Journal of Engineering and Technology

ISSN 2180-3811 eISSN 2289-814X https://jet.utem.edu.my/jet/index

DOI: https://doi.org/10.54554/jet.2025.16.2.013

ARTIFICIAL NEURAL NETWORK-BASED VOICEPRINT GENERATION MODELS FOR SPEAKER RECOGNITION

B. Sombo*[1], S. T. Apeh[1] and I. A. Edeoghon[1]

[1] Department of Computer Engineering, Faculty of Engineering, University of Benin, P.M.B. 1154, Ugbowo, Benin City, Edo State, Nigeria.

corresponding: sombopeter@gmail.com

Article history:
Received Date: 5 July 2025
Revised Date: 24 September 2025
Accepted Date: 3 October 2025

Keywords: Artificial Neural Networks, Cosine Similarity, Data Features

Abstract — Speaker recognition systems often do not prioritize generating high-quality voiceprints with minimal processing time, which can help reduce new user enrollment time while maintaining accuracy. Therefore, this study addressed the need for a model that can efficiently generate high-quality voiceprints, thus having the potential to improve system performance and enrollment speed when deployed in speaker recognition systems. Voice features, including Mel Frequency Cepstral Coefficients (MFCC), Gammatone Frequency Cepstral Coefficients (GFCC), Linear Predictive Coding (LPC) coefficients, and Perceptual Linear Prediction (PLP) coefficients, were extracted from clean

Extraction, Machine Learning, Speaker Recognition, Voice Biometrics, Voiceprint Generation

voice datasets collected from volunteers and the Mozilla Common Voice (MCV) database. Both Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks were then trained on these features for voiceprint generation. Evaluation using cosine similarity of voiceprints revealed that the MLP model trained with MFCC achieved the highest separation score (0.850553), outperforming the other models and this high value demonstrates its strong potential to enhance the accuracy and new user's enrollment time when deployed in speaker recognition systems.

I. Introduction

Speaker recognition is the process of identifying or verifying a person based on their voice characteristics, utilizing acoustic features that distinguish individuals. This technology is crucial in various applications, including security and access control for biometric authentication, forensic analysis in legal contexts, improving telecommunications services like voicemail, and monitoring patient health through voice analysis. The growing demand for voice-based interfaces and advancements in machine learning techniques have driven significant research interest in this field [1].

According to [2-7], speaker recognition includes speaker identification (SI) and verification (SV), each classified as text-dependent or text-independent. Speaker identification determines who is speaking by comparing the voice features of an unknown speaker against a database of enrolled templates in a 1:N matching process. In contrast, speaker verification confirms whether a speaker is who they claim to be by performing a 1:1 comparison between the input voice and a single stored template

associated with the claimed identity. Text-dependent methods require the speaker to utter specific phrases, such as passwords or PIN codes, for identification or verification, while text-independent methods rely on speech characteristics that are unrelated to the spoken content.

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and hybrid CNN-RNN models have been successfully trained as classifiers for speaker recognition tasks and have demonstrated strong performance [1, 8, 9-10]. However, these models typically require retraining or fine-tuning to enroll new users, resulting in long enrollment times and reduced scalability. This limitation poses a challenge for deploying speaker recognition systems in dynamic, real-world environments. To address this, the present study aims to develop a simple artificial neural network-based voiceprint generation model that enables faster enrollment while maintaining high-quality speaker representations.

To achieve this aim, the study set out to collect voice samples from volunteer participants and the Mozilla Common Voice (MCV) database, train Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) networks using extracted voice features such as MFCC, GFCC, LPC, and PLP for voiceprint generation and evaluate the quality of the generated voiceprints using cosine similarity scores to determine the model that produces the most effective speaker representations for recognition.

This paper contributes to the field of speaker recognition by presenting a simple artificial neural network-based voiceprint generation model capable of producing high-quality voiceprints. The proposed model has the potential to significantly reduce new user enrollment time, which hinders scalability and practical deployment.

