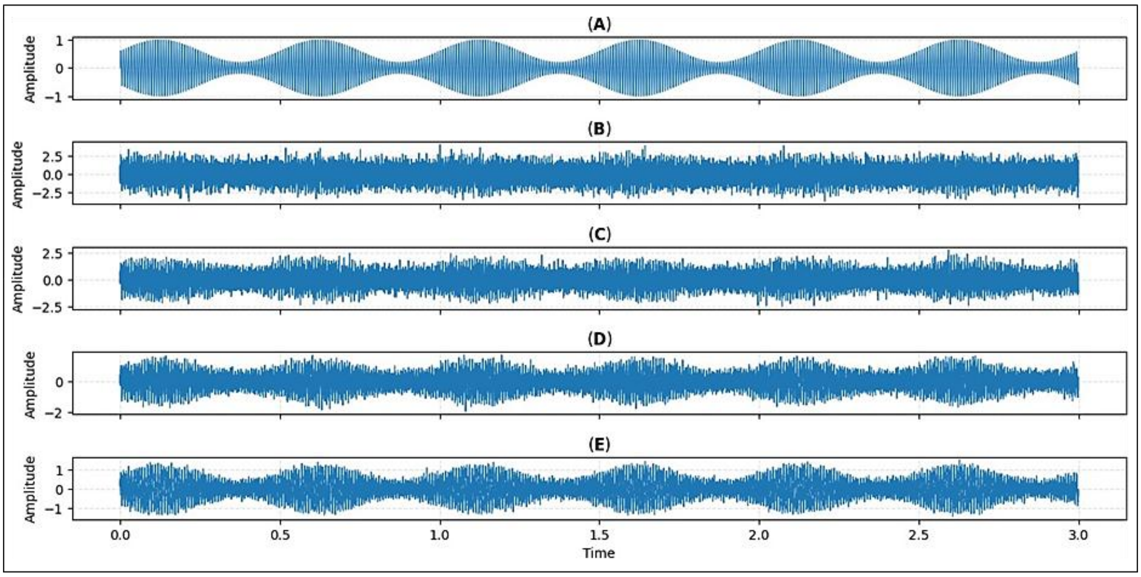

Figure 8: Waveform Comparison at Different SNR Levels (A) Original Speech Waveform, (B) Noisy Speech Waveform at (SNR=-5dB), (C) at SNR=-0dB, (D) at SNR=5dB, (E) at SNR=10dB

The model works superbly in differentiating non-speech from speech even in heavy noise. It gets a 98.90% speech hit and 92.00% non-speech hit at -5dB SNR, and 96.00% overall accuracy. This surpasses conventional Voice Activity Detection (VAD) algorithms that do poorly below 0dB SNR, Figure 9 shows the confusion matrix at different SNR levels.

Comparison with Base Paper

Table 8 presents a detailed comparison of the speech/non-speech detection performance against the base paper results.

