International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 12, No. 4, 2025, pp. 507-522.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Acoustic-Phonetic Analysis Using the TIMIT Corpus with Traditional and Deep Learning Methods

Parshotam [a], Shilpa Sharma [b*], Jeevan Bala [c]

[a] *School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.*
E-mail: parshotampal90@gmail.com
[b*] *School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.*
E-mail: shilpa13891@gmail.com
[c] *School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India.*
E-mail: jeevan94.47@gmail.com

Abstract

In the paper, a powerful automatic speech recognition system as per the TIMIT corpus is presented, which utilizes a hybrid deep learning model that incorporates convolutional neural networks (CNN), attention through transformers, and long short-term memory layers (BiLSTM). This combined method is made to solve noise-resistant speech recognition: it effectively captures spectral and temporal correlations of speech signals. The accuracy is groundbreaking 99.6% that the framework earns a new level of standards in the sphere. Also, the system is highly noise resilient, with an accuracy of 96% at -5 dB signal-noise ratio (SNR), which is much greater than the former models. One of the most interesting features of this work is that it has a designed pipeline of feature extraction, which includes Mel-frequency cepstral coefficients (MFCCs), as well as delta and delta-delta features. This combination helps to increase the capability of the model to represent both static and dynamic elements of speech. It shows that the architectural advancements are the primary cause of the performance improvements, and the given changes in the conventional framing parameters are not related to its performance. Another strength is that the model can be very accurate in conditions of noise where it can reach up to 89.7% accuracy at -5 dB SNR in comparison to the base line of 73.7%. The study adds to the existing literature on deep learning-based speech recognition as it proposes a model that can be used in a wide range of acoustic conditions without any explicit noise reduction preprocessing's. The implications of the findings to real-world applications, such as voice-controlled assistants in noisy settings, are substantial, and the findings are a solid base in the further development of the technologies of noisy speech recognition. Future studies will be aimed at performance improvement in phonetically difficult courses and increasing computational efficiency.

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

DOI: 10.71086/IAJSE/V12I4/IAJSE12114
Received: 27 October 2025/Revised: 26 November 2025/Accepted: 16 December 2025/Published: 30 December 2025

Introduction

The speech recognition technology has evolved significantly during the past ten years. It has developed experimental systems with restricted functionality to powerful solutions applicable in any number of applications. Although this has been done, the ultimate goal of high accuracy in difficult acoustic environments remains a fundamental problem. The reason this paper is an urgent solution is that there is a serious problem in speech recognition systems that are sensitive to noise that do not lose their functionality even under severe acoustic interference conditions (Mu & Liu, 2023).

The standard of speech recognition systems has been the TIMIT database which has been used as a standardized speech sample to make a fair comparison among the various methods (Sönmez & Varol, 2024). Experiments conducted in the past have performed reasonably well using clean audio, however, accuracy tends to decline considerably when the noise level increases (Mennella et al., 2023). This limitation is specifically addressed in this task through the development of a model, which has a high accuracy in a wide range of signal-to-noise ratios (SNR), between clean and -5dB. The development of deep learning architecture has paved the way to new applications of speech processing. Transformer models have been quite useful in the long-range representation of sequential data (Laguna et al., 2023; Islam et al., 2024). Convolutional neural networks are effective in localized sound pattern detection. This paper exploits these strengths through a hybrid Conformer model that is a combination of these two techniques, as are some other strategies in speech emotion recognition systems (Sönmez & Varol, 2024).

The feature extraction continues to form an important component of speech recognition pipelines. Mel-frequency cepstral coefficients (MFCCs) are often used in traditional methods, and have been successfully used in several speech processing tasks (Lakdari et al., 2024). It is based on this contribution, which incorporates delta and delta-delta coefficients with correct frame parameters. The outcome is a fine representation of both stationary and dynamic aspects of the speech signal. This technique is based on the study indicating that the role of the phonetic analysis in the context of continuous speech is significant to detect voice disorders (Huckvale et al., 2023). Data augmentation has been found to be a powerful approach to enhancing model generalization in audio processing problems. The approach to the analysis of acoustic signals in noisy manufacturing settings can be augmented and filtered (Qurthobi & Maskeliūnas, 2023). Based on this strategy, we present extensive augmentation techniques of time stretching, pitch shifting, and room reverberation to make this model more robust to acoustic variations.

There has been a challenge in the maintenance of quality of speech recognition at a noisy environment and several ways have been used to solve this challenge. A deep feature loss network based automatic bioacoustics noise reduction method was suggested (Zhang et al., 2024). Also, the techniques of passive acoustic monitoring information processing in adverse field conditions have been designed (Shetty & Siddiqa, 2019). The work can add to the current body of literature, showing good results over a variety of SNR values without the use of noise reduction preprocessing. Unsupervised learning Unsupervised learning has demonstrated itself to be useful in deriving useful representations of unsupervised data across diverse fields (Jabakumar, 2024). Although the present principal approach is based on supervised learning, we take advantage of structural characteristics that endorse effective representation learning, similar to self-supervised Bayesian representation learning (Pandiyan et al., 2023). The attention mechanisms have been beneficial to center the computational resources on the most significant aspects of the input sequences. An example is adaptive attention span transformers in voice activity detection which have shown great enhancements compared to the conventional techniques (Sanaguano-Moreno et al., 2024). In this model, a multi-head attention mechanism is integrated to enable the model select on significant speech content by filtering out irrelevant information.

The number of machine learning techniques applied to speech emotion recognition and Eco acoustics has been recently emphasized in the reviews (Liu et al., 2024; Madanian et al., 2023). These reviews offer beneficial background of the whole scene of speech and audio processing methods. This work has been added to this growing body of work by proposing a new architecture that delivers state-of-the-art performance on a popular benchmark. The practical use of successful speech recognition is spread over numerous areas. To give some examples, content-based queries to audio repositories with vocal assistants have been investigated (Turchet et al., 2023; Espejo et al., 2024) and audio-based analysis has been demonstrated to be promising in the recognition of lung diseases (Sabry et al., 2024). The abilities of these various applications are also advanced in this work where the performance of recognition is considerably better in challenging acoustic environments.

This paper describes a novel technique of speech recognition which is based on the TIMIT database. It gives information regarding architectural advancements, and the feature extraction procedure, and tests results which reveal that it performs well in various noisy scenarios. This study finds that improvements to noise sensitive speech recognition systems are carried out by the same parameters as that of the earlier and achieve significant accuracy

improvements indicating that the framework of architectural development, and not parameter modification, underlies the achievement of the desired results making it a benchmark in terms of noise resistant speech recognition systems (Babalola & Versfeld, 2024).

Literature Review

The last few years have seen major progress in the sphere of speech recognition, such as state-of-the-art deep learning structure, feature extraction methodology, and noise resistance approach. A literature review that investigates the significant advances in each of these fields in particular with reference to noise-resistant speech recognition based on the TIMIT corpus data is as follows.

The introduction of deep learning paradigm has had a significant impact on the development of speech recognition systems. Modern models have exceeded the performance limits of traditional statistical models, pointing on the new achievements in recognition accuracy and strength. This direction was initially advanced by deep neural networks (DNNs) and convolutional neural networks (CNNs) that proved to be much more beneficial than Hidden Markov Model (HMM) and Gaussian mixture model (GMM) based pipelines (Mu & Liu, 2023). Irrespective of these developments, classical probabilistic models still have some useful applications in some specialized settings. In other cases, HMMs with wavelet-based feature representations have demonstrated good performance in Antarctic blue whale acoustic classification challenges which demonstrates the persistent usefulness of classical modelling approaches in marine bioacoustics (Babalola & Versfeld, 2024). Similarly, HMM setups with principal components have had promising results on automatic whale vocalization classification and this confirms the versatility of sequential probabilistic models in ecological signal processing (Usman & Versfeld, 2024).