II. Literature Review

A. Artificial Neural Networks

A neural network (NN) is a computational model that mimics the interconnected network of

neurons found in biological brains [11]. In contrast to their biological counterparts, artificial neurons in neural networks are mathematical constructs designed to process information and extract insights from data.

In a generalized artificial neural network (ANN) model, the net input to a neuron is computed as the weighted sum of its input signals. This summation forms the basis for further processing within the network. Once the net input is obtained, an activation function is applied to determine the neuron's output response. This output, which depends on the specific activation function used, represents the final value propagated forward in the network. The mathematical expressions [11] for these steps are as Equation (1).

$$\text{Net input } y_{in} = x_1.w_1 + x_2.w_2 + x_3.w_3 + \ldots + x_m.w_m = \sum_{i=1}^{m} x_i.w_i \quad (1)$$

where:

y_{in} = net input to a neuron (the total weighted sum of all input signals)

x_i = i_{th} input signal to the neuron

w_i = weight associated with the i_{th} input, representing the strength or importance of the input

m = total number of inputs to the neuron

$\sum_{i=1}^{m} x_i.w_i$ = summation of all inputs multiplied by their corresponding weights

$$Output = YF(y_{in}) \quad (2)$$

where:

Output = final response of the neuron after applying the activation function

y_{in} = net input to the neuron

$YF(y_{in})$ = activation function applied to the net input

$$Output = activation\ function \times net\ input \quad (3)$$

Neural networks excel in pattern recognition and addressing complex problems that often elude traditional programming approaches [12]. Their ability to learn and adapt makes them indispensable tools across various domains, including speaker recognition, image recognition, natural language processing, and robotics.

The MLP and LSTM networks can be deployed in processing voice data. MLPs demonstrate proficiency in discerning intricate patterns across diverse data types, while LSTM networks are finely tuned for excelling in pattern recognition within sequential data structures.

B. Multi-Layer Perceptron (MLP) Networks

The MLP is a cornerstone in neural network technology, distinguished by its structured architecture comprising input, hidden, and output layers connected by weighted connections. Recognized for its adaptability, MLPs excel in various tasks like classification, regression, and pattern recognition, leveraging their ability to discern intricate relationships between input data and desired output. Despite challenges in training deep architectures due to the vanishing gradient problem, MLPs find extensive application across diverse domains including speaker recognition. In this domain, MLPs contribute to tasks such as accurately identifying speakers by analyzing their speech patterns. The general architecture of MLP neural network is shown in Figure 1.

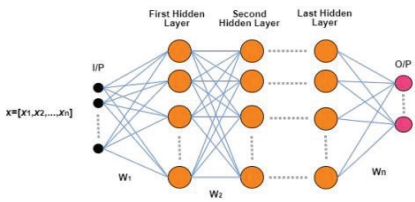

Figure 1: General Architecture of MLP Neural Network [13]

C. Long Short-Term Memory (LSTM) Networks

LSTM networks are specifically designed for handling sequential data where element order is crucial. LSTMs are adept at capturing long-term dependencies within sequences due to their unique cell structure, which incorporates gating mechanisms regulating information flow. These mechanisms enable LSTMs to effectively absorb and retain information over extended durations, making them indispensable for tasks like speech recognition and machine translation, where understanding word order and sentence context is vital. Additionally, LSTMs mitigate the vanishing gradient problem, facilitating successful

training in complex network architectures. Despite requiring more computational resources and sensitivity to hyperparameters, LSTMs find extensive application in various domains such as speech recognition, machine translation, time series forecasting, and video captioning, contributing to tasks like interaction with devices, precise translations, and accurate trend predictions. The fundamental LSTM neural network architecture is shown in Figure 2.

