JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-21, No.-08, August (2026) pp 198-215 ISSN (Print) 0973-8975

A CLOSE-SET MULTI-SPEAKER IDENTIFICATION via LIGHTWEIGHT CNN ARCHITECTURES LEVERAGING MFCC-DERIVED SPECTRAL FEATURES

N. K. KAPHUNGKUI

Electronics and Communication Engineering Department, Dibrugarh University, India-786004,

Email : ningshen@dibru.ac.in

https://doi.org/10.26782/jmcms.2026.08.00012

(Received: May 22, 2026; Revised: August 01, 2026; Accepted: August 12, 2026)

Abstract

Speaker identification is an important task in speech processing, with applications in security, authentication, forensics, and personalised human–computer interaction. This study presents a lightweight convolutional neural network (CNN)-based system for close-set identification of 40 speakers using Mel-Frequency Cepstral Coefficients (MFCCs). A dataset comprising 1,000 speech samples was collected from 40 speakers, with 850 samples used for training and 150 for validation. MFCC features were extracted from the speech signals and used as input to a compact CNN consisting of four convolutional blocks, three intermediate pooling operations, adaptive average pooling, and a fully connected classification layer. The proposed architecture contains 102,792 trainable parameters. Under the evaluated controlled conditions, the model achieved 100% validation accuracy, with all 150 validation samples correctly classified, and achieved an ROC-AUC of 1.00 for all 40 speaker classes. Pooling ablation experiments showed that removing any one of the three intermediate pooling operations reduced the best validation accuracy from 100.00% to 99.33%, supporting the use of the complete pooling configuration. An augmentation sensitivity analysis further showed that the unaugmented configuration achieved the highest validation accuracy of 100.00%, while pitch-shift augmentation resulted in 99.00% and Gaussian-noise augmentation reduced accuracy to 2.67%, both independently and when combined with pitch shifting. These findings indicate that the unaugmented configuration was the most effective setting for the proposed model under the evaluated conditions. Overall, the proposed lightweight CNN demonstrates effective closed-set speaker identification under controlled recording conditions, while further evaluation is required for open-set recognition, speaker verification, and greater recording variability.

Keywords: MFCC, CNN, Classification, Confusion matrix, Speaker Identification

I. Introduction

In recent years, speech has become one of the most natural and convenient means of human–computer interaction. Among various speech technologies, speaker identification systems have gained significant attention due to their wide range of applications in security, authentication, and personalised services. Unlike speech recognition, which focuses on understanding what is being said, speaker identification emphasises determining who is speaking by analysing the unique characteristics embedded in an individual's voice. These systems operate on the principle that every person's vocal tract, pitch, accent, and speaking style form a distinctive "voiceprint" that can be captured and modelled using advanced computational techniques.

A speaker identification system generally involves two major phases: training and testing (or recognition). In the training phase, a set of speech samples from a speaker is collected and processed to extract meaningful features such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), or spectral features. These features are then used to build a model representing the speaker's voice. During the testing phase, an unknown speech sample is compared against the stored models to identify the speaker. Depending on the application, the system may operate in two different modes: closed-set identification, where the speaker must be one among a predefined group, and open-set identification, where the system must also decide if the speaker is outside the enrolled database. The development of speaker identification systems has been motivated by the growing need for secure and user-friendly authentication methods. Traditional security mechanisms, such as passwords or identification cards, are prone to being lost, stolen, or forgotten. In contrast, a voice-based system provides a non-intrusive, easily accessible, and reliable means of verifying identity. This has made speaker identification particularly valuable in areas such as access control, forensics, banking transactions, and personalised customer support. Moreover, with the rise of smart devices and voice-activated assistants, the demand for accurate and robust speaker identification systems has expanded beyond security to include personalisation and adaptive user experiences.

Despite its potential, speaker identification faces several challenges. Variations in recording conditions, background noise, emotional state, health of the speaker, and channel distortions can significantly affect system performance. Researchers are addressing these issues by applying advanced machine learning and deep learning techniques, which improve robustness and adaptability. Methods such as convolutional neural networks, recurrent neural networks, and transformer-based models have recently shown great promise in handling variability and enhancing accuracy. This work will present a speaker identification system with an MFCC and CNN approach.

II. Literature Survey

Speaker recognition has developed from conventional statistical modelling approaches to modern deep-learning-based speaker representations. Early systems relied on statistical models to capture speaker-specific characteristics from speech. Reynolds et al. [XXVI] established the GMM-UBM framework for speaker verification, which became an important foundation for automatic speaker

recognition. However, such systems were affected by speaker, channel, and session variability. Kenny [XVI] addressed these limitations through Joint Factor Analysis (JFA), which explicitly modelled speaker and session variability. Probabilistic Linear Discriminant Analysis (PLDA) was subsequently introduced to model within-speaker and between-speaker variability [XIV], with later studies proposing improved PLDA formulations for speaker recognition [XX].

The introduction of i-vector representations provided a more compact way of representing variable-length speech utterances. Dehak et al. [V] proposed a factor-analysis-based approach for extracting fixed-dimensional speaker representations. Subsequent research investigated several improvements to i-vector systems, including length normalisation [IX] and multicondition PLDA training [X]. The UMD-JHU speaker recognition system further demonstrated the practical application of i-vector-based methods in large-scale evaluation settings [XI]. These developments established i-vector representations and PLDA as important approaches in speaker recognition before the widespread adoption of deep neural speaker embeddings.