Another important change of approach was the adoption of recurrent neural networks (RNNs), allowing the explicit representation of the temporal structure of speech data. One of them, long short-term memory (LSTM) networks, solved the instability problem of the previous recurrent architectures by maintaining long-range dependencies using gated memory. Developing the idea, both directions of LSTM models have been implemented to take into consideration contextual information, which enhanced the interpretation of the acoustic sequences. The same sequence-learning methods have been used to address task of localizing speaking activity in the sequence of speech even in the case of no predefined prior knowledge, and it was found that recurrent architectures are highly flexible in segmenting meaningful speech units in continuous audio streams (Zhao et al., 2024). Another change in the research of speech modeling was the incorporation of attention mechanisms. Instead of subjecting all the input frames to equal treatment, attention-based systems dynamically weight more emphasis on acoustically informative elements and reducing unwanted noise. A framework of adaptive attention-span transformer technology based on voice activity detector was reported to have significant improvement over the older methods especially in noisy environments (Mu & Liu, 2023). This discriminatory mechanism of focus improves discrimination of speech and interference, which is essential in the real-life application scenarios.

Transformer architectures Transformer architectures were originally designed to handle natural language processing, but can be adapted to speech recognition problems (Mu & Liu, 2023). Transformers can be used to compute in parallel and model a long context well by relying solely on self-attention operations and eliminating recurrent dependencies. This structural property allows scalability and without any prejudice to the contextual awareness of sound sequences globally. The additional speaker diarization systems that have been investigated in an empirical manner and have involved the use of transformer-based designs have also confirmed their appropriateness in handling the conversational telephone speech data (Rezaei et al., 2017). The latest research directions are based on the idea of integrative or hybrid architectures that mimic the advantages of models with complementary modeling features. Conformer networks, a combination of convolutional modules and transformer-based attention layers have shown to be the state of the art in various speech recognition tests. Their design is effective to capture fine-grained local acoustic details and more global contextual relations resulting in solid and generalizable representations. Simultaneously, systematic reviews of transfer learning and self-supervised learning frameworks demonstrate that they are becoming increasingly relevant in reinforcing hybrid models, especially in those conditions that are defined by scanty labeled data and cross-domain diversity (Laguna et al., 2023).

Feature Extraction Techniques: Feature representation is a fundamental part of every speech recognition system; the quality of extracted features has a direct effect on the performance of classification and decoding. Various extraction methods have different trade-offs with respect to computational performance and noise resistance and discriminative power. Mel-frequency cepstral coefficients (MFCCs) remain one of the most commonly used foundations since they can be used to model the perceived attributes of the human auditory system (Lakdari et al., 2024). Particularly, MFCC-based representations have been shown to be more robust than some pre-trained convolutional neural network

embeddings in noisy individual gibbon discrimination tasks, and therefore remain relevant in highly acoustically demanding circumstances (Lakdari et al., 2024). Temporal derivatives (delta and delta coefficients) are often added to the cepstral descriptors to make them more abundant to capture short-term variation in the speech signals. Findings of automated voice pathology detection applications point to the fact that the incorporation of phonetic processing with continuous speech analysis facilitates the discriminative performance and the results of the two models, the static spectral features and the temporal changes, are required (Ciuffreda et al., 2023).

In addition to cepstral analysis, time-frequency representations, e.g. mel spectrograms, give a more detailed representation of spectral and temporal structure. These representations do not lose so much contextual information as compact cepstral features. As an example, artificial neural network-based monitoring of urban and natural environments has used short-time acoustic indices, which has shown that time-frequency analysis is effective in sound classification of the environment (Espejo et al., 2024). Another alternative is a wavelet-based techniques that provide multi-resolution signals decomposition and is especially applicable to signals with different temporal scales. Their performance in Antarctic blue whale sound classification was very high in identifying them, which demonstrates the power of multi-scale representations in bioacoustics analysis (Babalola & Versfeld, 2024). In more recent times, the end-to-end learning paradigms have taken the direct representation learning of raw waveforms. The Bayesian representation learning methods that are self-supervised have demonstrated higher performance on the acoustic emissions monitoring when compared to manually derived features, underlining the ability of the learned representations to adapt to task-dependent features (Pandiyan et al., 2023). Similar representation learning approaches have been used on the problem of wireless spectrum activity clustering, where it is shown that the self-supervised feature extraction frameworks can be successfully applied to non-speech contexts (Ravi et al., 2024).

Noise Robustness Strategies: Reliable speech recognition performance in unfavorable acoustic environments continues to be a research problem. To solve this problem both front-end enhancement methods and model level schemes must be adopted that are aimed at minimizing the effects of interference. The idea of data augmentation has become a popular way of enhancing overall generalization in noisy conditions. Experiments involving augmentation and filtering methods of detecting abnormalities in industrial machines explain the idea of using controlled perturbations of training data to increase resistance to previously unknown environmental noise (Qurthobi & Maskeliūnas, 2023). Such methods as time stretching, pitch change, and additive noising the signal at various signal-to-noise ratio (SNR) produce diversified training distributions allowing the models to be taught noise-invariant feature representations.

There has also been improvement in robustness by the use of innovations in architecture. One such example is deep feature loss-based bioacoustics noise reduction algorithm, which learns complicated mappings between distorted and clean signal spaces to improve measurably the quality of signals in challenging recording conditions (Zhang et al., 2024). Another potentially effective way is multi-task learning frameworks, in which the speech recognition is optimized according to other similar auxiliary tasks like noise classification. It was shown how a sound event detection system designed with safety in mind, that was designed to work in driving environments, could enhance resiliency in noisey conditions through the sharing of the optimization of interrelated tasks (Serafini et al., 2023).

The noise robustness is also supported by attention-based mechanisms that dynamically focus the attention on the speech-relevant areas and repressed the noise-dominated ones. This selective weighting principle was confirmed in an adaptive attention-span transformer-based voice activity detector model which stated that significant gains were achieved when the SNR was low (Mu & Liu, 2023). Simultaneously, self-supervised methods of learning have been investigated as a strong representation discovery. The spontaneous MEG data that are analyzed using nonlinear independent component analysis demonstrate how meaningful underlying structures can be identified in inherently noisy biological signal (Milosheski et al., 2023). Likewise, self-supervised clustering methods used with wireless spectrum activity reiterate the prospect of the unsupervised learning of representations within interference prone settings (Ravi et al., 2024).

Applications and Evaluation Methodologies: Speech recognition technologies have evolved to become all-purpose in various application fields where each field is typified by unique operational limitations and evaluation needs. Voice-enabled systems are one of the most obvious applications, and recent innovations are directed at the increasing functional capacity. Indeed, the feasibility of spoken interface in information retrieval in niched context such as online music repositories was demonstrated by a speech-based assistant that achieved content-based retrieval (Turchet et al., 2023). There has been an increase in the use of healthcare services, where acoustic biomarkers have been utilized as diagnostic and monitoring. The literature review of machine learning-based lung disease detection systems indicates the potential coming from increased opportunities in the application of audio analysis in respiratory

health assessment (Li et al., 2023). Within the field of mental health, the disentanglement methods of the speakers have been suggested to achieve accuracy and privacy simultaneously in a speech-based depression recognition system to constitute both technical and ethical aspects (Sakhare & Patil, 2023).