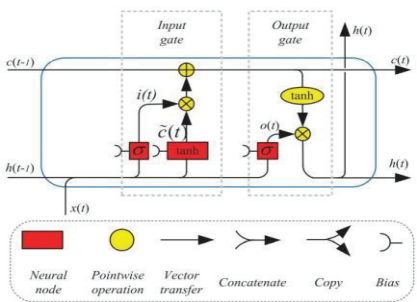

Figure 2: Fundamental LSTM Architecture [14]

D. Related Works

Despite significant advances in speaker recognition using deep learning techniques, most existing approaches prioritize accuracy and robustness without explicitly addressing a critical bottleneck in real-world applications which is enrollment time. For speaker recognition systems to be scalable and user-friendly, particularly in biometric authentication and security domains, the ability to quickly onboard new users with minimal voice samples are essential. The related reviews are critically examined with a focus on their applicability to enrollment time reduction.

[15] proposed a speaker recognition approach for intelligent home service robots. SincNet-based raw waveform processing was integrated with an ANFIS classifier enhanced by fuzzy c-means clustering. The model was evaluated on a custom noisy home-environment dataset with TV and robot motion sounds, and it outperformed traditional CNN, CNN-ANFIS, and standalone SincNet models in accuracy, demonstrating robust performance and transparency for practical robot applications. However, the study did not consider false acceptance or rejection rates or speaker enrollment time, which are vital for real-world deployment. enrollment.

[16] worked on improving speaker identification in

reverberant environments using MFCCs and comb filtering with neural network classification. A lightweight framework was developed by integrating comb filtering for reverberation suppression, MFCCs for feature extraction, and a neural network classifier for recognition. Experiments conducted under varying reverberation times (RT60 = 0.3-0.9s) and noise levels (SNR = 30-0dB) showed that the system achieved 97.6% accuracy in low-reverberation scenarios and 85.4% accuracy at high reverberation (RT60 = 0.9s), compared to 70.2% for the baseline. The study offers an effective, secure, and practical solution for real-time speaker recognition in challenging acoustic environments although the study is limited to speaker identification under reverberant conditions.

[17] proposed investigating the potential of multi-stage score fusion in spoofing-aware speaker verification. ECAPA-TDNN (ASV) and AASIST (CM) models were employed alongside support vector machine and logistic regression classifiers, with an additional auxiliary score from RawGAT (CM) incorporated to strengthen the system. Experimental evaluation on the SASV2022 dataset shows that the framework achieves an equal error rate (EER) of 1.30%, reflecting a 24% improvement over the baseline. The study shows that multi-stage ASV and CM integration greatly enhances security and robustness against spoofing. However, the work is limited to addressing spoofing attacks in speaker verification.

[18] worked on the performance analysis of machine learning approaches for developing a real-time speaker recognition system. CNN, KNN, and SVM classifiers were trained on MFCC and LPC features extracted from 160 audio files. The system was validated in real time with live microphone input and hardware feedback, achieving high accuracies of 91.67% (KNN), 97.92% (CNN), and 95.83% (SVM), thereby demonstrating the practical usability of machine learning for real-time speaker identification. However, the study does not address false acceptance/rejection rates or speaker enrollment time,

limiting its contribution to these important aspects of system performance and usability.

[19] presents an approach for enhancing speaker recognition models using noise-resilient feature optimization strategies. Their methodology involves extensive experiments on multiple speech datasets with varying speaker populations and noise conditions, employing classifiers such as KNN and LD. The study achieved speaker identification accuracies of up to 95.2% and equal error rates as low as 0.13%. These results demonstrate that feature optimization boosts both accuracy and computational speed, making the method suitable for large-scale speaker recognition applications. However, further work could explore ways to improve accuracy even more.

[1] proposed a speaker identification model combining 2D CNNs for spatial voiceprint feature extraction and stacked GRUs for temporal modeling. Evaluated on the Aishell-1 dataset, the model achieved 98.96% accuracy, outperforming CNN, RNN, and LSTM baselines. This integration significantly improved identification performance. However, the approach does not consider enrollment time, as deep GRUs and spectrogram-based input likely require long utterances and high computation. This limits its suitability for real-time or low-resource applications where rapid identification is essential.