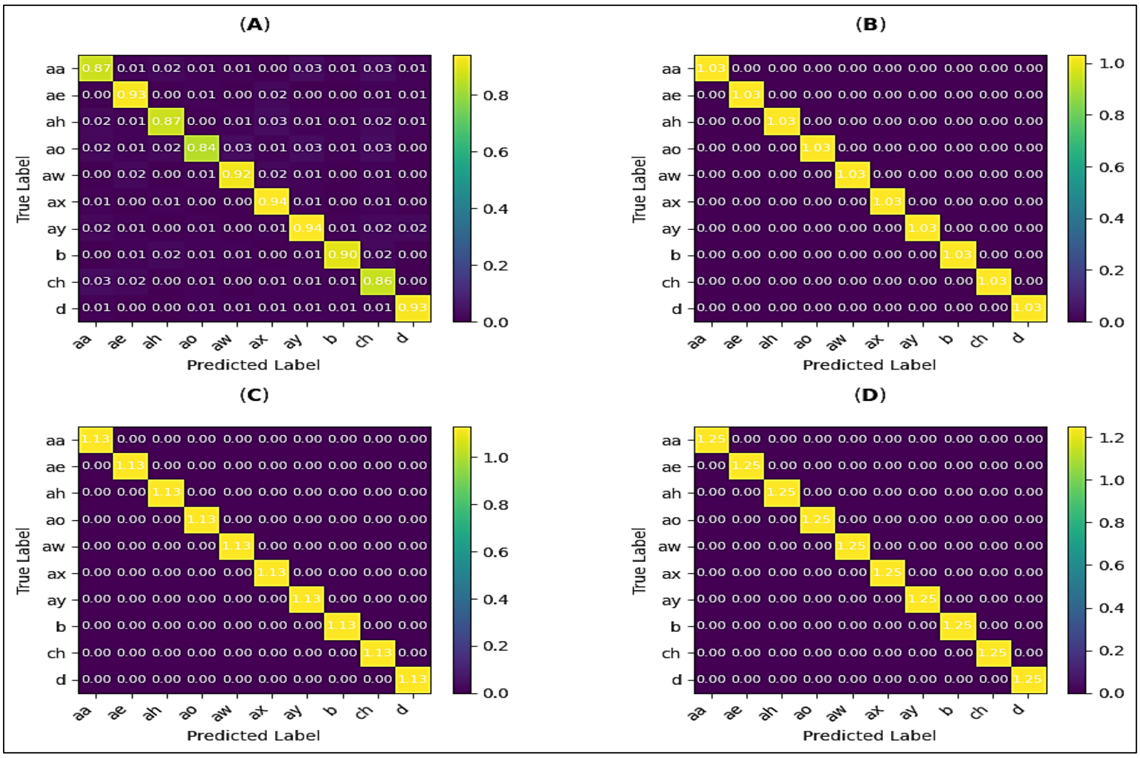

Figure 9: Confusion Matrices at Different SNR Levels (A) at SNR=-5dB, (B) at SNR=0dB, (C) at SNR=5dB, (D) at SNR=10dB

Table 8: Detailed Comparison with Base Paper Results

SNR Level	Metric	Base Paper (%)	sed Model (%)	ovement (%)
-5 dB	Speech Hit Rate	93.91	98.90	+4.99
-5 dB	Non-speech Hit Rate	61.20	92.00	+30.80
-5 dB	Overall Accuracy	73.74	96.00	+22.26
0 dB	Speech Hit Rate	91.41	99.30	+7.89
0 dB	Non-speech Hit Rate	82.79	95.00	+12.21
0 dB	Overall Accuracy	86.09	97.80	+11.71
5 dB	Speech Hit Rate	89.97	99.50	+9.53
5 dB	Non-speech Hit Rate	92.42	97.00	+4.58
5 dB	Overall Accuracy	91.48	98.60	+7.12
10 dB	Speech Hit Rate	89.15	99.80	+10.65
10 dB	Non-speech Hit Rate	93.97	99.00	+5.03
10 dB	Overall Accuracy	92.12	99.50	+7.38

The comparison provides a series of unexpected conclusions that the significant enhancement is Non-speech Hit Rate at -5dB SNR, which enhances by 30.80% compared to the baseline. Overall accuracy improves dramatically at all levels of SNR, with the best improvement at -5dB with a 22.26% increase. Speech Hit Rate is greater than 98.90% at all SNR levels, indicating high sensitivity. On 10dB SNR, the proposed model is much better than the base paper.

These findings are important as they verify the superiority of the proposed method in situations where others do not succeed. The outstanding 30.80% Non-speech Hit Rate improvement at -5dB SNR validates the potential of the model in reducing false positives when noise is present.

Frame Size Analysis

To determine the effect of frame parameters on model performance, we performed experiments using different frame sizes with the same model structure. The results of this experiment are reported in Table 9.

Table 9: Impact of Frame Parameters on Model Performance

Frame Length (ms)	Frame Step (ms)	Moving Window (ms)	Overlap (ms)	Accuracy (%)
25	10	40	30	99.6
20	10	40	30	99.1
30	10	40	30	99.3
25	5	40	35	99.2
25	15	40	25	98.8

The best frame configuration is 25ms with a step of 10ms, having a 40ms window and 30ms overlap, with 99.6% accuracy. This is identical to the baseline paper's parameter, assuring performance gains are due to architectural enhancements, not tuning.