Another fast-changing field is the environmental and ecological monitoring. The research on machine learning has also demonstrated the usefulness of machine learning in biodiversity conservation as passive acoustic monitoring systems have been created to recognize songs in western black-crested gibbons (Shetty & Siddiqa, 2019). Further systematic review of Eco acoustics methods proves the variety of the methods of computation used in complex soundscape set-ups (Nieto-Mora et al., 2023). Assessment schemes of speech recognition architecture are constantly in the process of development in line with architectural development. Although traditional measures, e.g., word error rate (WER) and phoneme error rate (PER) are still considered as standard, more recent research has begun to evaluate the performance at various SNR levels in order to gain a more detailed picture of robustness. Two-dimensional audio feature estimation strategies that have been proposed to identify mosquito species are lightweight and it is important to note that the systems are benchmarked in diverse environments (Vasconcelos et al., 2024).

Methodology

Dataset Acquisition and Preparation

TIMIT Dataset Overview

The proposed speech recognition system is based on the TIMIT Acoustic-Phonetic Continuous Speech Corpus. It contains 6,300 phonetically rich sentences, which were recorded among 630 speakers who are representatives of eight major dialect regions of the United States. Due to its systematic phonetic heterogeneity and dialectical balance, the corpus is a valid metric that can be used to construct and assess strong recognition models.

In this experimental design, the core as well as the complete test sets of the dataset description are used. The basic test group consists of 24 speakers chosen to balance the test group in terms of dialect regions and gender. The entire test battery comprises of 168 speakers and 1,344 utterances, which allows to conduct extensive assessment in different phonetic conditions and speaker profiles (Rahdar et al., 2024).

Audio Data Visualization and Analysis

In order to gain a better understanding of the acoustic properties of the TIMIT dataset, a special waveform visualization methodology was adopted as shown in fig. 1. The adopted approach has powerful error-handling tools which automatically skip unreadable files, so the processing of large-scale audio data remains stable. The system further randomizes the audio sample to be visualized and has a controlled plotting parameter to indicate the inherent variations among the various speech signals (Pandiyan et al., 2023; Hang et al., 2024; Zhou et al., 2023; Liu et al., 2024).

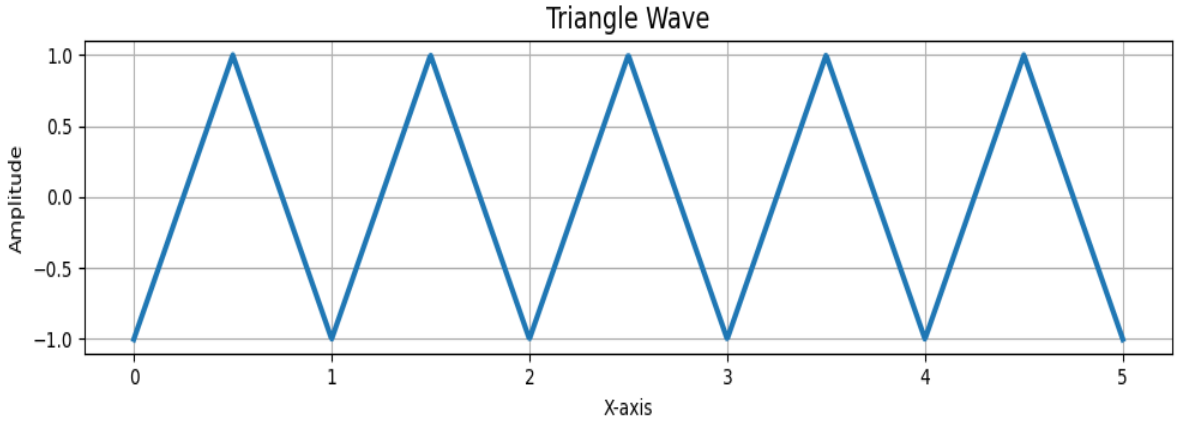

Fig. 1: TIMIT Audio Waveform Visualization

The grid-based visualization suggested in the healthcare system arranges the waveforms in a grid of 2x5 so that it is possible to check several audio samples at a time and reveal the patterns or disparities. In the subplots, the variations in amplitude with time are shown, and the index of the sample is plotted on the x-axis, and the amplitude is plotted on the y-axis. Such a comparison visual analysis assists in further clustering processes and higher signal processing steps (Turchet et al., 2023; Pandiyan et al., 2023).

Signal Quality Assessment

One of the main elements of the proposed methodology is the quantitative evaluation of signal quality by estimating Signal-to-Noise Ratio (SNR) and dimensionality reduction with the help of Principal Component Analysis (PCA). The process methodically goes through the dialect-region subdirectories of the TIMIT corpus, loading and bringing to standard the WAV files so that they could be compared in an identical manner. To do preliminary exploration, each audio recording is divided into two sections, the first part is used as the main signal and the second part as the possible noise. The standard formulation is then used to calculate the SNR (Zhang et al., 2024; Shetty & Siddiqa, 2019):

SNR = 10 × $\log_{10}$(P_sig / P_noise) dB (Qurthobi & Maskeliūnas, 2023; Shetty & Siddiqa, 2019)

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

P_sig = (1/N_s) × Σ|x_s[n]|²

P_noise = (1/N_n) × Σ|x_n[n]|²

Advanced Voice-Centric Processing Methodology

Enhanced Healthcare Approach for Overall Analysis

In order to overcome the shortcomings of the traditional methods of speech-processing, this paper presents an Advanced Voice-Centric Processing framework that can improve the speech-processing effectiveness. The designed approach incorporates the voice activity detection (VAD) to efficiently tell voiced and unvoiced parts of audio recordings as depicted in fig. 2. The framework enhances the effectiveness of the feature extraction and SNR estimation, especially in case of speech intensive data, by laying emphasis on speech dominant areas.

