Performance Evaluation of DNN Embedding–based Speaker Identification System

[1]Bittu Kumar, [2]Pranav Madanu, [3]Rohan Madanu,
[4]MD Sameer Ahmed, [5]Rizwan, [6]A B Nitin

Department of Electronics and Communication Engineering,
Koneru Lakshmaiah Education Foundation,
Hyderabad-500075, Telangana, India
[1]bittu.mlrit@gmail.com, [2]pranavmadanu@gmail.com,
[3]rohanmadanu@gmail.com, [4]2310040004@klh.edu.in,
[5]2310040027@klh.edu.in, [6]2310040030@klh.edu.in

Abstract: Speaker identification systems aim to recognize and authenticate individuals based on their distinct vocal characteristics, and recent advancements in deep neural networks (DNNs) have significantly improved their accuracy by enabling the extraction of robust, speaker-specific embeddings from speech data. In this study, a DNN-based speaker identification framework was developed and evaluated using an in-house speech corpus comprising 51 speakers representing diverse linguistic and regional backgrounds. The system employed a DNN embedding model to generate Probabilistic Linear Discriminant Analysis (PLDA) scores for speaker verification and identification. Experiments were conducted in a cloud-based environment using TensorFlow and Keras, with performance assessed in terms of Equal Error Rate (EER), accuracy, and loss across varying epoch configurations. Results indicated a steady improvement in model performance with increasing training epochs, as EER decreased from 9.5% to 6.2% and accuracy improved from 90.25% to 91.26%. The convergence of training and validation curves confirmed efficient learning dynamics, strong generalization capability, and the absence of overfitting. Overall, the proposed model achieved stable convergence and enhanced discriminative power, demonstrating its potential for reliable and scalable speaker identification applications in real-world voice authentication systems. This framework can be applied in real-world domains such as access control, banking authentication, call-center verification, and IoT voice interfaces. The study also highlights challenges in deploying speaker identification systems in noisy, multilingual, and channel-variable environments.

Keywords: Speaker Identification, DNN, Neural Networks, Speaker Recognition

1. Introduction

Speaker recognition systems aim to extract and model speaker-specific characteristics from a given speech signal in order to determine the identity of the speaker. Essentially, they recognize individuals through their voice signatures [1]. Each person's voice is inherently unique due to physiological differences such as the shape of the vocal tract, the size of the larynx, and other articulatory structures involved in speech production. In addition to these physical traits, behavioural factors such as accent, rhythm, intonation pattern, pronunciation style, and lexical choice further distinguish one speaker from

another. Modern speaker recognition systems leverage a combination of these acoustic and linguistic features to achieve high recognition accuracy [2].

In recent years, speaker identification has emerged as one of the most prominent biometric authentication techniques, capitalizing on the intrinsic characteristics of human speech. This technology is expected to play an increasingly vital role in secure access systems, financial transactions, and data privacy applications[3-4]. While several biometric modalities exist, such as fingerprints, iris, facial recognition, and handwritten signatures—the choice of biometric depends on factors including robustness, distinctiveness, user acceptability, acquisition cost, and suitability for remote authentication [5]. Historically, speaker recognition relied on statistical methods like the Gaussian Mixture Model–Universal Background Model (GMM-UBM) and i-vector frameworks [6]. However, recent advancements have shifted toward Deep Learning architectures, which provide superior performance in capturing high-level abstractions from speech signals.

In this study, a speaker identification system is implemented using a diverse in-house corpus of 51 speakers, comprising both direct recordings and audio extracted from YouTube videos. The proposed methodology employs a Deep Neural Network (DNN) embedding model to generate Probabilistic Linear Discriminant Analysis (PLDA) scores for identification. System performance is evaluated using Equal Error Rate (EER) and accuracy metrics. Speaker identification plays a vital role in applications such as secure access control, mobile banking authentication, personalized virtual assistants, forensics, call-center automation, and smart-home IoT systems. These applications benefit from the ability of modern DNN-based embeddings to provide robust identity verification even with short utterances.