Deep learning subsequently changed the way speaker representations were learned. Early studies investigated deep neural networks for text-dependent speaker verification [XXXI], while end-to-end approaches were developed to learn speaker-discriminative representations directly from speech [XIII][XXXIV]. The x-vector framework introduced fixed-dimensional speaker embeddings using neural networks and statistics pooling, providing a widely adopted approach for speaker recognition [XXIX]. Related work also investigated x-vector representations for spoken language recognition [XXVIII]. More advanced architectures, such as ECAPA-TDNN, further improved speaker representation learning through enhanced attention and feature aggregation mechanisms [VI]. Attentive statistics pooling [XXII] and large-margin fine-tuning [VII] have also been explored to improve the discriminative quality of learned speaker embeddings.

The availability of large speech datasets has further supported the development of deep speaker recognition systems. VoxCeleb provided a large-scale dataset containing speech collected from unconstrained conditions [XXI], while VoxCeleb2 expanded the number of speakers and speech samples available for training [IV]. These datasets enabled neural speaker recognition systems to be trained and evaluated using more diverse speech data. In parallel, toolkits such as BOSARIS [I] and Kaldi [XXIV] have supported system development, scoring, calibration, and reproducible evaluation.

Researchers have also investigated the effects of acoustic variability and limited training data on speaker recognition performance. Data augmentation has been used to introduce additional variability during training [XII], while domain adaptation techniques have been explored to reduce performance degradation between training and evaluation conditions [XXXIII]. Calibration methods have been investigated to improve the reliability of speaker recognition scores [VIII]. More recent studies have also considered few-shot and data-efficient speaker recognition approaches when only limited speech data are available for new speakers [XIX]. Bayesian approaches have additionally been explored to incorporate uncertainty into speaker recognition models [XXXII].

Security and forensic applications have introduced additional requirements for speaker recognition systems. The ASVspoof challenges have provided benchmarks for evaluating spoofing attacks and countermeasures in automatic speaker verification [XXXV] [XXXVI]. Tan et al. [II] reviewed presentation attack detection methods and discussed important challenges affecting the security of automatic speaker verification systems. In forensic applications, Sigona et al. [XXVII] investigated the validation of an ECAPA-TDNN-based system, demonstrating the application of modern speaker embeddings in forensic automatic speaker recognition.

Despite the progress of deep speaker embedding methods, conventional acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) remain useful because of their relatively simple extraction and computational requirements. MFCC-based voice recognition has been investigated in traditional feature-based systems [III], while Khan et al. [XVII] demonstrated the use of MFCC features with a probabilistic neural network for text-independent speaker recognition. These studies indicate that MFCCs can still provide useful speaker-discriminative information when combined with suitable machine-learning models.

Overall, the literature shows a progression from statistical speaker models to i-vector representations and, more recently, deep speaker embeddings such as x-vectors and ECAPA-TDNN [XXVI][V][VI]. Despite the strong performance of modern deep speaker-embedding systems, MFCC-based approaches remain useful when a compact and computationally lightweight solution is required. Motivated by this observation, the present study investigates a lightweight CNN using MFCC feature maps for close-set identification of 40 speakers under controlled recording conditions.

III. Methodology

The proposed system follows a structured pipeline beginning with dataset preparation, preprocessing, feature extraction, and neural network training. A dataset containing audio samples from 40 unique speakers was employed. The voice samples were collected from different age groups in a relatively noise-free environment with the same recording devices. Dataset preprocessing involved amplitude normalisation and segmentation of audio waveforms. MFCC features were extracted from the preprocessed audio waveforms and used as input to the CNN for speaker classification. A lightweight convolutional neural network (CNN) architecture was employed for speaker classification. The dataset was divided into training and validation sets, with training conducted over 40 epochs.

Speaker Pattern Classification Using CNN with MFCC Input

Convolutional Neural Networks (CNNs) have shown great success in processing speech signals for speaker classification. Instead of relying purely on handcrafted features, CNNs can learn hierarchical representations from MFCC feature maps, making them suitable for speaker classification. This report outlines the process of extracting features from raw audio and training a CNN to classify speaker patterns.

Feature Extraction

Although CNNs can directly process raw waveform samples, it is often beneficial to use MFCCs as input features. MFCCs mimic the human auditory system by

emphasising perceptually important frequency bands. Mel-Frequency Cepstral Coefficients (MFCC) are among the most widely adopted techniques for speech feature extraction, valued for their straightforward implementation and strong performance [XXXVI]. Their effectiveness largely stems from their design, which emulates the characteristics of the human auditory system. Although MFCCs have some limitations in capturing signals above 1 kHz, they continue to demonstrate remarkable robustness, particularly in noisy environments. The extraction involves the following steps, as shown in Figure 1.

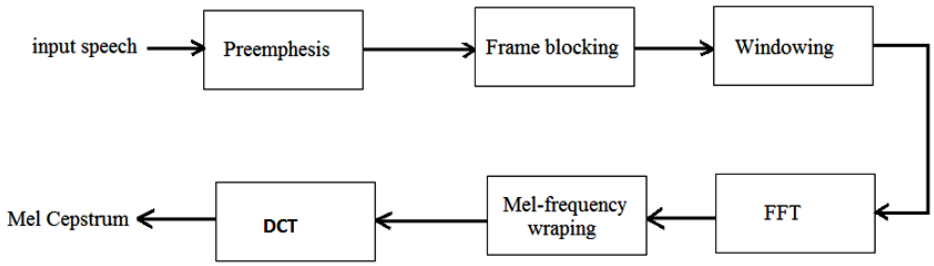

Fig. 1. MFCC extraction steps

1. Pre-emphasise the speech signal: When the speech signal S[n] is passed to a high-pass low-order filter, the output So[n] is expressed as

$$So[n] = S[n] - aS[n-1] \quad (1)$$

where a is a constant and lies between 1> a > 0.9.