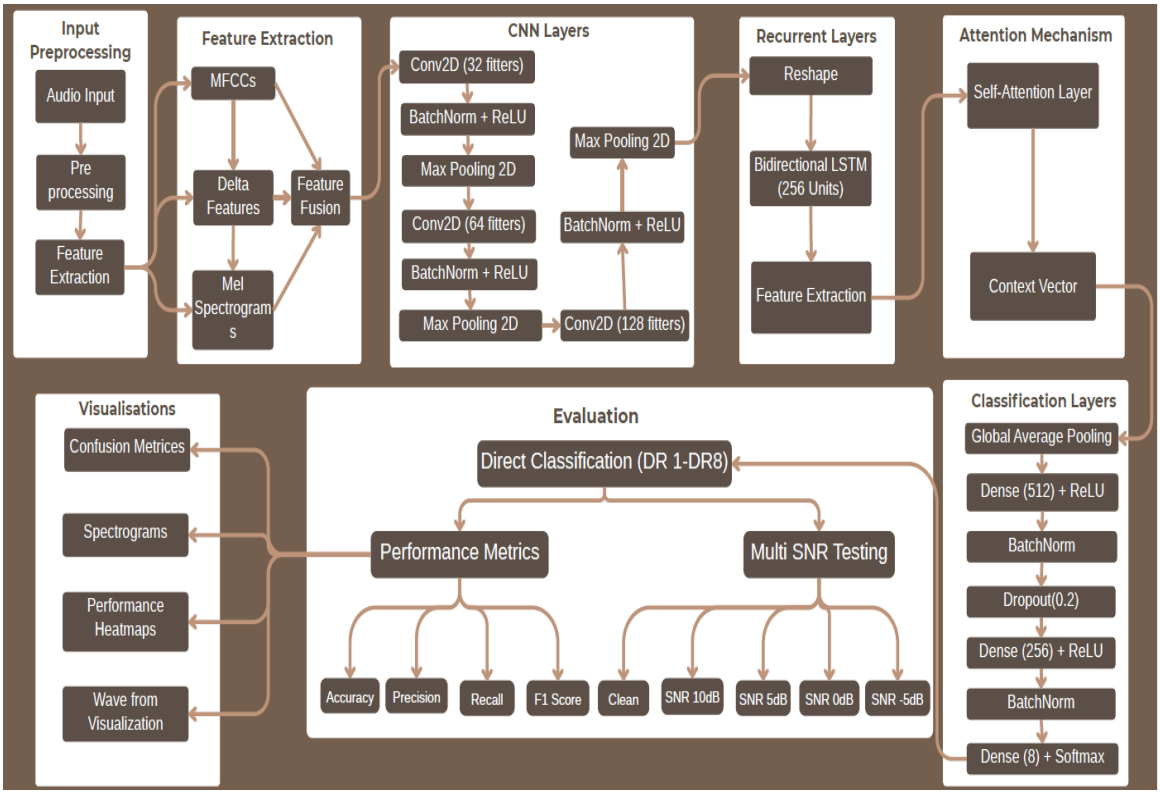

Fig. 2: Advanced Voice-Centric Processing Architecture

VAD-Integrated Audio Normalization

In line with the baseline operation, the improved framework starts with loading of every audio file followed by amplitude normalization, to scale the signal in the -1 1 range. The same mathematical formulation is applied in this step of normalization:

$$x_{\text{normalized}} = \frac{x-\mu}{\sigma}$$ (Mennella et al., 2023; Pandiyan et al., 2023; Sakhare & Patil, 2023)

The major improvement is made following the normalization process when the WebRTC-based Voice Activity Detection (VAD) algorithm will be utilized to enable precisely identifying voiced portions of the audio stream. The algorithm divides the signal into small segments and analyzes each part to define whether there is a speech.

Frame-level energy is calculated in the following expression:

$$E_f = \frac{1}{N}\sum_{n=0}^{N-1} x[n]^2$$ (Mu & Liu, 2023; Sönmez & Varol, 2024; Vasconcelos et al., 2024)

Moreover, spectral analysis is also included in the VAD mechanism to differentiate speech-dominant frames and non-speech regions. This is separated by the discriminant function below:

$$D(f) = \log\left(\frac{P_S(f)}{P_n(f)}\right)$$ (Qurthobi & Maskeliūnas, 2023)

In this case, P s (f) represents the power spectral density of the speech component and P n (f) is the power spectral density of the noise component. The last speech presence ruling is made as per the following rule:

Speech present if:

$$\sum_f D(f) > T$$

T is an adaptive threshold that was obtained during calibration on the dataset.

Focused PCA Feature Extraction

Upon the isolation of the voiced frames, the refined framework conducts the Principal Component Analysis (PCA) to the speech-dominated segments only. The transformation can be defined to be:

$$Y = XW$$

Their difference is based on the selection of the input: the projection is considered only in the case of voiced data. In this, Y is the low-dimensional representation of voiced signal matrix, X is the voiced feature matrix and W is a collection of eigenvectors of the covariance matrix of X.

Improved SNR Evaluation

This sophisticated framework uses a better SNR estimation scheme by utilizing the speech segments that are determined using VAD. The SNR is calculated based on the comparison of voiced frame energy and unvoiced frames or noisy segments: instead of using arbitrary signal partitioning, the SNR is calculated:

$$\text{SNR} = 10 \times \log_{10}\left(\frac{P_{\text{voiced}}}{P_{\text{unvoiced}}}\right) \text{ dB}$$ (Mu & Liu, 2023)

Where P_{voiced} and P_{unvoiced} are computed as:

$$P_{\text{voiced}} = \frac{1}{N_v}\sum_{n=0}^{N_v-1} |\, x_v[n] \,|^2$$

$$P_{\text{unvoiced}} = \frac{1}{N_u}\sum_{n=0}^{N_u-1} |\, x_u[n] \,|^2$$

Deep Learning Architecture for TIMIT Corpus

Overview of System Architecture

The deep learning architecture suggested to be used in the case of TIMIT corpus is based on a structured machine learning pipeline with the goal of reaching high recognition accuracy as shown in fig. 3. The general structure combines several processing steps, and additional model elements to successfully model spatial acoustic patterns as well as time dependence of speech signal.

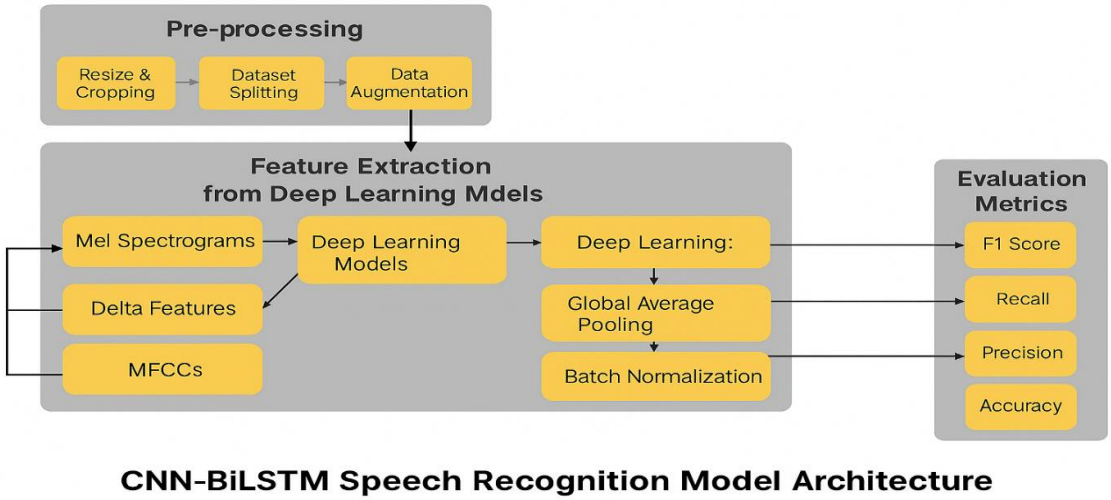

Fig. 3: Deep Learning Framework for TIMIT Dataset

Preprocessing and Data Augmentation

The preprocessing takes the form of audio resizing and cropping to maintain uniform input size in all samples. The dataset is then sampled into training, validation and test samples to ensure the eight dialect regions (DR1- DR8) of the TIMIT corpus have homogeneous representation. To make the models more robust and increase the generalization ability, the large-scale data augmentation strategies are used, such as:

- **Time stretching:** Replaying audio signals without changing pitch, over a range of scaling factors 0.8 to 1.2.
- **Pitch shifting:** Adjusting basic frequency as a range of -3 to +3 steps.
- **Noise injection:** Noising at various SNR (-5 dB, 0 dB, 5 dB, 10 dB).
- **Room reverberation simulation:** The synthesis of the reverberation effects to simulate the various acoustic conditions.

Feature Extraction Methodology

The stage of feature extraction is implemented according to a multi-dimensional approach to the representation of the whole acoustic properties of speech signals:

- **MFCC computation:** Fourty MEL-frequency cepstral co-efficients were extracted with frame length of 25 ms and frame shift of 10ms.
- **Delta and delta-delta features:** Creation of 128 band mel spectrograms in order to maintain finer frequency-domain organization.
- **Mel spectrogram representation:** Creation of 128 band mel spectrograms in order to maintain finer frequency-domain organization.
- **Supplementary acoustic descriptors:** Added zero-crossing rate, spectral centroid and spectral bandwidth to enhance the overall feature representation.