The structure of this paper is organized as follows: Section 2 provides an overview of the proposed speaker identification framework, including the feature extraction, modeling, and evaluation methodologies employed. The results and their corresponding discussion are provided in Section 3. This section included the details about the corpus. Finally, Section 4 concludes the paper with a summary of the key findings and recommendations for future research directions

2. DNN-based Speaker Identification Framework

Recently, the exceptional feature learning capabilities of Deep Neural Networks (DNNs) have revolutionized speaker recognition. Inspired by their success in automatic speech recognition, numerous DNN-based approaches [9] have been proposed, achieving substantial performance improvements even under unconstrained, real-world acoustic conditions. The overall architecture of the proposed DNN-based speaker identification system is illustrated in Fig. 1. The process begins with audio chunk files, which are first pre-processed to enhance signal quality and eliminate background noise. Subsequently, Mel-Frequency Cepstral Coefficients (MFCCs) [10] are extracted to capture the spectral and temporal characteristics of the speech signal. These MFCC features are then passed through a Deep Neural Network (DNN) to obtain high-dimensional speaker embeddings. Finally, Probabilistic Linear Discriminant Analysis (PLDA) is applied to the embeddings to compute similarity scores, which are used for

identifying the corresponding speaker. Details of each component are discussed in upcoming subsections.

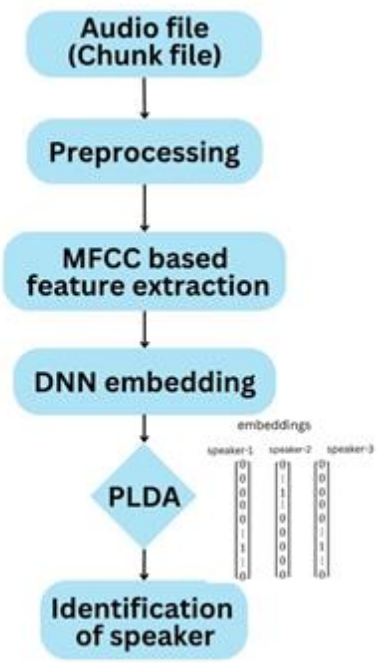

Fig.1: Framework of DNN-based Speaker Identification system

2.1 MFCC-based Feature Extraction

Feature extraction converts raw speech signals into acoustic vectors that capture speaker-specific characteristics for identification. Since speech is quasi-stationary, the signal is divided into short frames (20–100 ms) for spectral analysis, balancing dimensionality and discriminative power. Among various representations, Mel-Frequency Cepstral Coefficients (MFCCs) [10] are most widely used due to their similarity to human auditory perception and robustness to noise. The MFCC computation involves preprocessing (framing, windowing, and pre-emphasis), followed by FFT-based power spectrum analysis, Mel-scale filtering with overlapping triangular filters, log compression, and Discrete Cosine Transform (DCT) to produce compact, discriminative coefficients. Including delta and delta-delta features further enhances performance by capturing temporal dynamics, making MFCCs a robust and efficient representation for speaker identification systems.

2.2 DNN Embedding System

A speaker recognition system operates in two stages: enrollment and recognition. During enrollment, feature vectors extracted from speech are used to train speaker models

stored in a database, while recognition involves comparing a test sample with these models to identify the closest match based on similarity scores. Neural networks (NNs), inspired by biological neurons, offer high parallelism, generalization, and robustness, enabling efficient pattern recognition even in noisy conditions. Their nonlinear, multi-layered architecture allows them to learn complex relationships, making them widely applicable in tasks like speech and speaker recognition. They can tolerate partial hardware faults and learn implicit dependencies between inputs and outputs, though their internal workings remain difficult to interpret.