The Z transform of the filter is given by $H(Z) = 1 - aZ^{-1}$

2. Framing the speech signal into shorter frames (ms): Here, the continuous speech signal is blocked into several frames of N samples. The adjacent frames are separated by M (N> M). The initial frame consists of the first N samples. The second frame starts after the first frame, which overlaps it by N - M samples, and so on. This process continues until the entire speech length is covered

$$\text{Frame size} = (\text{Sampling frequency}) \times (\text{Frame duration}) \quad (2)$$

$$\text{Total no of frames} = \text{speechlength/framesize} \quad (3)$$

3. Windowing: After framing the entire length of the speech signal, the next step is to window every individual frame. This step is done in order to minimise the signal discontinuities at the starting and ending points of each frame. The aim of windowing is the minimisation of the spectral distortion to taper the signal gradually to zero at the beginning and ending points of each frame signal. The mathematical expression

of Hamming windows is expressed (here, a window is L frames long)

$$\text{Hamming } w[n] = 0.54 - 0.4\{\cos(2.\pi.\, n)/L\} \qquad 0 \leq n \leq L - 1 \quad (4)$$

$$\text{Hamming } w[n] = 0 \text{ Otherwise} \quad (5)$$

4. Fourier Transform: The input of the Discrete Fourier Transform is a windowed signal x[n]...x[m], and the output for each of N discrete frequency bands is a complex number X [k] representing the magnitude and phase of that frequency component in

the original signal. A commonly used algorithm for computing the DFT is the Fast Fourier Transform or FFT. It is defined as

$$X[k] = \sum_{n=0}^{N-1}\left(x[n]e^{-(j2\pi nk)/N}\right) \qquad k= 0, 1 , 2\ldots\ldots N\text{-}1 \qquad (6)$$

Generally X[k] is complex numbers and their absolute values are only considered

5. Mel Filter Bank: The FFT result will give the information about the amount of energy at each frequency band. The human ear hearing system, however, is not equally sensitive at all frequency bands. At higher frequency ranges, roughly above 1 kHz, our ears are less sensitive. Hence, modelling this property of the human ear hearing system during the feature extraction enhances and improves the performance of the speech recognition system. The particular form of model which is used in MFCCs is the warping of frequencies output by the Fast Fourier Transform into a mel scale. The mapping between frequency in Hertz and the mel scale is linear below 1 KHz and after that it follows the logarithmic scale above 1 kHz. The mel frequency can be computed from the raw acoustic frequency as follows.

$$Mel(f) = 1127\ ln\ (1+f/700) \qquad (7)$$

6. Discrete Cosine Transform (DCT): Decorrelate filter bank energies:

$$c_k=\sum_{m=1}^{M} \log E_m .\ \cos\{(\pi k)/m .\ (m-1/2)\} \qquad K=1,2,\ldots.k \qquad (8)$$

k is the index of the cepstral coefficient (usually 12–13 are kept)

M is typically 20–40 filter banks

K<M to reduce dimensionality

Em = log energy output of the mmm-th Mel filter

ck = the k^{th} MFCC coefficient

Figure 2 represents the Speaker CNN model, which takes MFCC feature maps as input and outputs the speaker identity. The proposed CNN contains four convolutional blocks with three intermediate max-pooling operations, followed by adaptive average pooling and a fully connected classification layer. The network contains **102,792 trainable parameters**. The theoretical receptive field progressively increases through the convolutional and pooling operations, reaching approximately **38 × 38 feature-map positions** at the output of the fourth convolutional layer. This progressive expansion enables the network to capture broader spectral-temporal patterns while maintaining a compact representation.

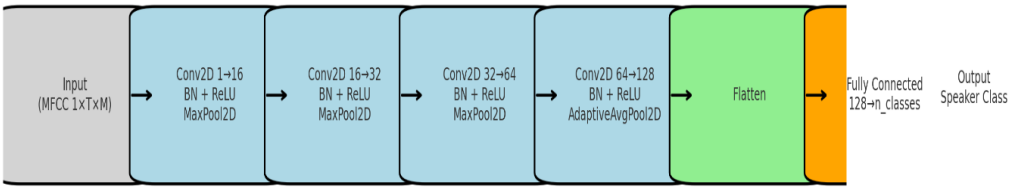

Fig. 2. Architecture of CNN

1. Input Layer (MFCC Features): Input is a 2D array of MFCCs, shaped as 1 × T × M where 1 represents the channel (like a grayscale image), T represents the number of time frames, and M represents the number of Mel frequency bins. This acts like an "image" where the CNN can detect patterns across time and frequency.

2. Convolutional + BatchNorm + ReLU + Pooling Layers: The CNN has four convolutional blocks, each learning progressively more abstract features:

Conv2D (1→16): Learns low-level patterns (edges, energy bursts).

Conv2D (16→32): Detects short-term temporal patterns.

Conv2D (32→64): Captures more complex combinations of spectral features.

Conv2D (64→128): Extracts high-level representations specific to a speaker's voice. Each block contains: BatchNorm: Normalises activations to stabilise learning. ReLU: Non-linear activation introducing sparsity. MaxPool2D: Reduces feature map size, keeping only important information. The last block uses AdaptiveAvgPool2D to shrink feature maps to 1×1, summarising global features.

3. Flatten Layer: Converts the pooled feature maps (shape 128 × 1 × 1) into a 128-dimensional vector. This vector represents the learned 128-D speaker representation. The 128-dimensional feature vector represents the learned speaker representation used for the enrolled-speaker classification task.

4. Fully Connected Layer (128 → n_classes): Maps the 128-dimensional feature vector to n_classes, where each class represents a different speaker

5. Output Layer (Speaker Class): The final softmax layer outputs probabilities across all speaker classes. The class with the highest probability is predicted as the speaker.

Figure 3 illustrates the pipeline from raw WAV input to speaker classification using MFCC feature extraction and a CNN model.