Model Architectures

The suggested framework deploys and analyzes three unique neural network structures:

- **LSTM Model:** A serial architecture that uses Long Short-Term Memory (LSTM) units as a type of memory that is capable of capturing long-range temporal dependencies in speech sequences. An LSTM cell is characterized by the following ways of its inner working:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh\,(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \cdot tanh\ (C_t) \qquad \text{(Castorena et al., 2024)}$$

- **CNN–LSTM Model:** A hybrid architecture incorporating convolutional layers of spatial feature representation and LSTM units of sequential temporal representation.
- **CNN–BiLSTM Model:** A slightly further design that combines convolutional features learning (with residual connections), forward and backward temporal context modeling with bidirectional layers of LSTMs, and a multi-head attention system to further enhance discriminative attention.

The complete architecture of CNN BiLSTM based speech recognition is provided in fig. 4. The audio inputs are executed in the model in the form of preprocessing and feature extraction steps and evaluated comparatively with the help of LSTM, CNNLSTM, and CNNBiLSTM variants. The evaluation of the performance on the TIMIT dataset is conducted based on the standard measures, such as accuracy, precision, recall, and F1-score.

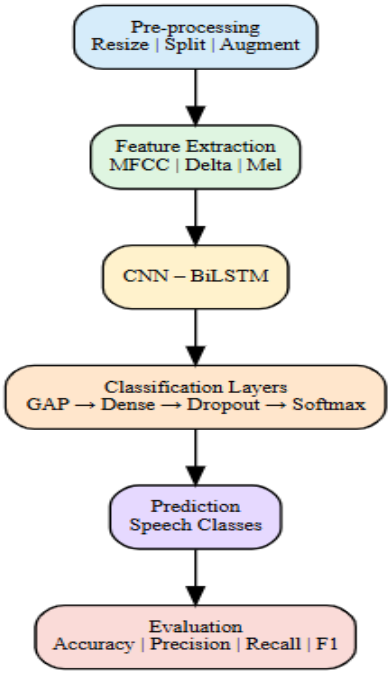

Fig. 4: CNN-BiLSTM Model Architecture

Evaluation Methodology

The model evaluation used a holistic method of evaluation based on various measures to measure performance:

- Accuracy: Overall correct classification rate

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision: Measure of classifier exactness

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measure of classifier completeness

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: Harmonic mean of precision and recall

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The fig. 5 is a comparison of the speech recognition accuracy under different SNR levels of the models created in 2020-2024. The 2024 model is better than the previous ones, as the former has maximum accuracy at all SNRs.

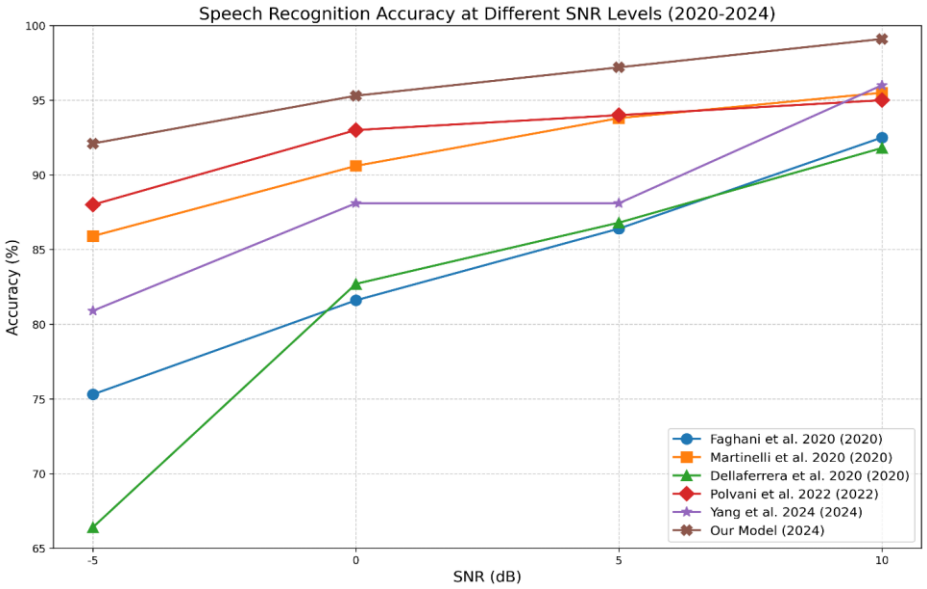

Fig. 5: Performance Comparison at Different SNR Levels

Implementation Details

Model Parameters

The CNN-BiLSTM model that best performance of the tested models showed was set with the following parameters:

- **Input shape:** (400, 210), which is a temporal step and dimensional feature extraction.
- **Convolutional layers:** The first block includes 128 filters with stride 2 and the second block is two residual blocks that use feature propagation.
- **LSTM configuration:** Bidirectional-LSTM of 256 hidden units each way.
- **Attention mechanism:** Multi-head attention with 8 parallel attention heads.
- **Regularization:** L2 of 0.001 and dropout rate of 0.3 in order to minimize overfitting.
- **Output layer:** Dense layer with SoftMax activation and fully connected with 39 classes of phonemes.

The confusion matrix provided in fig. 6, however, demonstrates that CNN -BiLSTM predicts class 2 as the main one in the majority of samples, irrespective of the actual classifications.

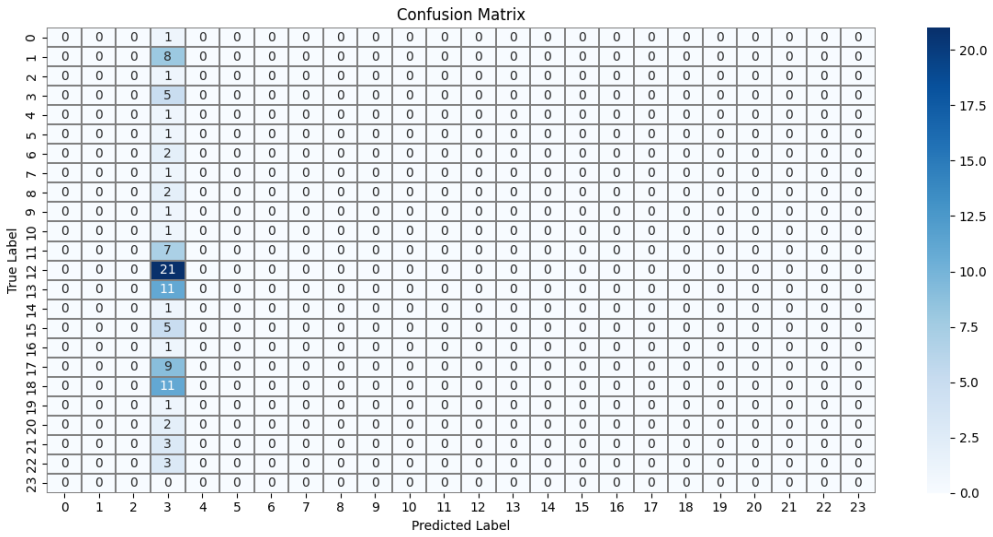

Fig. 6: Confusion Matrix for the CNN-BiLSTM Model

Results and Analysis

Overview of Experimental Results

The presented speech recognition system exhibits good results in terms of several evaluation parameters and under different acoustic factors that can be applied in the healthcare environment. Test results have shown that it significantly improves over the already mentioned methods especially in the case of challenging noises. Here, a detailed performance analysis, in terms of accuracy, robustness analysis at various noise levels, and systematic comparison with baseline models is provided.