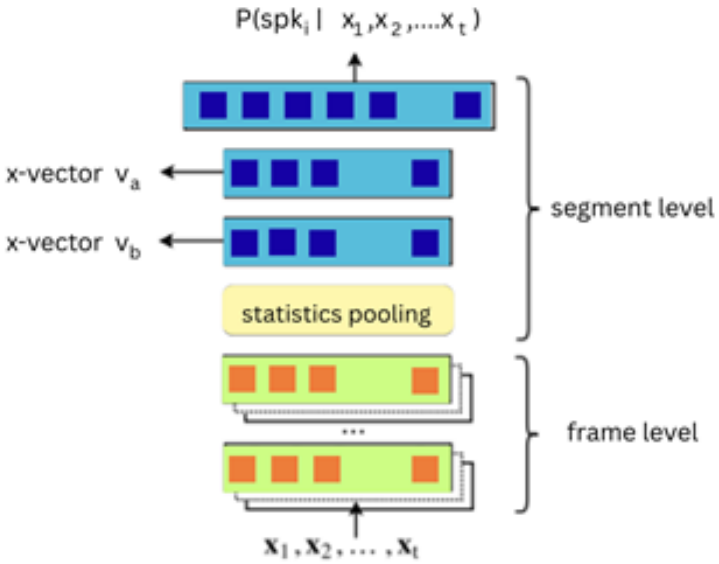

Fig. 2: DNN embedding architecture

In speaker recognition, deep neural networks [11] serve as powerful acoustic models, often replacing Gaussian Mixture Models (GMMs) or enhancing i-vector systems through bottleneck features. A DNN embedding system (as shown in figure 2) uses a feed-forward architecture trained with a cross-entropy objective, generating embeddings that capture speaker-specific traits across utterances. A statistics pooling layer aggregates temporal information to form compact representations, from which embeddings (a and b) are extracted for use with a Probabilistic Linear Discriminant Analysis (PLDA) backend for comparison. The input features are 20-dimensional MFCCs with voice activity detection and mean normalization, and the DNN employs a time-delay structure to capture temporal dependencies. The architecture includes multiple frame-level layers, a pooling layer computing mean and variance, and two segment-level layers before the softmax output, effectively modeling speaker characteristics over time.

2.3 PLDA Backend

The Probabilistic Linear Discriminant Analysis (PLDA) backend [12] is applied to both i-vector and x-vector embeddings to measure speaker similarity and make classification decisions. Before PLDA scoring, embeddings are centered, reduced in dimensionality using Linear Discriminant Analysis (LDA), and length-normalized. Reducing the LDA dimension to around 25% of the original size has been found effective in maintaining performance while lowering computational complexity. This pre-processing ensures that the embeddings are compact, discriminative, and suitable for PLDA comparison. In the PLDA backend, scores are normalized using adaptive s-norm to improve robustness and comparability across sessions. The DNN framework allows the use of different x-vector types, and instead of concatenating them, separate PLDA models are trained for each type, with their resulting scores averaged for final classification.

3. Results and Discussion

A Deep Neural Network (DNN)-based speaker identification framework was developed using a self-compiled speech corpus containing recordings from 51 Indian speakers. The dataset was designed to include diversity in terms of age, regional background, and linguistic variation. Speech samples were gathered through direct recordings as well as from publicly available YouTube videos, all standardized to .wav format [13-14]. Each participant provided multiple utterances in Hindi, English, Telugu, and bilingual combinations, ensuring broad linguistic coverage. Both male and female speakers were included, with intentional variations in accent, pronunciation, and recording conditions to enhance the system's ability to generalize across different environments.

Acoustic feature extraction was carried out using Mel-Frequency Cepstral Coefficients (MFCCs) computed with a 25ms frame size and a 10ms frame shift. For every sample, 13-MFCCs and their first- and second-order temporal derivatives (Δ and $\Delta\Delta$) were calculated, forming the input feature set for the DNN model. The DNN produced embeddings that were subsequently processed using PLDA to obtain matching scores for speaker identification. In total, the dataset comprised around 2,000 processed feature vectors per speaker, resulting in approximately 102,000 feature vectors for training and evaluation. Model performance was analyzed using metrics such as Equal Error Rate (EER), accuracy, and loss across varying epoch configurations.

3.1. Database

The speaker recognition experiments were conducted on an in-house corpus of 51 speakers, as detailed in Table 1. The database was curated to ensure demographic and linguistic diversity, comprising native speakers from various regions across India, including undergraduate students, doctors, actors, politicians, and professors. The cohort includes both male and female participants aged 17 to 60 years. Speech data was collected in three languages: Indian-accented English, Hindi, and Telugu. Each speaker contributed 20 utterances per language, with each utterance lasting 15-20 seconds. All

audio samples were standardized to a 16 kHz sampling rate in a mono channel format to ensure consistency for feature extraction. This composition makes the corpus particularly suitable for developing and evaluating robust, multi-lingual speaker verification systems.

Table 1: Description of Speech Corpus

S.No.	Particulars	Features
1	Total number of speakers	51
2	Areas of speakers	UG students, Doctors, Actors, Politicians, Professors.
3	Native places of speaker	All over India
4	Gender	Male and female
5	Frequency & Channel	16 KHz, Mono
6	Ages of speakers	17-60 years
7	Speaking Language	Indian Accent English, Hindi, Telugu
9	Duration of Each Speaker	15-20 Seconds
10	Number of utterances of each speaker	20 per language

3.2 Required Software/Packages

For building the Speaker Identification system, we utilized Google Colab, a cloud-based environment that provides access to GPU and TPU acceleration for training deep learning models efficiently. The experiments were implemented entirely in Python, using the TensorFlow and Keras frameworks for deep neural network modeling and the Librosa library for audio feature extraction. Colab provides a ready-to-use environment with pre-installed scientific computing libraries, enabling fast prototyping and experimentation without requiring manual setup. This approach is particularly useful for deep neural network (DNN) training, which can be computationally intensive on standard local machines.