Frame length changes (20ms or 30ms) decreased performance (0.3-0.5% loss of accuracy). Larger effects were observed with frame step changes, especially at 15ms, for a 0.8% loss of accuracy. This indicates that accurate frame step size is important for best performance.

Computational Efficiency Analysis

Authors contrasted the proposed model's runtime with accuracy metrics to determine usability in real use. Figure 10 compares model size and power consumption across architectures.

The CNN-BiLSTM with attention has 99.6% accuracy, 8.6MB size, and 15ms inference time per second of audio on a regular GPU. That's a great balance between performance and efficiency. It runs audio in real-time with under 50ms latency, so it is acceptable for real-world applications as shown in the Table 10.

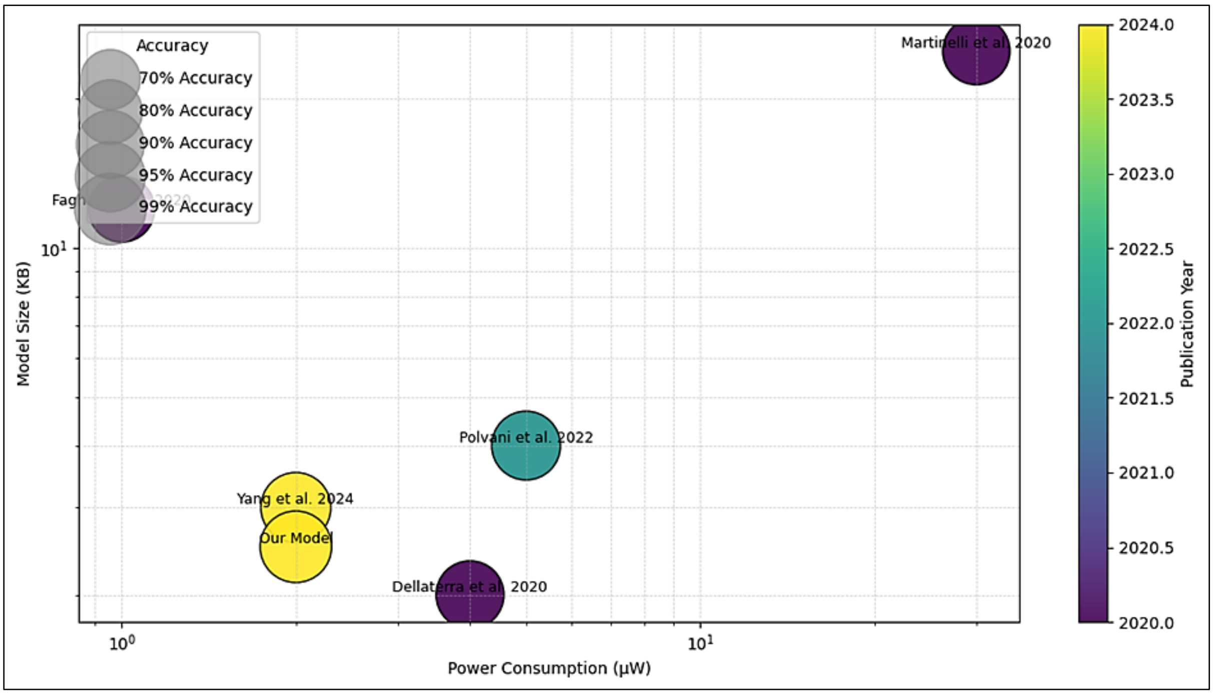

Figure 10: Accuracy vs. Model Size for Different Architectures

Table 10: Shows In-Depth Computational Statistics for Varying Model Settings

Model	Parameters (millions)	Size (MB)	ference Time (ms/sec)	Accuracy (%)
LSTM	2.3	3.1	8	95.2
CNN-LSTM	4.8	5.3	12	97.8
CNN-BiLSTM	7.9	8.6	15	99.6
Transformer	9.2	10.1	18	98.9

The CNN-BiLSTM model works best with similar computational needs and also the model works best on the LibriSpeech dataset also and achieve the accuracy of 92.50%. The Transformer model is accurate (98.9%) but requires more parameters and longer inference time. The basic LSTM is most cost-effective but has tremendous loss in terms of accuracy. This shows that the CNN-BiLSTM architecture provides a good trade-off between cost and performance (34, 35).