[20] provides a comprehensive review of deep learning methods in speaker recognition, covering subtasks such as verification, identification, diarization, and robust recognition. The study explores core components including input features, network architectures, pooling strategies, and objective functions, while also highlighting recent advances in supervised and end-to-end systems, including online diarization. It emphasizes deep learning's ability to extract abstract speaker characteristics, achieving superior performance over traditional methods. While the review establishes deep learning as a foundation for future developments in the field, it does not address strategies for reducing speaker enrollment time, which

remains a key challenge for real-time or low-resource applications.

[21] proposed a hybrid speaker recognition method combining Artificial Neural Networks and Self Organizing Feature Maps (SOFM) for improved accuracy. MFCCs were used for feature extraction, followed by dimensionality reduction with SOFM and classification using an MLP with Bayesian Regularization. The system was trained and tested on the Multivariability speaker database with 10 speakers, achieving a 93.33% recognition rate. This demonstrates the method's effectiveness and potential for real-world speaker recognition applications. However, the study does not address reducing enrollment time for new users, limiting its immediate applicability in fast or dynamic environments. Further research is needed to improve real-time enrollment capabilities.

The reviewed literature contributes significantly to enhancing the accuracy and generalizability of speaker recognition systems. However the works do not directly address the challenge of minimizing new user enrollment time using neural network-based models. The absence of such considerations represents a crucial gap, particularly for applications requiring fast and scalable speaker identification.

[22] examined speaker recognition using x-vector neural embeddings across datasets including SITW, CMN2, and the mismatched VAST. Their approach evaluated TDNN-based architectures, pooling methods, training losses, and adaptation strategies. Results showed x-vectors significantly outperformed i-vectors, particularly under matched conditions. Despite enhancements from learnable dictionary encoders and back-end techniques like PLDA and AS-Norm, performance declined under domain mismatch. The study established x-vectors as a new standard but did not address fast enrollment. Model complexity and data demand present challenges for rapid or low-resource deployment, indicating future work is needed on enrollment efficiency.

[23] proposed a CNN-based speaker recognition model for 1D speech signals, introducing *convVectors* to extract speaker-specific features. Unlike standard 2D CNNs, their architecture learns filters tailored to vocal characteristics. Evaluated on the THUYG-20 SRE dataset under clean and noisy conditions, the model achieved a 43% improvement over the baseline and a low EER of 1.05%. This demonstrates the potential of adapting CNNs to speech-based tasks. However, the study does not address enrollment time reduction, and the computational demands of training may hinder real-time applications. Further research is needed to assess its feasibility in low-resource scenarios.

III. Methodology

The various activities undertaken to achieve the aim of the research are clearly formulated and structured in a workflow diagram presented in Figure 3.

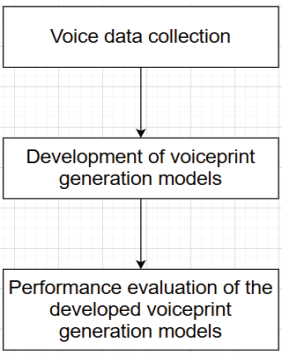

Figure 3: Workflow Diagram of the Research

A. Collection of Voice Data

A clean and diverse Voice data was collected from volunteer participants using Audacity software and high-quality microphones to minimize background noise. Each of the 20 volunteers contributed 50 recordings. To ensure sufficient representation for each speaker, data augmentation techniques such as time stretching and pitch shifting were applied as needed. A MATLAB script was developed to label all voice samples with their respective speaker identities, assigning labels like "1," "2," "3," and so on for the first, second, and third speakers, respectively. This organized dataset of labelled voice recordings served as the foundation for training neural network models.