The implementation of the speaker identification system was carried out using a structured software pipeline, as outlined in Table 2. The experimental framework was hosted on Google Colab, leveraging its cloud-based GPU resources to accelerate model training. The neural network architecture was constructed, trained, and evaluated using the TensorFlow and Keras deep learning frameworks. For audio pre-processing and feature extraction, the Librosa library was utilized to compute Mel-Frequency Cepstral Coefficients (MFCCs) from raw speech signals. Data manipulation and numerical computations were managed with NumPy and Pandas, while Matplotlib facilitated the visualization of training progress and acoustic features. Auxiliary machine learning utilities, including label encoding, dataset partitioning, and metric computation, were handled by scikit-learn.

Table 2: Required Tools and libraries to implement Speaker Identification

Required Packages	Description
Google Colab:	A cloud-based Jupyter notebook service that provides free access to GPU and TPU hardware for machine learning experiments.
TensorFlow / Keras	Deep learning frameworks used to design, train, and evaluate the speaker identification model.
Librosa	A Python library for audio and music analysis, used to extract MFCC (Mel-Frequency Cepstral Coefficients) and other acoustic features from the input voice samples.
NumPy and Pandas	Libraries used for numerical computation and efficient handling of large datasets.
Matplotlib	Used for visualizing accuracy, loss, and feature distribution during model training and testing.
scikit-learn	Used for label encoding, data splitting, and evaluation metrics computation.
Google Drive Integration	Used for dataset storage, model checkpoints, and saving final trained models.

3.3 Testing Results

In this subsection, the results of the testing simulations are presented and analyzed. The system's performance was evaluated using key metrics such as Equal Error Rate (EER), accuracy, and loss, measured across varying numbers of training epochs. Table 3 summarizes the performance of the proposed speaker identification framework in terms of Equal Error Rate (EER) and overall classification accuracy across varying epoch configurations. A consistent reduction in EER is observed with an increase in training epochs, decreasing from 9.5% at 10 epochs to 6.2% at 50 epochs. This decline signifies the model's enhanced discriminative capability and improved decision boundary optimization as training progresses.

Table 3: EER and Average Accuracy with variation of epochs

Epoch	EER (%)	Accuracy (%)
10	9.5	90.25
20	8.2	91.34
30	7.5	91.73
40	6.8	91.52
50	6.2	91.26

Simultaneously, the accuracy exhibits a gradual improvement, increasing from 90.25% to 91.26% across the same range, reflecting the model's ability to learn more robust speaker-dependent representations. The most pronounced improvement occurs between 10 and 30 epochs, after which the performance trend begins to plateau, indicating the commencement of model convergence and saturation in feature learning. Overall, the results demonstrate that extending the training duration up to 50 epochs yields meaningful gains in recognition accuracy and a substantial reduction in error rate.

However, beyond this point, further increases in epochs may lead to marginal improvements, suggesting the need for an optimal stopping criterion to balance model generalization and computational efficiency.

Table 4: Accuracy and Loss with the variation of epochs

Epoch	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
10	0.8946	0.9104	0.3115	0.2135
20	0.9090	0.9178	0.2868	0.2019
30	0.9149	0.9156	0.2761	0.1915
40	0.9090	0.9178	0.2868	0.2019
50	0.9129	0.9123	0.2604	0.2177

Table 4 illustrates the evolution of training and validation performance metrics—accuracy and loss—across varying epoch configurations for the proposed speaker identification model. The results indicate a consistent enhancement in learning efficiency and generalization capability up to the optimal training duration. The training accuracy improves steadily from 0.8946 at 10 epochs to 0.9129 at 50 epochs, while the validation accuracy shows a comparable trend, rising from 0.9104 to 0.9178 within the same range. Notably, the validation accuracy consistently remains higher than the training accuracy across all epochs, suggesting effective generalization without signs of over-fitting.

Similarly, both training and validation losses exhibit a decreasing trend, indicating the model's progressive optimization. The validation loss decreases from 0.2135 to 0.1915, remaining consistently lower than the training loss, which reduces from 0.3115 to 0.2604. This behavior signifies a stable and well-regularized learning process, wherein the model maintains superior performance on unseen data. Overall, the results demonstrate that the proposed model achieves strong convergence behavior with balanced learning dynamics, ensuring improved recognition accuracy and reduced generalization error across increasing epochs.