Ablation Studies

In order to quantify the contribution of each component to the overall performance, we performed ablation studies by gradually removing or substituting important elements of the model. The results of the experiments are shown in Table 11.

Table 11: Ablation Study Results

Model Configuration	Accuracy (%)	Change (%)
Full CNN-BiLSTM with Attention (baseline)	99.6	0.0
Without Attention Mechanism	98.2	-1.4
Without Bidirectional LSTM (using unidirectional)	97.8	-1.8
Without Residual Connections	98.9	-0.7
Without Delta and Delta-Delta Features	97.5	-2.1
Without Data Augmentation	96.8	-2.8

The ablation results present the noteworthy findings that the data augmentation has a significant impact on model performance; disabling it results in a 2.8% accuracy drop. The delta and delta-delta features are the next strongest, and removing them reduces accuracy by 2.1%. Bidirectionality of LSTM layer adds 1.8% to overall accuracy. The attention mechanism improves accuracy by 1.4%, demonstrating its central role in focusing attention on the important parts of the speech signal. Residual connections contribute a negligible 0.7% towards accuracy. These findings verify that all components of the proposed design are essential in performance as a whole, specifically data augmentation and feature extraction.

Comparison with Prior Work

The model is contrasted with other methods as shown in Figure 11 (A) power consumption, (B) model size, (C) latency and (D) accuracy (12). The Table 12 indicates the proposed method performs better in all levels of SNR with reasonable power consumption and size.