B. Development of Voiceprint Generation Models

Features were extracted from voice samples because raw voice samples are too complex and noisy to use directly. Spectral features, such as MFCC and GFCC, which describe how energy is distributed across frequencies, were used because they mimic human hearing. Vocal tract characteristic features, including LPC and PLP, which describe the shape and behaviour of the speaker's vocal tract, were also employed, as everyone's vocal tract exhibits unique anatomical and physiological traits. These extracted features served as inputs for training Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM) neural networks. The choice of MLP and LSTM networks for speaker recognition tasks was based on their capabilities: MLPs are adept at learning intricate patterns and relationships in input data, while LSTMs excel at modeling temporal dependencies inherent in sequential speech signals. Each feature extraction technique was applied to derive feature vectors from the audio data of volunteer speakers, and these vectors were then used to train the MLP and LSTM architectures for generating voiceprints.

C. Performance Evaluation of the Developed Voiceprint Generation Models

The developed neural networks for voiceprint generation were evaluated using voice data from 20 speakers. Voiceprints were generated for each speaker, and cosine similarity scores [24] as shown in Equation (4) were computed between pairs of voiceprints which one from the same speaker and others from different speakers.

$$Cosinesimilarity = cos\emptyset = \frac{A.B}{||A||.||B||} \quad (4)$$

where:

$A \cdot B$ = dot product

$||A||$ = magnitude (or norm) of vector A, calculated as $||A|| = \sqrt{\sum_{i=1}^{n} A_i^2}$

$||B||$ = magnitude (or norm) of vector B, calculated as $||B|| = \sqrt{\sum_{i=1}^{n} B_i^2}$

This evaluation assessed the models' ability to generate distinct

voiceprints for everyone. By analyzing the difference in similarity scores between voiceprints from the same speaker and those from different individuals, the models' performance in distinguishing between speakers was measured, providing valuable insights into their effectiveness for speaker recognition tasks.

IV. Results and Discussion

The results of the performance evaluation of the developed models for voiceprint generation are presented in Table 1.

Table 1: Performance Evaluation of the Developed Models for Voiceprint Generation

Trained Networks	Mean Similarity Score from Same Speaker	Mean Similarity Score from Different Speakers	Difference Between Mean Similarity Scores
MLP network trained with MFCC	0.9455	0.094947	0.850553
MLP network trained with GFCC	0.8695	0.083526	0.785974
MLP network trained with LPC	1.00	1.00	0.00
MLP network trained with PLP	0.854	0.12639	0.72761
LSTM network trained with MFCC	0.9595	0.18637	0.77313
LSTM network trained with GFCC	0.918	0.1865	0.7315
LSTM network trained with LPC	0.9455	0.58389	0.36161
LSTM network trained with PLP	0.83	0.20924	0.62076

Table 1 shows the mean cosine similarity scores of the speakers' voiceprints obtained from the trained neural networks for 20 volunteer participants. The table summarizes the developed fast voiceprint generation models, including: the mean cosine similarity scores of voiceprints from the same speakers, the mean cosine similarity scores of voiceprints from different speakers, and the difference between these scores. This difference signifies the quality of the voiceprints generated by the models, the greater the difference between the mean similarity scores of voiceprints from the same and different speakers, the higher the quality of the voiceprint for speaker recognition.

The comparison of the quality of voiceprints generated by the developed models is presented in Figure 4. The bars represent the absolute differences in mean cosine similarity scores of voiceprints between same and different speakers (i.e., the quality of the voiceprint), generated by the MLP and LSTM ANN-based models trained with MFCC, GFCC, LPC, and PLP voice feature vectors. The bar representing the quality of voiceprints generated by the MLP ANN-based model trained with MFCC has the greatest height, indicating that this model generates the most distinctive voiceprints between same and different individuals.