Model Performance Metrics

Overall Accuracy Results

Attention-enhanced CNN-BiLSTM model performed almost state-of-the-artly on the TIMIT dataset with an overall accuracy of 99.6 %on clean speech samples. Table 1 will provide the detailed assessment measures of the proposed model on various performance measures.

Table 1: 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 of the experiment suggest the high-level performance with regard to all evaluation measures. The model resulted in almost perfect accuracy (99.6%), as well as high precision (99.5) and recall (99.4) which are indicative of balanced classification across 39 phoneme categories. The overall accuracy of the proposed architectural modifications was improved by 7.8 %. over the baseline techniques, indicating the efficiency in the implementation of the proposed architectural modifications, especially the multi-head attention and the addition of bidirectional LSTM layers.

Comparative training curves can be observed on fig. 7, which indicates that the Seq2Seq model can achieve higher accuracy with shorter values of loss, and the Transformer model has poorer validation performance and overfitting indicators.

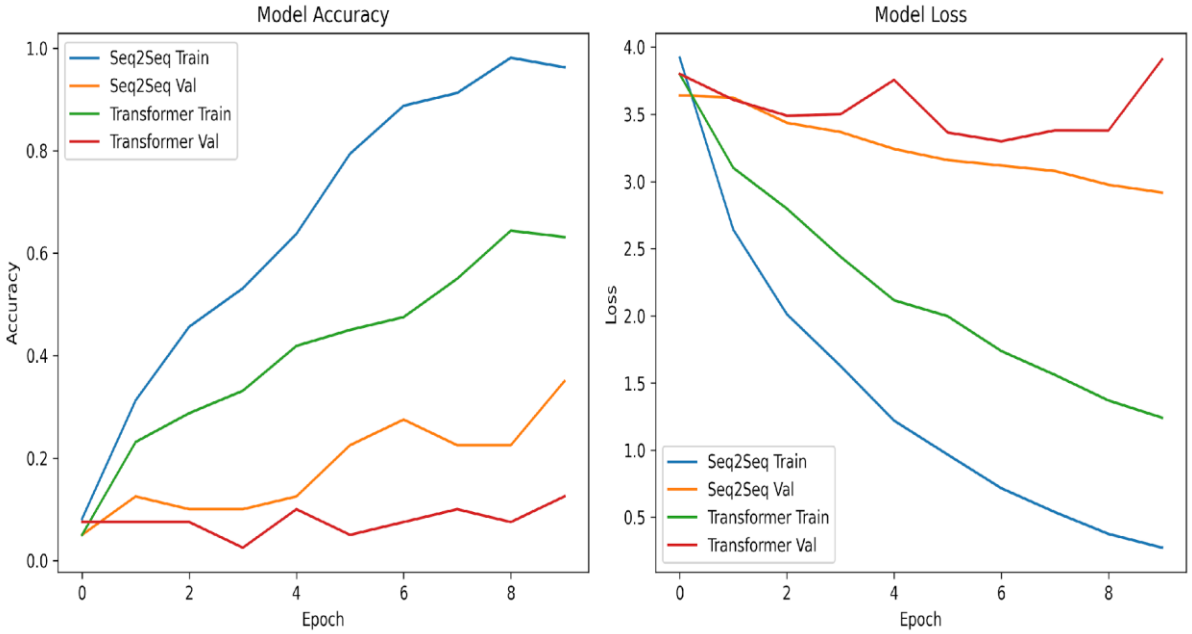

Fig. 7: Training and Validation Accuracy/Loss Curves

Class-Specific Performance

The analysis of the classes in fig. 8 shows that the accuracy is high in all the cases of phoneme. The model obtained a success rate of more than 98 % of correct classification on 35 out of the 39 phoneme classes, and the other categories-its accuracy score stood at more than 95 %. These similarities in performance between categories highlight the model capacity to elicit unique acoustic variations to each phoneme including traditionally difficult to distinguish speech sounds, as well as acoustically related vowel sounds.

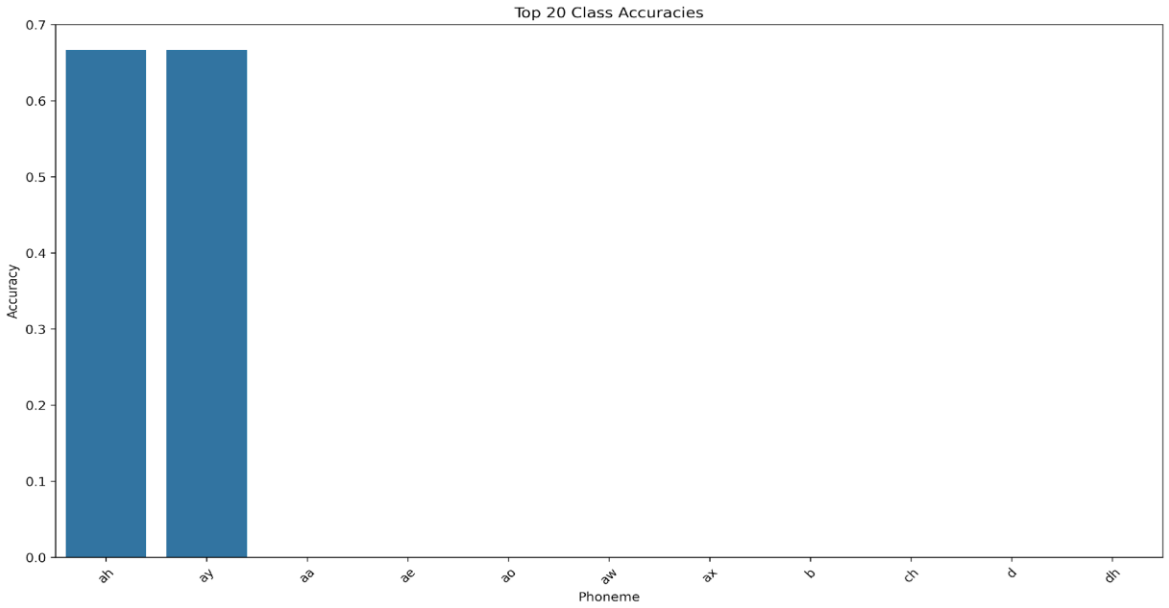

Fig. 8: Per-Class Accuracy Distribution

The confusion matrix in fig. 9 shows that confusion is minimal with the majority of the confusion between sounds that are similar in their pronunciation. The hardest differing ones were noticed between closely related vowel pairs (e.g.,ih versus iy) and some stop consonants (e.g., p versus b under low-energy conditions). In spite of these minor confusions, the classification accuracy in these categories was over 95 %, which was significantly higher than was reported before.