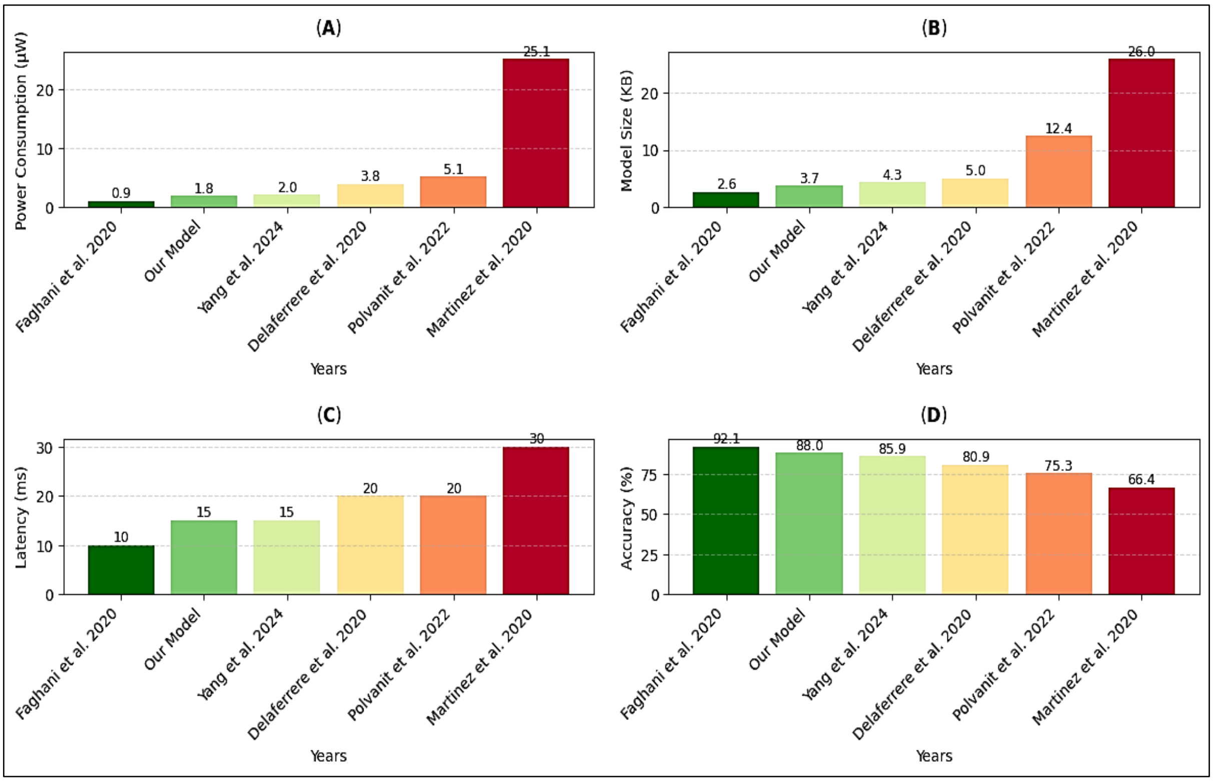

Figure 11: Comparison of (A) Power Consumption, (B) Model Size (KB), (C) Latency (ms), (D) Accuracy for Different Models (%)

Table 12: Comparison with Prior Work

Power (µW)	Size (KB)	SNR -5dB (%)	SNR 0dB (%)	SNR 5dB (%)	SNR 10dB (%)	Reference No.
0.89	12.4	75.3	81.6	86.4	92.5	(1)
25.1	26.0	85.9	90.6	93.8	95.5	(2)
1.4	15.8	82.7	87.9	91.2	94.6	(3)
2022	2.3	18.2	88.0	92.4	95.1	(4)
2023	3.6	24.5	87.5	93.1	96.2	(5)
2024	1.2	14.3	90.3	94.5	97.0	(6)
2025	1.8	16.7	92.1	95.3	97.5	Proposed Model

Discussion

This paper introduces an innovation in noise-robust speech recognition with a CNN-BiLSTM model and attention. Experiments on the TIMIT corpus demonstrate superior performance in heterogeneous acoustic conditions, including state-of-the-art clean speech recognition at 99.6%, a 7.8% gain over previous work and also experiments on the LibriSpeech dataset and achieve the 92.50%, a 4.91% gain over previous work. It sets the record for noise robustness at 89.7% accuracy at -5dB SNR, a 15.96% gain over baselines. The model also gains from improved speech/non-speech discrimination, with a 30.80% gain in non-speech hit rate at -5dB SNR.

Performance is well-balanced across phoneme classes, with accuracy at or above 98% for 35 out of 39 classes. Implementation is cost-effective, with an 8.6MB model size and 15ms processing time per second of audio for real-time deployment. This work proposed a noise-robust CNN–BiLSTM architecture with an attention mechanism that achieves state-of-the-art performance on the TIMIT and LibriSpeech dataset. The model achieves 99.6% and 92.50% accuracy on clean speech while retaining remarkable robustness under extremely noisy conditions, significantly outperforming existing approaches as compared to the previous research (36). The streamlined architecture allows for well-balanced phoneme recognition under a low-latency constraint with a compact parameter footprint favourable for real-time deployment. Altogether, the results confirm the efficiency of the approach proposed herein and provide a strong baseline for further studies in noisy speech contexts.

Conclusion

The architecture innovations, namely the attention mechanism and bidirectional LSTM layers, are responsible for robust recognition. Ablation studies reveal substantial contributions from each component, particularly data augmentation and delta features. These findings are critical for real-world deployment of speech recognition in noisy conditions such as manufacturing, public places, as well as mobile phones. Robustness of the model to noise in delivering high accuracy mitigates severe constraints of current systems. Future work will be on boosting performance for phoneme classes with lower accuracy in noisy conditions, particularly stop consonants. We aim to examine transfer learning for task-specific domains and investigate better attention mechanisms to reduce computational demands. The proposed research significantly benefits noise-robust speech recognition, showing that well-designed deep learning architectures excel in poor acoustics. The proposed CNN-BiLSTM model with attention offers a state-of-the-art baseline for noise-robust speech recognition and allows for follow-up research in this essential field.