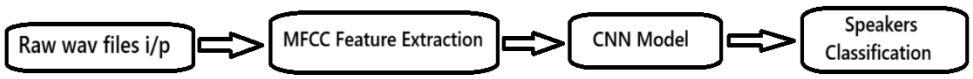

Fig. 3. Workflow Diagram

Step 1: Load raw audio files

Step 2: Extract MFCC features

Step 3: Prepare the dataset

Step 4: Define CNN model, Compiling, Training and Evaluating model performance

Step 5: Test the model with new unseen voice samples and identify the speaker

Data augmentation was evaluated as a sensitivity experiment. Gaussian noise was applied with a probability of 0.5 using a scale factor of 0.005, while random pitch shifting was applied with a probability of 0.5 over a range of −2 to +2 semitones. These augmentation configurations were applied only to the training data, while validation samples remained unmodified. The resulting configurations were

compared with the baseline model without augmentation to determine their effect on validation performance.

IV. Results and Discussion

The system was evaluated using multiple performance metrics, including accuracy, confusion matrix, loss convergence, and ROC curves. The results are presented as follows:

Figure 4 illustrates the training confusion matrix, where correct predictions lie along the diagonal, showing excellent classification consistency. The confusion matrix obtained from the CNN-based speaker identification system for 40 speakers demonstrates exceptional performance, with every prediction lying perfectly on the diagonal and all off-diagonal entries being zero. This indicates that the model was able to correctly classify every training sample without a single misclassification, showcasing its ability to effectively capture and differentiate the unique vocal characteristics of each speaker. The balanced distribution of samples across the 40 speaker classes, each with approximately 21 or 22 utterances, further confirms the robustness of the model in learning speaker-specific features such as timbre, pitch, and spectral patterns. The absence of overlap or confusion between classes highlights the discriminative capability of the CNN architecture, while the consistent validation performance indicates strong classification performance under the evaluated closed-set conditions. Overall, the confusion matrix provides compelling evidence that the proposed CNN-based approach achieves near-perfect speaker separation, making it effective for the evaluated closed-set speaker identification task.

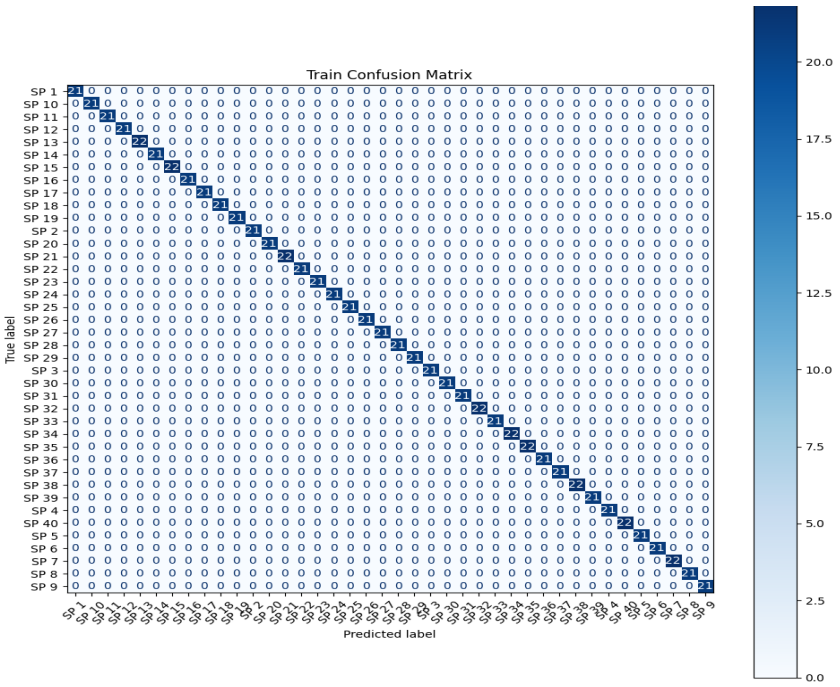

Fig. 4. Training confusion matrix

Figure 5 presents the validation confusion matrix for the 40 speaker classes. All non-zero entries occur along the main diagonal, with each speaker having either three or four correctly classified validation samples. The variation between three and four samples per class results from the distribution of the validation samples across the 40 speakers. Importantly, no off-diagonal entries are present, indicating that none of the validation samples were misclassified. Thus, all 150 validation samples were correctly identified, corresponding to a validation accuracy of 100%.