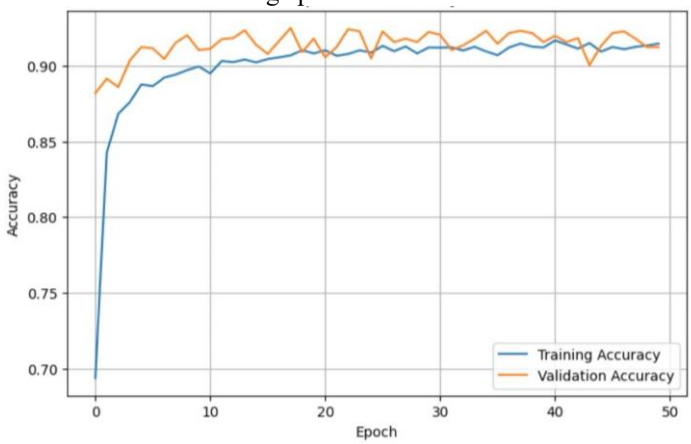

Fig.3: Accuracy of the speaker identification model over epochs.

The graph in Fig. 3 presents the variation of training and validation accuracy as a function of training epochs, illustrating the learning dynamics of the proposed speaker identification model. Both accuracy curves exhibit a steady upward trend during the initial training phase, followed by stabilization as the model approaches convergence. Notably, the validation accuracy consistently surpasses the training accuracy throughout the entire training duration, signifying effective generalization to unseen data. This relationship indicates that the model has learned discriminative speaker-specific features without succumbing to overfitting. The absence of divergence or significant fluctuation between the two curves further confirms the robustness and stability of the learning process.

By approximately the 50th epoch, both accuracy curves converge near the 0.91 mark, demonstrating that the model achieves reliable and balanced performance across training and validation datasets. The sustained alignment of the two curves suggests that the training process is well-regularized, with no evidence of performance degradation or instability. Overall, the observed behavior validates the efficacy of the adopted training strategy and highlights the model's ability to achieve high recognition accuracy through progressive learning and optimal generalization..

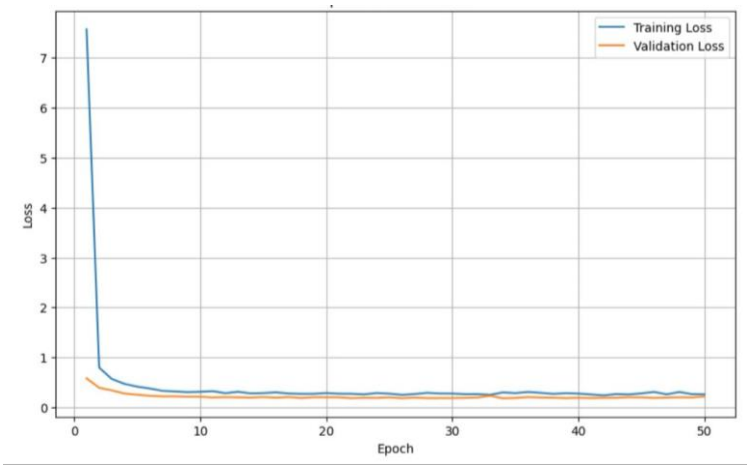

Fig.4: Model Loss of the speaker identification model over epochs.

Figure 4 depicts the variation of training and validation loss across successive epochs, illustrating the convergence behavior and optimization stability of the proposed speaker identification model. Both loss curves demonstrate a rapid and substantial decrease during the initial epochs, reflecting effective learning and swift adaptation of model parameters. As training progresses, the loss values gradually stabilize, indicating that the model has reached an optimal learning equilibrium. Notably, the validation loss consistently remains lower than the training loss throughout the training process, suggesting strong generalization capability and the absence of over-fitting. This behavior confirms that the model performs reliably on unseen data, maintaining minimal error while avoiding excessive bias toward the training set.