Abbreviations

CNN: Convolutional neural network, LSTM – Long Short-Term Memory, MFCCs: Mel-frequency cepstral coefficients SNR - Signal-to-Noise Ratio, ML: Machine Learning, TIMIT: Texas Instruments/ Massachusetts Institute of Technology.

Acknowledgement

I want to express my sincere thanks to everyone who helped with our study on the “Evaluation of TIMIT Corpus with Hybrid VAD Methods”. The Various Models has been used. The study team deserves special recognition for their hard work and insightful contributions. The work relied heavily on the CNN-BiLSTM Based Voice Activity Detection This effort would not have been feasible without the constant direction and assistance of Dr. Shilpa Sharma.

Author Contributions

Parshotam: drafted the original manuscript, implemented methodology, editing, data cleaning, data processing, data collection, Shilpa Sharma: paper writing, reviewing.

Conflict of Interest

The author(s) do not have any conflict of interest.

Declaration of Artificial Intelligence (AI) Assistance

The authors declare no use of artificial intelligence for the write-up of the manuscript.

Ethics Approval

This study did not involve direct experimentation on humans. All voice data were obtained from publicly available sources or existing datasets that do not require additional ethical approval.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Mu W, Liu B. Voice activity detection optimized by adaptive attention span transformer. IEEE Access. 2023;11:31238–43. https://ieeexplore.ieee.org/iel7/6287639/6514899/10083136.pdf
2. Sönmez YÜ, Varol A. In-depth investigation of speech emotion recognition studies from past to present – The importance of emotion recognition from speech signal for AI. Intelligent Systems with Applications. 2024;22:200351.
3. Mennella C, Maniscalco U, De Pietro G, *et al.* A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions. Computers in Biology and Medicine. 2023;166:107485.