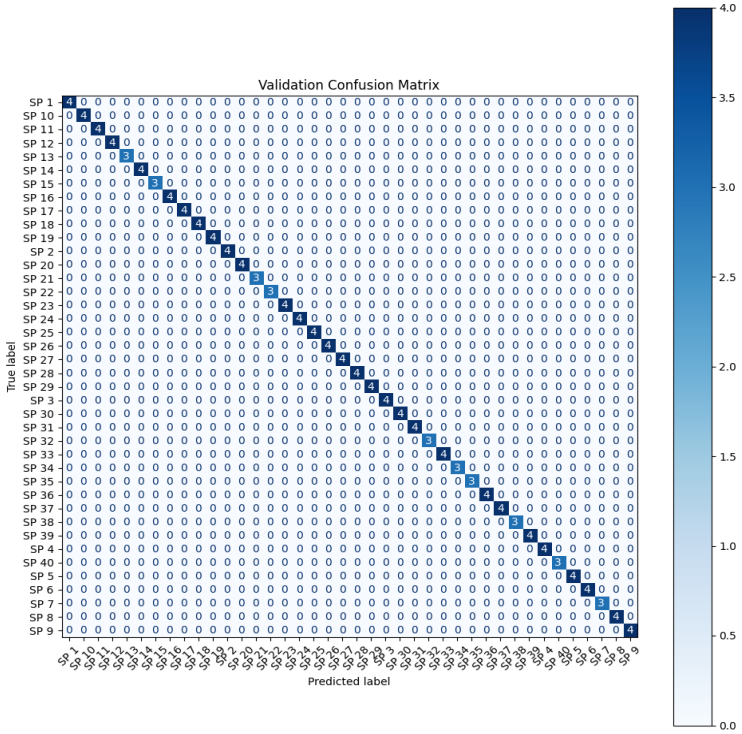

Fig. 5. Validation confusion matrix

Figure 6 depicts the training and validation loss curves, along with validation accuracy. The training and validation performance curves further reinforce the reliability of the model, providing deeper insights into its learning behaviour over successive epochs. As observed in the loss graph, both training and validation losses exhibit a rapid decline during the initial epochs, indicating that the model quickly captured meaningful patterns from the dataset. The close alignment between the training and validation loss curves throughout the epochs reflects that the model effectively generalised without significant overfitting. The validation accuracy curve demonstrates a sharp increase during the early epochs, reaching above 90% accuracy within the first ten epochs, and stabilising around near-perfect accuracy values (close to 100%) as training progresses. The occasional minor fluctuations in validation accuracy are expected and can be attributed to inherent variations in the validation set, yet the overall stability indicates consistent validation performance under the evaluated conditions. These findings directly correlate with the confusion matrix,

where nearly all speakers were classified correctly with minimal misclassifications, confirming that the model not only converged efficiently but also maintained exceptional predictive accuracy across diverse classes. Together, these results demonstrate the robustness and consistency of the model for the evaluated speaker classification task.

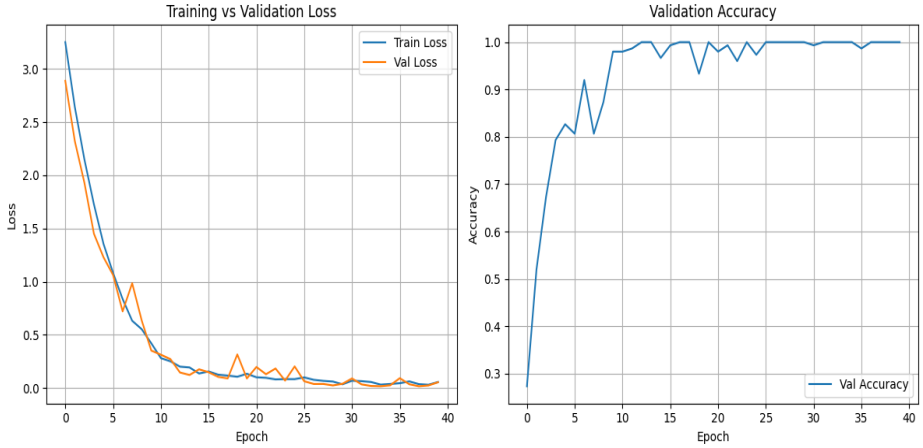

Fig. 6. Training and validation loss curves and validation accuracy.

Figure 7 presents the ROC curves for all 40 speaker classes. Each class achieves an AUC score of 1.00, confirming perfect discrimination across speakers. The Receiver Operating Characteristic (ROC) curve provides an additional perspective on the classification performance of the model, particularly in distinguishing between different speakers in a multiclass one-vs-rest setting. As illustrated in the figure, each speaker's class (SP 1 to SP 40) achieved an Area Under the Curve (AUC) score of 1.00, representing perfect classification capability. This outcome indicates that the model was able to flawlessly discriminate between positive and negative instances for every individual class without overlap in prediction errors. The ROC curve lines, which closely align with the top-left boundary of the plot, emphasise the model's ability to maximise true positive rates while maintaining zero false positive rates across all classes. The diagonal reference line represents random classification, and the significant separation of the actual ROC curves from this line further highlights the superior performance of the model. These results are consistent with the confusion matrix and validation accuracy findings, which collectively demonstrate near-perfect precision and recall across all categories. The ROC analysis thus provides strong statistical confirmation of the model's robustness, reinforcing its applicability for highly accurate speaker recognition tasks in practical scenarios.

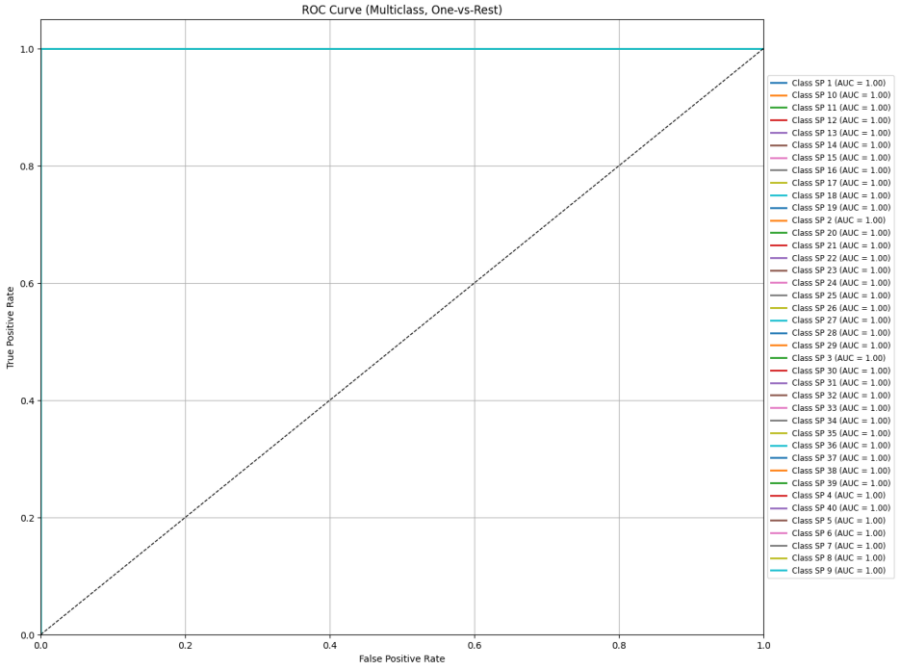

Fig. 7. ROC curves for all 40 speaker classes.