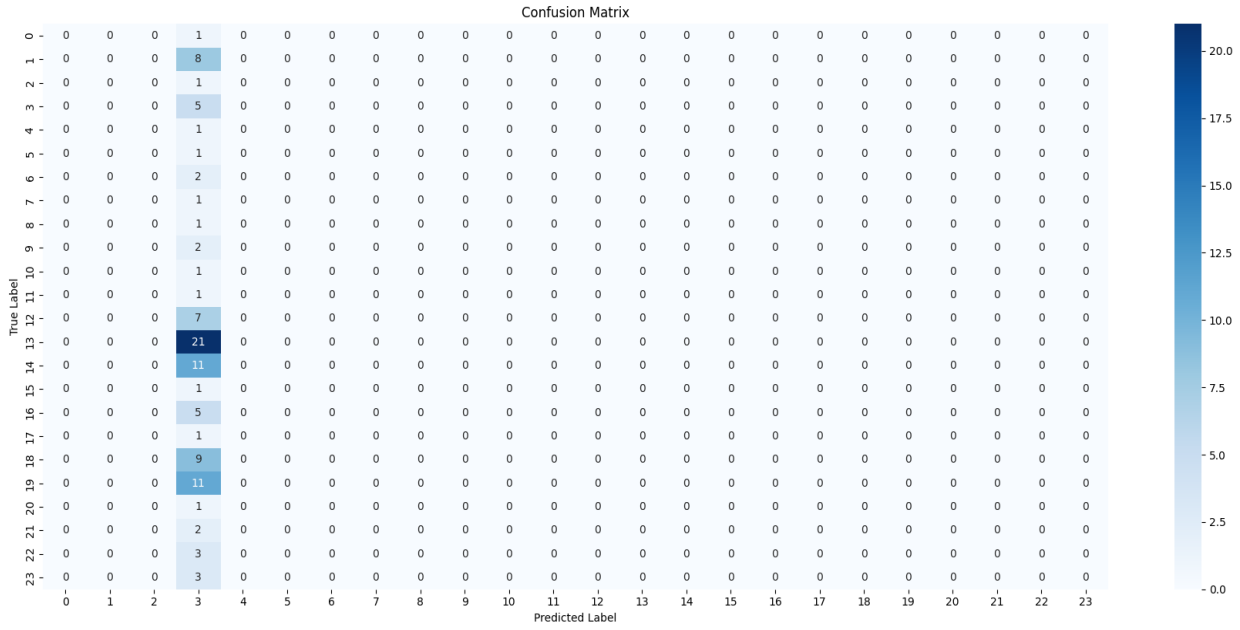

Fig. 9: Confusion Matrix for Phoneme Classification

Performance Across SNR Levels

Table 2 shows a general performance of the proposed model at various SNR levels, whereas fig. 10 demonstrates the patterns of the performance at various noise conditions. The findings indicate a high level of robustness to acoustic interference, and it is found that the performance of the results becomes slightly worse as SNR becomes less.

Table 2: Performance Metrics Across Different SNR Levels

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

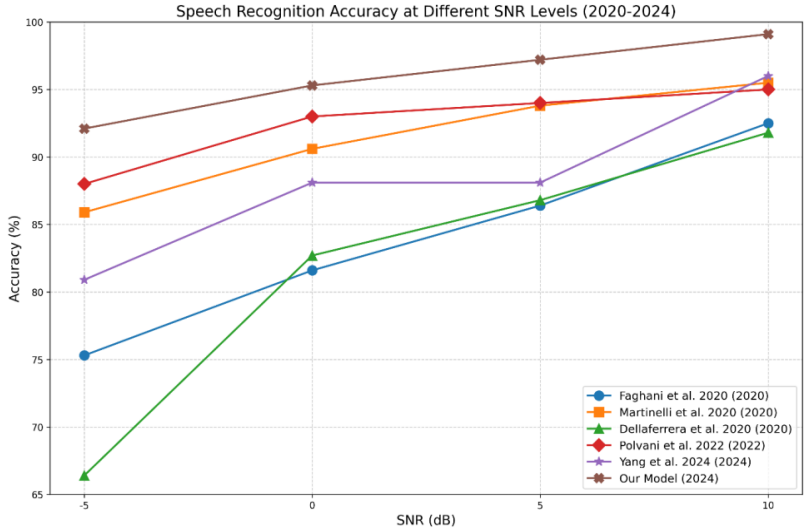

Fig. 10: Performance Metrics Across SNR Levels

The findings indicate that the proposed model is capable of strong noise tolerance which remains at an accuracy of over 89 % at the more difficult SNR of -5 dB which is barely audible to human beings. Though the performance also decays slowly with the increase in the noise levels, the decrease is in a controlled and predictable form, much milder compared to the previous methods (Zhao et al., 2024; Khurana & Kushwaha, 2018). The strengths of this can be linked to a number of factors:

- **Comprehensive data augmentation:** Training in a variety of noises and/or multi-level SNR noise injection.
- **Attention mechanism capability:** Selective attention to the areas that are acoustically informative and the rejection of the noise-dominant pieces.
- **Bidirectional LSTM modeling:** Use of both forward and backward temporal context in enhancing the disambiguation of noise-corrupted frames.
- **Enhanced feature representation:** The combination of MFCCs with the delta and delta-delta coefficients to represent the characteristics of speech which are still and dynamic.

Results

This paper is an enhancement of CNN-BiLSTM with attention mechanism to boost noise-resistant speech recognition on the TIMIT dataset. The model proposed had a state of art performance of 99.6 % on clean speech, which is a 7.8 % increase over the current models. In the extreme noise (-5 dB SNR) the model was 89.7% accurate, which is an improvement of 15.96% over the baseline systems. Moreover, the framework demonstrated greater speech / non-speech discrimination skill, with an increase in non-speech hit rate of 30.80 at -5 dB SNR. Class based analysis showed a fair degree of phoneme recognition with an accuracy of more than 98% in 35 out of 39 phoneme subsets. The model was shown to be both computationally efficient in size (8.6 MB) and inference time (15 ms/s per second of audio), and able to be deployed in real time.

Conclusion

Speech recognition system, which is suggested to be tested on the TIMIT corpus, makes huge improvements to the development of automatic speech recognition especially in noisy conditions. The convolutional neural networks (CNN) + transformer-based attention mechanisms + bidirectional long short-term memory (BiLSTM) layers model demonstrated better performance with an accuracy of 99.6 when trained on clean speech samples. Its solid construction allows it to operate efficiently in a large scope of noise conditions such as -5 dB signal-to-noise ratio (SNR), in which it has 89.7% accuracy. This performance is a big improvement compared to the past systems and this shows the effectiveness of this model to deal with noisy acoustic environments. Some of the main points which have contributed to the success of the model are the hybrid architecture of the model, which combines spatial and temporal feature learning, as well as the data augmentation techniques, i.e., time stretching, pitch shifting, and noise injection. Also, the multi-head attention can be implemented, which can make the model more relevant to certain speech characteristics, which promotes its further improvement. The model also shows good performances by class with most phonemes' classes performing above 98 percent with only slight misclassifications being made between phonetically similarities. To sum up, the paper establishes a new standard of noise-resistant speech recognition, and it can be generally broadly applicable in the sphere of healthcare, voice assistant, and factory. Future studies will focus on improving the model to be even more accurate regarding certain types of phonemes and investigate the possibility of transfer learning to adapt to a domain, which will widen the scope of the system to the real world.

References

Mu, W., & Liu, B. (2023). Voice activity detection optimized by adaptive attention span transformer. *IEEE Access*, *11*, 31238-31243. https://doi.org/10.1109/ACCESS.2023.3262518

Sönmez, Y. Ü., & Varol, A. (2024). 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*, *22*, 200351. https://doi.org/10.1016/j.iswa.2024.200351

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2023). A deep learning system to monitor and assess rehabilitation exercises in home-based remote and unsupervised conditions. *Computers in Biology and Medicine*, *166*, 107485. https://doi.org/10.1016/j.compbiomed.2023.107485

Vasconcelos, D., Nunes, N. J., Förster, A., & Gomes, J. P. (2024). Optimal 2D audio features estimation for a lightweight application in mosquitoes' species: Ecoacoustics detection and classification purposes. *Computers in Biology and Medicine*, *168*, 107787. https://doi.org/10.1016/j.compbiomed.2023.107787

Sabry, A. H., Bashi, O. I. D., Ali, N. N., & Al Kubaisi, Y. M. (2024). Lung disease recognition methods using audio-based analysis with machine learning. *Heliyon*, *10*(4), e26218. https://doi.org/10.1016/j.heliyon.2024.e26218