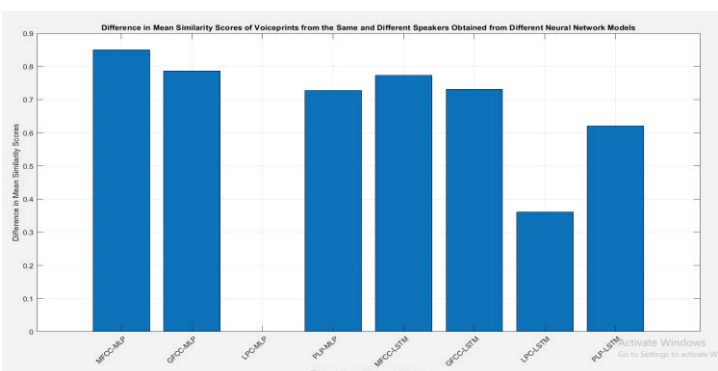

Figure 4: The Comparison of the Difference in Mean Similarity Scores of Voiceprints Between the Same and Different Speakers for All Trained Neural Networks

The high-quality voiceprints, reflecting the more distinct separation between same and different individuals, clearly show that a speaker recognition system deploying the developed fast voiceprint generation model will achieve high accuracy in recognizing speakers, as well as fast enrollment time for new users.

Score ranges depend strongly on factors such as dataset, embedding type (e.g., i-vector versus neural embeddings), preprocessing and score normalization. For instance, in the study by [25], a cosine similarity threshold of 0.13 was used to flag highly similar voiceprints from the same speaker. Previous works on i-vector and x-vector embeddings generally report overlap between same-speaker and different-speaker score distributions, with Equal Error Rates (EERs) ranging from 6.73% to 9.33% [22, 26-28]. However, the absolute difference between cosine similarity scores of same-speakers and different-speaker pairs is rarely reported directly. In this study, we obtained a difference of 0.85, which exceeds the separation typically observed in benchmark systems, demonstrating the generation of high-quality voiceprints suitable for speaker recognition tasks.

V. Conclusion

Voiceprint generation models were developed using Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) neural networks, trained with MFCC, GFCC, LPC, and PLP voice feature vectors. These models rapidly produced voiceprints, whose quality was evaluated by measuring the cosine similarity differences between same-speaker and different-speaker samples, with larger differences indicating higher discriminative power.

Among the models, the MLP trained with MFCC features achieved the best performance, reaching a difference of 0.850553 between the mean cosine similarity scores of same-speaker and different-speaker voiceprints, demonstrating strong potential for speaker recognition tasks.

The ability to quickly generate distinct and reliable voiceprints shows that the proposed models

can significantly reduce false acceptance and rejection rates and shorten speaker enrollment time, thereby contributing to more efficient and accurate voice-based authentication systems.

This research work is limited to the investigation of high-quality voiceprints ANN based models, with short processing time. Further study would consider deployment of the ANN-based voiceprint generation model in speaker recognition, considering the system's accuracy in recognizing speakers as well as new user enrollment time that will be achieved.

VI. References

[1] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Applied Sciences*, vol. 11, 2021.

[2] S. P. Todkar, S. S. Babar, R. U. Ambike, P. B. Suryakar and J. R. Prasad, "Speaker recognition techniques: A review," in *Proc. Int. Conf. for Convergence in Technology (I2CT)*, pp. 1-5, 2018.

[3] S. M. Kamruzzaman, A. N. Karim, M. S. Islam and M. E. Haque, "Speaker identification using MFCC-domain support vector machine," *Int. J. Engineering and Applied Sciences*, vol. 5, no. 3, pp. 274-278, 2020.

[4] A. Sajjad, A. Shirazi, N. Tabassum, M. Saquib and N. Sheikh, "Speaker identification and verification using MFCC and SVM," *Int. Res. J. Engineering and Technology (IRJET)*, vol. 4, no. 2, pp. 1950-1953, 2017.