Figure 8 provides waveform visualisations for multiple speaker testing samples and their predicted speakers. The final stage of evaluation involved testing the trained model using previously unseen audio samples from all registered speakers, and the results demonstrate its exceptional robustness and reliability. The waveform plots presented for multiple speakers show the raw audio signals, each annotated with its true identity and the model's predicted label. Importantly, the predictions align perfectly with the ground truth across all cases, confirming that the model successfully identified every speaker without misclassification. This outcome demonstrates accurate speaker identification on previously unseen samples from the enrolled speakers under the evaluated conditions. The consistent accuracy across diverse speaker waveforms correlates strongly with earlier findings from the confusion matrix, training-validation performance curves, and ROC analysis, all of which indicated near-perfect classification capability. Together, these results demonstrate strong closed-set speaker identification performance under the controlled evaluation conditions.

V. Speaker Representation and Unseen-Sample Identification Analysis

The 128-dimensional speaker representation was examined as an intermediate representation learned by the closed-set classification model. After training and validation, the model was evaluated on new, unseen samples from the enrolled speakers, and the results show that it correctly predicted the test speaker, as shown in Figure 8.

Testing sample	Predicted Speaker	Testing sample	Predicted Speaker
File: SP10_26.wav -> Predicted Speaker: SP 10		File: SP29_26.wav -> Predicted Speaker: SP 29	
File: SP11_26.wav -> Predicted Speaker: SP 11		File: SP2_26.wav -> Predicted Speaker: SP 2	
File: SP12_26.wav -> Predicted Speaker: SP 12		File: SP30_26.wav -> Predicted Speaker: SP 30	
File: SP13_26.wav -> Predicted Speaker: SP 13		File: SP31_26.wav -> Predicted Speaker: SP 31	
File: SP14_26.wav -> Predicted Speaker: SP 14		File: SP32_26.wav -> Predicted Speaker: SP 32	
File: SP15_26.wav -> Predicted Speaker: SP 15		File: SP33_26.wav -> Predicted Speaker: SP 33	
File: SP16_26.wav -> Predicted Speaker: SP 16		File: SP34_26.wav -> Predicted Speaker: SP 34	
File: SP17_26.wav -> Predicted Speaker: SP 17		File: SP35_26.wav -> Predicted Speaker: SP 35	
File: SP18_26.wav -> Predicted Speaker: SP 18		File: SP36_26.wav -> Predicted Speaker: SP 36	
File: SP19_26.wav -> Predicted Speaker: SP 19		File: SP37_26.wav -> Predicted Speaker: SP 37	
File: SP1_26.wav -> Predicted Speaker: SP 1		File: SP38_26.wav -> Predicted Speaker: SP 38	
File: SP20_26.wav -> Predicted Speaker: SP 20		File: SP39_26.wav -> Predicted Speaker: SP 39	
File: SP21_26.wav -> Predicted Speaker: SP 21		File: SP3_26.wav -> Predicted Speaker: SP 3	
File: SP22_26.wav -> Predicted Speaker: SP 22		File: SP40_26.wav -> Predicted Speaker: SP 40	
File: SP23_26.wav -> Predicted Speaker: SP 23		File: SP4_26.wav -> Predicted Speaker: SP 4	
File: SP24_26.wav -> Predicted Speaker: SP 24		File: SP5_26.wav -> Predicted Speaker: SP 5	
File: SP25_26.wav -> Predicted Speaker: SP 25		File: SP6_26.wav -> Predicted Speaker: SP 6	
File: SP26_26.wav -> Predicted Speaker: SP 26		File: SP7_26.wav -> Predicted Speaker: SP 7	
File: SP27_26.wav -> Predicted Speaker: SP 27		File: SP8_26.wav -> Predicted Speaker: SP 8	
File: SP28_26.wav -> Predicted Speaker: SP 28		File: SP9_26.wav -> Predicted Speaker: SP 9	

Fig. 8. Multiple speaker testing samples and their predicted speakers

The results highlight the exceptional performance of the proposed system, achieving an excellent classification accuracy across 40 speakers. The confusion matrix confirms the robustness of the model, with zero misclassifications observed in the training and validation phases. The rapid convergence of training and validation loss curves reflects the efficiency of the architecture, while the ROC analysis indicates perfect one-vs-rest class separability for the evaluated speaker classes. These results suggest that the system is highly effective for controlled speaker identification tasks. Experimental reproducibility was ensured through deterministic initialisation and controlled randomisation settings.

Pooling Ablation Analysis

A separate pooling sensitivity experiment was conducted to examine the contribution of the three intermediate pooling operations in the proposed CNN. The findings are shown in Table 1.

Table 1

Model	Pooling Configuration	Accuracy (%)
Baseline	Pool 1 + Pool 2 + Pool 3	**100.00**
No Pool 1	Pool 2 + Pool 3	**99.33**
No Pool 2	Pool 1 + Pool 3	**99.33**
No Pool 3	Pool 1 + Pool 2	**99.33**

The baseline model, which retains all three pooling operations, achieves the highest validation accuracy of 100.00%. Removing any one of the three pooling operations results in a marginal reduction in accuracy to 99.33%. Although the decrease is small (0.67 percentage points), the consistent reduction across all three ablated configurations indicates that retaining all pooling stages provides the best-performing configuration for the proposed architecture.