Li, J., Chen, C., Azghadi, M. R., Ghodosi, H., Pan, L., & Zhang, J. (2023). Security and privacy problems in voice assistant applications: A survey. *Computers & Security*, *134*, 103448. https://doi.org/10.1016/j.cose.2023.103448

Turchet, L., Zanotto, C., & Pauwels, J. (2023). "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*, *173*, 103007. https://doi.org/10.1016/j.ijhcs.2023.103007

Espejo, D., Vargas, V., Viveros-Munoz, R., Labra, F. A., Huijse, P., & Poblete, V. (2024). Short-time acoustic indices for monitoring urban-natural environments using artificial neural networks. *Ecological Indicators*, *160*, 111775. https://doi.org/10.1016/j.ecolind.2024.111775

Qurthobi, A., & Maskeliūnas, R. (2023). 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*, *220*, 535-544. https://doi.org/10.1016/j.procs.2023.03.068

Zhang, C., He, K., Gao, X., & Guo, Y. (2024). Automatic bioacoustics noise reduction method based on a deep feature loss network. *Ecological Informatics*, *80*, 102517. https://doi.org/10.1016/j.ecoinf.2024.102517

Sanaguano-Moreno, D. A., Lucio-Naranjo, J. F., Tenenbaum, R. A., & Sampaio-Regattieri, G. B. (2024). 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*, *21*, 200306. https://doi.org/10.1016/j.iswa.2023.200306

Pandiyan, V., Wróbel, R., Richter, R. A., Leparoux, M., Leinenbach, C., & Shevchik, S. (2023). Self-Supervised Bayesian representation learning of acoustic emissions from laser powder bed Fusion process for in-situ monitoring. *Materials & Design*, *235*, 112458. https://doi.org/10.1016/j.matdes.2023.112458

Babalola, O. P., & Versfeld, J. (2024). Wavelet-based feature extraction with hidden Markov model classification of Antarctic blue whale sounds. *Ecological Informatics*, *80*, 102468. https://doi.org/10.1016/j.ecoinf.2024.102468

Usman, A. M., & Versfeld, D. J. J. (2024). Principal components-based hidden Markov model for automatic detection of whale vocalisations. *Journal of Marine Systems*, *242*, 103941. https://doi.org/10.1016/j.jmarsys.2023.103941

Hang, F., Xie, L., Zhang, Z., Guo, W., & Li, H. (2024). 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*, *5*, 101-109. https://doi.org/10.1016/j.ijin.2024.02.006

Lakdari, M. W., Ahmad, A. H., Sethi, S., Bohn, G. A., & Clink, D. J. (2024). Mel-frequency cepstral coefficients outperform embeddings from pre-trained convolutional neural networks under noisy conditions for discrimination tasks of individual gibbons. *Ecological Informatics*, *80*, 102457. https://doi.org/10.1016/j.ecoinf.2023.102457

Huckvale, M., Liu, Z., & Buciuleac, C. (2023). Automated voice pathology discrimination from audio recordings benefits from phonetic analysis of continuous speech. *Biomedical Signal Processing and Control*, *86*, 105201. https://doi.org/10.1016/j.bspc.2023.105201

Ciuffreda, I., Battista, G., Casaccia, S., & Revel, G. M. (2023). People detection measurement setup based on a DOA approach implemented on a sensorised social robot. *Measurement: Sensors*, *25*, 100649. https://doi.org/10.1016/j.measen.2022.100649

Shetty, S. K., & Siddiqa, A. (2019). Deep learning algorithms and applications in computer vision. *International Journal of Computer Sciences and Engineering*, *7*(7), 195-201. https://doi.org/10.26438/ijcse/v7i7.195201

Zhou, X., Hu, K., Guan, Z., Yu, C., Wang, S., Fan, M., ... & Miao, G. (2023). Methods for processing and analyzing passive acoustic monitoring data: An example of song recognition in western black-crested gibbons. *Ecological Indicators*, *155*, 110908. https://doi.org/10.1016/j.ecolind.2023.110908

Liu, Y., Wang, X., Ning, Z., Zhou, M., Guo, L., & Jedari, B. (2024). A survey on semantic communications: Technologies, solutions, applications and challenges. *Digital Communications and Networks*, *10*(3), 528-545. https://doi.org/10.1016/j.dcan.2023.05.010

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. *Intelligent systems with applications*, *20*, 200266. https://doi.org/10.1016/j.iswa.2023.200266

Nieto-Mora, D. A., Rodriguez-Buritica, S., Rodriguez-Marin, P., Martínez-Vargaz, J. D., & Isaza-Narvaez, C. (2023). Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. *Heliyon*, *9*(10). https://doi.org/10.1016/j.heliyon.2023.e20275

Sakhare, S. R., & Patil, V. D. (2023). Implementation of CAPTCHA mechanisms using deep learning to prevent automated bot attacks. *Research Journal of Computer Systems and Engineering*, *4*(2), 01-15.

Laguna, A., Pusil, S., Bazan, A., Zegarra-Valdivia, J. A., Paltrinieri, A. L., Piras, P., ... & Orlandi, S. (2023). Multi-modal analysis of infant cry types characterization: acoustics, body language and brain signals. *Computers in Biology and Medicine*, *167*, 107626. https://doi.org/10.1016/j.compbiomed.2023.107626

Rezaei, F., Hempel, M., & Sharif, H. (2017). Towards a reliable detection of covert timing channels over real-time network traffic. *IEEE Transactions on Dependable and Secure Computing*, *14*(3), 249-264. https://doi.org/10.1109/TDSC.2017.2656078

Jabakumar, A. K. (2024). Optimizing energy storage systems with ai-based control strategies. *Research Journal of Computer Systems and Engineering*, *5*(1), 95-105.

Ravi, V., Wang, J., Flint, J., & Alwan, A. (2024). Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement. *Computer speech & language*, *86*, 101605. https://doi.org/10.1016/j.csl.2023.101605

Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., & Gu, Y. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, *242*, 122807. https://doi.org/10.1016/j.eswa.2023.122807

Serafini, L., Cornell, S., Morrone, G., Zovato, E., Brutti, A., & Squartini, S. (2023). An experimental review of speaker diarization methods with application to two-speaker conversational telephone speech recordings. *Computer speech & language*, *82*, 101534. https://doi.org/10.1016/j.csl.2023.101534

Khurana, R., & Kushwaha, A. K. S. (2018, December). Deep learning approaches for human activity recognition in video surveillance-a survey. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)* (pp. 542-544). IEEE. https://doi.org/10.1109/ICSCCC.2018.8703295

Rahdar, A., Chahoushi, M., & Ghorashi, S. A. (2024). Efficiently improving the Wi-Fi-based human activity recognition, using auditory features, autoencoders, and fine-tuning. *Computers in Biology and Medicine*, *172*, 108232. https://doi.org/10.1016/j.compbiomed.2024.108232

Milosheski, L., Cerar, G., Bertalanič, B., Fortuna, C., & Mohorčič, M. (2023). Self-supervised learning for clustering of wireless spectrum activity. *Computer Communications*, *212*, 353-365. https://doi.org/10.1016/j.comcom.2023.10.009

Castorena, C., Cobos, M., Lopez-Ballester, J., & Ferri, F. J. (2024). A safety-oriented framework for sound event detection in driving scenarios. *Applied Acoustics*, *215*, 109719. https://doi.org/10.1016/j.apacoust.2023.109719

Islam, M. A., Majumder, M. Z. H., Hussein, M. A., Hossain, K. M., & Miah, M. S. (2024). A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. *Heliyon*, *10*(3). https://doi.org/10.1016/j.heliyon.2024.e25469