[5] B. Squires and C. Sammut, "Automatic speaker recognition: An application of machine learning," in *Proc. Int. Conf. on Machine Learning*, pp. 515-521, 1995.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[7] Z. Yuan, B. Xu and C. Yu, "Binary quantization of feature vectors for robust text-independent speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 70-8, 1999.

[8] Y. H. Chen, L. M. Sainath, V. Mirkó, A. Raziel and P. Carolina, "Locally-connected and convolutional neural networks for small footprint speaker recognition," in *Proc. Interspeech*, Dresden, Germany, 2015.

[9] S. Parveen, A. Qadeer and P. Green, "Speaker recognition with recurrent neural networks," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[10] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *Proc. IEEE Spoken Language Technology Workshop*

(SLT), San Diego, USA, pp. 165-170, 2016.

[11] S. Saminu et al., "Development of a low-cost and accessible hand tremor rehabilitation game for unhealthy patients," *Journal of Engineering and Technology*, vol. 15, no. 2, pp. 137-151, 2024.

[12] R. A. A. Dunne, *Statistical Approach to Neural Networks for Pattern Recognition*. John Wiley & Sons, 2007.

[13] A. Botalb, M. Moinuddin, U. M. Al-Saggaf and S. S. A. Ali, "Contrasting convolutional neural network (CNN) with multi-layer perceptron (MLP) for big data analysis," in *Proc. Int. Conf. Intelligent and Advanced System (ICIAS)*, 2018.

[14] Y. Yu, X. Si, C. Hu and J. Zhang, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, pp. 1-36, 2019.

[15] S. H. Kim, T. W. Kim and K. C. Kwak, "Speaker recognition based on the combination of SincNet and neuro-fuzzy for intelligent home service robots," *Electronics*, vol. 14, no. 18, 2025.

[16] A. Hassan et al., "Improving speaker identification in reverberant environments using MFCCs and comb filtering with neural network classification," 2025.

[17] O. Kurnaz, T. H. Kinnunen and C. Hanilçi, "Investigating the potential of multi-stage score fusion in spoofing-aware speaker verification," in *Proc. Signal Processing and Communications Applications Conf. (SIU)*, Istanbul, Turkiye, pp. 1-4, 2025.

[18] A. Saxena and P. Garg, "Performance analysis of machine learning approaches for developed real-time speaker's voice recognition system," in *Proc. Int. Conf. Augmented Reality, Intelligent Systems, and Industrial Automation (ARIIA)*, Manipal, India, pp. 1-7, 2024.

[19] N. Chauhan, T. Isshiki and D. Li, "Enhancing speaker recognition models with noise-resilient feature optimization strategies," *Acoustics*, vol. 6, no. 2, pp. 439-469, 2024.

[20] Z. Bai and X. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65-99, 2021.

[21] K. J. Devi, N. H. Singh and K. Thongam, "Automatic speaker recognition from speech signals using self-organizing feature map and hybrid neural network," *Microprocessors and Microsystems*, vol. 79, 2020.

[22] . Villalba et al., "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, 2020.

[23] S. Hourri, N. S. Nikolov and J. Kharroubi, "Convolutional neural network vectors for speaker recognition," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 389-400, 2021.

[24] P. Miesle, "What is cosine similarity: A comprehensive guide," Datastax, Sep. 2023.

[25] Z. Tan, Y. Yang, E. Han and A. Stolcke, "Improving speaker identification for shared devices by adapting embeddings to speaker subsets," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1124-1131, 2021.

[26] M. I. Mandasari, M. L. McLaren and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," in *Proc. Interspeech*, 2011, pp. 21-24, 2011.

[27] Z. Bai, X. L. Zhang and J. Chen, "Cosine metric learning for speaker verification in the I-vector space," in *Proc. Interspeech*, pp. 1126-1130, 2018.

[28] W. Xia, J. Huang and J. H. Hansen, "Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation," in *Proc. IEEE ICASSP*, pp. 5816-5820, 2019.