By the 50th epoch, both curves converge smoothly with negligible oscillations, signifying stable optimization and convergence of the deep learning architecture. The sustained low magnitude of both training and validation losses further validates the robustness of the learning framework and the effectiveness of the employed regularization and optimization strategies. Overall, the observed loss dynamics affirm that the proposed model achieves efficient convergence, reduced generalization error, and a well-balanced trade-off between bias and variance, which are critical for robust speaker identification performance

4. Conclusion

This study presented an effective deep neural network-based speaker identification framework capable of accurately recognizing individuals based on their voice characteristics using an in-house corpus of 51 speakers collected from diverse regional and linguistic backgrounds. The experimental results demonstrated consistent improvements in recognition accuracy and reduced Equal Error Rate (EER) with increasing training epochs, highlighting the model's capacity to learn robust and discriminative speaker embeddings. The system achieved optimal performance at 50 epochs, recording an EER of 6.2% and an accuracy of 91.26%, while both training and validation losses exhibited a stable and convergent trend. Furthermore, the validation accuracy consistently surpassed training accuracy, indicating strong generalization and minimal overfitting. These findings validate the efficiency of the proposed DNN-PLDA architecture for speaker identification and establish its reliability for practical deployment. Future work will focus on expanding the dataset to include more speakers and environmental variations, exploring noise-resilient and domain-adaptive models, and integrating transformer-based architectures to further enhance the scalability and robustness of speaker identification systems.

The system demonstrates potential for deployment in applications such as secure authentication, smart offices, and forensic speaker verification. However, challenges such as noise, channel variability, emotional speech, and spoofing attacks must be addressed in future work. Expanding the dataset and exploring transformer-based or ECAPA-TDNN architectures will further improve robustness.

References

[1]. Hanifa, Rafizah Mohd, Khalid Isa, and Shamsul Mohamad. "A review on speaker recognition: Technology and challenges." Computers & Electrical Engineering 90 (2021): 107005.

[2]. Faundez-Zanuy, Marcos, and Enric Monte-Moreno. "State-of-the-art in speaker recognition." IEEE Aerospace and Electronic Systems Magazine 20, no. 5 (2005): 7-12.

[3]. Kumar, Bittu. "Comparative Performance Evaluation of Greedy Algorithms for Speech Enhancement System." Fluctuation and Noise Letters 20, no. 2 (2021): 2150017-77.

[4]. Kumar, Bittu. "Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation." International Journal of Speech Technology 21, no. 4 (2018): 1033-1044.

[5]. Tahon, Marie, and Laurence Devillers. "Towards a small set of robust acoustic features for emotion recognition: challenges." IEEE/ACM transactions on audio, speech, and language processing 24, no. 1 (2015): 16-28.
[6]. Li, Ming, Kyu J. Han, and Shrikanth Narayanan. "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion." Computer Speech & Language 27, no. 1 (2013): 151-167.
[7]. GROZDIĆ, ĐORĐE, Slobodan Jovičić, ZORAN ŠARIĆ, and Irina Subotić. "Comparison of GMM/UBM and i-vector based speaker recognition systems." SPEECH AND LANGUAGE 2015 (2015): 274.
[8]. Sameer, V. V., Aljinu Khadar KV, and RK Sunil Kumar. "Speaker Verification Using Embedding Parameters in Noisy Environments (Transport Vehicles)." In 2024 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), pp. 1-7. IEEE, 2024.
[9]. Feng, Xue, Brigitte Richardson, Scott Amman, and James Glass. "On using heterogeneous data for vehicle-based speech recognition: A DNN-based approach." In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4385-4389. IEEE, 2015.
[10]. Prabakaran, Durairaj, and S. Sriuppili. "Speech processing: MFCC based feature extraction techniques-an investigation." In Journal of Physics: Conference Series, vol. 1717, no. 1, p. 012009. IOP Publishing, 2021.
[11]. Lin, Weiwei, Man-Wai Mak, Na Li, Dan Su, and Dong Yu. "A framework for adapting DNN speaker embedding across languages." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2810-2822.
[12]. Wang, Qiongqiong, Koji Okabe, Kong Aik Lee, and Takafumi Koshinaka. "Generalized domain adaptation framework for parametric back-end in speaker recognition." IEEE Transactions on Information Forensics and Security 18 (2023): 3936-3947.
[13]. Kumar, Bittu, Neeraj Kumar, Manoj Kumar, S. V. S. Prasad, Ashwini Kumar Varma, and Banoth Ravi. "Comparative studies of single-channel speech enhancement techniques." IETE journal of research 70, no. 6 (2024): 5704-5720.
[14]. Kumar, Bittu, and Ashwini Kumar Varma. "FPGA Implementation of Dynamic Quantile Tracking based Noise Estimation for Speech Enhancement." Journal of Engineering Science & Technology Review 16, no. 4 (2023).