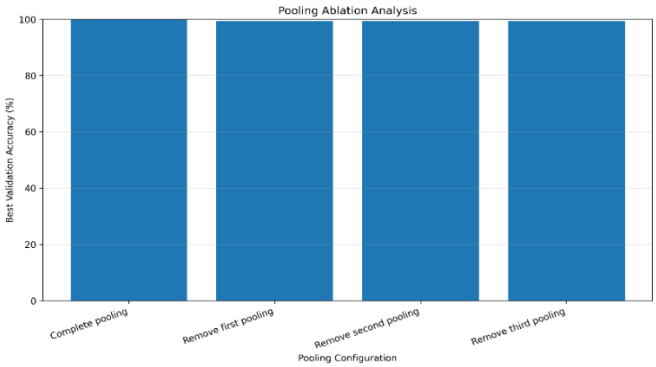

Fig. 9. Comparison of best validation accuracy across complete and single-pooling-removal configurations.

Figure 9 further illustrates the pooling ablation results, showing that the complete pooling configuration achieves the highest validation accuracy of 100.00%. Removing any individual pooling layer produces a small reduction to 99.33%, indicating that retaining all three pooling stages provides the best overall configuration for the proposed model.

Augmentation Sensitivity Analysis

Table 2 presents the augmentation sensitivity analysis of the proposed model. The model achieves the highest validation accuracy of 100.00% without augmentation, while pitch-shift augmentation results in a marginal decrease to 99.00%. In contrast, the use of Gaussian noise, either alone or combined with pitch shifting, substantially reduces the validation accuracy to 2.67%. These results indicate that the unaugmented configuration provides the most effective setting, while Gaussian-noise augmentation is unsuitable for the proposed model under the evaluated conditions.

Table 2

Configuration	Augmentation	Best Validation Accuracy (%)
No augmentation	None	100.00
Pitch shift only	Pitch	99.00
Gaussian noise only	Noise	2.67
Pitch shift + Gaussian noise	Both	2.67

Figure 10 illustrates the augmentation sensitivity analysis, showing that the model achieves the highest validation accuracy of 100.00% without augmentation. Pitch-shift augmentation causes only a marginal decrease to 99.00%, whereas Gaussian noise, either alone or combined with pitch shifting, reduces accuracy substantially to 2.67%. These results indicate that Gaussian-noise augmentation is unsuitable under the evaluated conditions, while the unaugmented configuration provides the best overall performance.

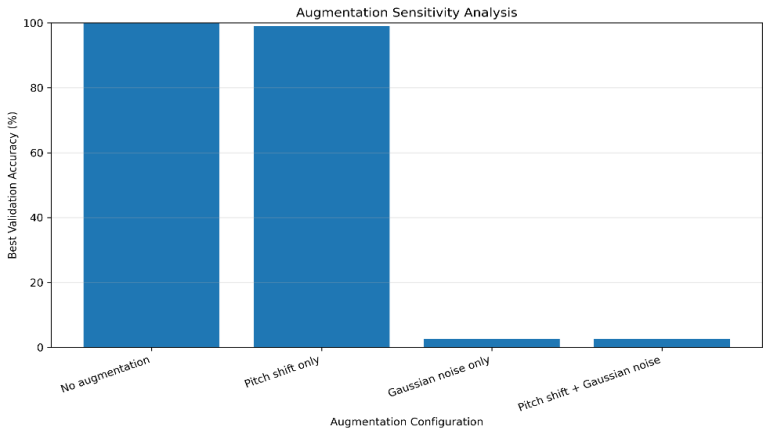

Fig. 10. Augmentation sensitivity analysis showing the best validation accuracy under different augmentation configurations.

VI. Conclusions

This study presented a lightweight CNN-based speaker identification system using MFCC features for close-set classification of 40 speakers. The proposed model contains 102,792 trainable parameters and achieved 100% validation accuracy, with an ROC-AUC of 1.00 for all 40 speaker classes under the evaluated controlled conditions. The pooling ablation analysis showed that retaining all three intermediate pooling operations achieved the best validation accuracy of 100.00%, while removing any one pooling operation reduced accuracy to 99.33%. The augmentation sensitivity analysis showed that the unaugmented configuration provided the best performance at

100.00%, compared with 99.00% for pitch shifting and 2.67% for Gaussian noise, either alone or combined with pitch shifting.

Overall, the proposed lightweight MFCC-CNN model demonstrates effective closed-set speaker identification under controlled conditions. However, the results do not establish open-set or general-purpose biometric verification capability. Future work will focus on speaker verification, open-set evaluation, and robustness to greater speaker and recording variability.

. **Conflict of Interest:**

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

I. Brümmer, Niko, and Johan de Villiers. "The BOSARIS Toolkit and Calibration Methods for Speaker Recognition Scoring." *Proceedings of the 2011 NIST Speaker and Language Recognition Workshop*, 2011.

II. C. B. Tan, M. H. A. Hijazi, N. Khamis, P. N. E. binti Nohuddin, Z. Zainol, F. Coenen, and A. Gani, "A survey on presentation attack detection for automatic speaker verification systems: State-of-the-art, taxonomy, issues and future direction," *Multimedia Tools and Applications*, vol. 80, nos. 21–23, pp. 32725–32762, 2021. doi: 10.1007/s11042-021-11235-x

III. Chakraborty, Koustav, Asmita Talele, and Savitha Upadhya. "Voice Recognition Using MFCC Algorithm." *International Journal of Innovative Research in Advanced Engineering*, vol. 1, no. 10, 2014, pp. 158–161.

IV. Chung, Joon Son, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition." *Interspeech 2018*, 2018, pp. 1086–1090. 10.21437/Interspeech.2018-1929.

V. Dehak, Najim, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. "Front-End Factor Analysis for Speaker Verification." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, 2011, pp. 788–798. 10.1109/TASL.2010.2064307.

VI. Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification." *Interspeech 2020*, 2020, pp. 3830–3834. 10.21437/Interspeech.2020-2650.