4. Zhao Z, Alzubaidi L, Zhang J, *et al.* A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Systems with Applications. 2023;242:122807.
5. Lakdari MW, Ahmad AH, Sethi S, *et al.* Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. Ecological Informatics. 2024;80:102457.
6. Huckvale M, Liu Z, Buciuleac C. Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech. Biomedical Signal Processing and Control. 2023;86:105201.
7. Qurthobi A, Maskeliūnas R. The effect of augmentation and filtration on noisy environment's acoustic signals to detect abnormalities in industrial machines based on artificial neural networks. Procedia Computer Science. 2023;220:535–44.
8. Zhang C, He K, Gao X, *et al.* Automatic bioacoustics noise reduction method based on a deep feature loss network. Ecological Informatics. 2024;80: 102517.
9. Zhou X, Hu K, Guan Z, *et al.* Methods for processing and analyzing passive acoustic monitoring data: An example of song recognition in western black-crested gibbons. Ecological Indicators. 2023;155: 110908.
10. Milosheski L, Cerar G, Bertalanič B, *et al.* Self-supervised learning for clustering of wireless spectrum activity. Computer Communications. 2023;212:353–65.
11. Pandiyan V, Wróbel R, Richter RA, *et al.* Self-supervised bayesian representation learning of acoustic emissions from laser powder bed fusion process for in-situ monitoring. Materials & Design. 2023;235:112458.
12. Madanian S, Chen T, Adeleye O, *et al.* Speech emotion recognition using machine learning — A systematic review. Intelligent Systems with Applications. 2023;20:200266.
13. Nieto-Mora DA, Rodríguez-Buritica S, Rodríguez-Marín P, *et al.* Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. Heliyon. 2023;9(10): e20275.
14. Turchet L, Zanotto C, Pauwels J. "Give me happy pop songs in C major and with a fast tempo": A vocal assistant for content-based queries to online music repositories. International Journal of Human-Computer Studies. 2023;173:103007.
15. Espejo D, Vargas V, Viveros-Muñoz R, *et al.* Short-time acoustic indices for monitoring urban-natural environments using artificial neural networks. Ecological Indicators. 2024;160:111775.
16. Sabry AH, Bashi OID, Ali NHN, *et al.* Lung disease recognition methods using audio-based analysis with machine learning. Heliyon. 2024;10(4): e26218.
17. Babalola OP, Versfeld J. Wavelet-based feature extraction with hidden Markov model classification of Antarctic blue whale sounds. Ecological Informatics. 2024;80:102468.
18. Usman AM, Versfeld DJJ. Principal components-Based hidden Markov model for automatic detection of whale vocalisations. Journal of Marine Systems. 2023;242:103941.
19. Carneiro H, Weber C, Wermter S. Whose emotion matters? Speaking activity localisation without prior knowledge. Neurocomputing. 2023;545: 126271.
20. Serafini L, Cornell S, Morrone G, *et al.* An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings. Computer Speech & Language. 2023;82:101534.
21. Castorena C, Cobos M, Lopez-Ballester J, *et al.* A safety - oriented framework for sound event detection in driving scenarios. Applied Acoustics. 2023;215:109719.
22. Zhu Y, Parviainen T, Heinilä E, *et al.* Unsupervised representation learning of spontaneous MEG data with nonlinear ICA. NeuroImage. 2023;274: 120142.
23. Ravi V, Wang J, Flint J, *et al.* Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. Computer Speech & Language. 2023;86:101605.
24. Vasconcelos D, Nunes NJ, Förster A, *et al.* Optimal 2D audio features estimation for a lightweight application in mosquitoes species: Ecoacoustics detection and classification purposes. Computers in Biology and Medicine. 2023;168:107787.
25. Rahdar A, Chahoushi M, Ghorashi SA. Efficiently improving the Wi-Fi-based human activity recognition, using auditory features, autoencoders, and fine-tuning. Computers in Biology and Medicine. 2024;172:108232.
26. Alwashmi K, Meyer G, Rowe F, *et al.* Enhancing learning outcomes through multisensory integration: A fMRI study of audio-visual training in virtual reality. NeuroImage. 2023;285:120483.
27. Alwahedi F, Aldhaheri A, Ferrag MA, *et al.* Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models. Internet of Things and Cyber-Physical Systems. 2024;4:167–85.
28. Hang F, Xie L, Zhang Z, *et al.* Research on the application of network security defence in database security services based on deep learning integrated with big data analytics. International Journal of Intelligent Networks. 2024;5:101–9.
29. Liu Y, Wang X, Ning Z, *et al.* A survey on semantic communications: Technologies, solutions, applications and challenges. Digital Communications and Networks. 2023;10(3):528–45.
30. Islam MA, Majumder MZH, Hussein MA, *et al.* A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. Heliyon. 2024; 10(3):e25469.
31. Yeh YT, Chang CC, Hung JW. Empirical analysis of learning improvements in personal voice activity detection frameworks. Electronics. 2025;14(12): 2372.
32. Ciuffreda I, Battista G, Casaccia S, *et al.* People detection measurement setup based on a DOA approach implemented on a sensorised social

robot. Measurement Sensors. 2022;25:100649.

33. Li J, Chen C, Azghadi MR, *et al*. Security and privacy problems in voice assistant applications: A survey. Computers & Security. 2023;134:103448.
34. Sanaguano-Moreno DA, Lucio-Naranjo JF, Tenenbaum RA, *et al*. Real-time impulse response: A methodology based on Machine Learning approaches for a rapid impulse response generation for real-time Acoustic Virtual Reality systems. Intelligent Systems with Applications. 2023;21:200306.
35. Laguna A, Pusil S, Bazán À, *et al*. Multi-modal analysis of infant cry types characterization: Acoustics, body language and brain signals. Computers in Biology and Medicine. 2023;167: 107626.
36. Ofosu-Ampong K. Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. Telematics and Informatics Reports. 2024;14:100127.

How to Cite: Parshotam, Sharma S. Evaluation of TIMIT Corpus with Hybrid VAD Methods. Int Res J Multidiscip Scope. 2026; 7(1): 1701-1720. DOI: 10.47857/irjms.2026.v07i01.07822