VII. Desplanques, Brecht, Jenthe Thienpondt, and Kris Demuynck. "Large Margin Fine-Tuning for Speaker Recognition." *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 10.1109/ICASSP40776.2020.9054465.

VIII. Ferrer, Luciana, Matthew McLaren, and Niko Brümmer. "Robust Back-End Calibration and Adaptation Techniques for Speaker Recognition." *Computer Speech & Language*.

IX. Garcia-Romero, Daniel, and Carol Y. Espy-Wilson. "Analysis of i-Vector Length Normalization in Speaker Recognition Systems." *Interspeech 2011*, 2011, pp. 249–252. 10.21437/Interspeech.2011-53.

X. Garcia-Romero, Daniel, and Aaron McCree. "Multicondition Training of Gaussian PLDA Models in i-Vector Space for Noise and Reverberation Robust Speaker Recognition." *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4257–4260. 10.1109/ICASSP.2012.6288859.

XI. Garcia-Romero, Daniel, et al. "The UMD-JHU 2011 Speaker Recognition System." *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4225–4228. 10.1109/ICASSP.2012.6288852.

XII. He, Y., et al. "Exploration of Data Augmentation and Augmentation Strategies for Speaker Recognition." Conference paper.

XIII. Heigold, Georg, Ignacio Moreno, Samy Bengio, and Noam Shazeer. "End-to-End Text-Dependent Speaker Verification." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5115–5119. 10.1109/ICASSP.2016.7472652.

XIV. Ioffe, Sergey. "Probabilistic Linear Discriminant Analysis." *Computer Vision—ECCV 2006 Workshops*, Springer, 2006, pp. 531–542. 10.1007/11744085_41.

XV. Ji, R., et al. "An End-to-End Text-Independent Speaker Identification Framework." *Interspeech 2018*, 2018, pp. 3267–3271.

XVI. Kenny, Patrick. "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms." Technical Report CRIM-06/08-13, 2005.

XVII. Khan, Suhail Ahmad, Anil S. Thosar, Jagannath H. Nirmal, and Vinay S. Pande. "A Unique Approach in Text Independent Speaker Recognition Using MFCC Feature Sets and Probabilistic Neural Network." *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, IEEE, 2015, pp. 1–6.

XVIII. Kinnunen, Tomi, and Haizhou Li. "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors." *Speech Communication*, vol. 52, no. 1, 2010, pp. 12–40. 10.1016/j.specom.2009.08.009.

XIX. Lee, Kong Aik, et al. "Data-Efficient Speaker Recognition and Few-Shot Adaptation Techniques." Conference paper.

XX. Lei, Yun, et al. "A Novel Scheme for Speaker Recognition Using a PLDA Variant." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 1695–1699.

XXI. Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. "VoxCeleb: A Large-Scale Speaker Identification Dataset." *Interspeech 2017*, 2017, pp. 2616–2620. 10.21437/Interspeech.2017-950.

XXII. Okabe, Koji, Takafumi Koshinaka, and Koichi Shinoda. "Attentive Statistics Pooling for Deep Speaker Embedding." *Interspeech 2018*, 2018, pp. 2252–2256. 10.21437/INTERSPEECH.2018-993.

XXIII. Okada, H., et al. "Improvements in i-Vector PLDA Scoring and Calibration for NIST SREs." IEEE workshop/conference paper.

XXIV. Povey, Daniel, et al. "The Kaldi Speech Recognition Toolkit." *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

XXV. Ravanelli, Mirco, and Yoshua Bengio. "Speaker Recognition from Raw Waveform with SincNet." *Interspeech 2018*, 2018, pp. 3698–3702. 10.21437/Interspeech.2018-2492.

XXVI. Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker Verification Using Adapted Gaussian Mixture Models." *Digital Signal Processing*, vol. 10, nos. 1–3, 2000, pp. 19–41. 10.1006/dspr.1999.0362.

XXVII. Sigona, Francesco, et al. "Validation of an ECAPA-TDNN System for Forensic Automatic Speaker Recognition." Speech Communication, 2024. 10.1016/j.specom.2024.103045.

XXVIII. Snyder, David, et al. "Spoken Language Recognition Using X-Vectors." *Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 105–111. 10.21437/Odyssey.2018-15.

XXIX. Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. "X-Vectors: Robust DNN Embeddings for Speaker Recognition." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333. 10.1109/ICASSP.2018.8461375.

XXX. Sztahó, Dániel, Gábor Szaszák, and Anna Beke. "Deep Learning Methods in Speaker Recognition: A Review." *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 65, no. 4, 2021, pp. 310–328. 10.3311/PPee.17024.

XXXI. Variani, Ehsan, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez. "Deep Neural Networks for Small-Footprint Text-Dependent Speaker Verification." *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056. 10.1109/ICASSP.2014.6854363.

XXXII. Villalba, Jesús, and Niko Brümmer. "Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance." *Interspeech 2011*, 2011, pp. 505–508. 10.21437/Interspeech.2011-142.

XXXIII. Villalba, Jesús, et al. "Domain Adaptation Techniques for PLDA and Neural Back-Ends." ICASSP/Interspeech proceedings.

XXXIV. Wan, Li, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. "Generalized End-to-End Loss for Speaker Verification." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4879–4883. 10.1109/ICASSP.2018.8462665.

XXXV. Wu, Zhizheng, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. "ASVspoof 2015: The First Automatic Speaker Verification Spoofing and Countermeasures Challenge." *Interspeech 2015*, 2015, pp. 2037–2041. 10.21437/Interspeech.2015-462.

XXXVI. Wu, Zhizheng, et al. "ASVspoof 2017/2019 Challenge Series: Challenge Overview and Datasets." *Interspeech Proceedings*.