iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** XII **Month of publication:** December 2025

DOI: https://doi.org/10.22214/ijraset.2025.76683

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

Forensic Speaker Recognition: CNN-Based Inter and Intra Speaker Performance Analysis

Kavita Waghmare[1], Mohammad Basil Abdul Kareem[2], Bharti Gawali[3]
[1]*Research Student,* [3]*Senior Professor, Dr Babasaheb Ambedkar Marathwada, Univeristy Chh. Sambhaji Nagar*
[2]*Assistant Professor, University of Anabar*

Abstract: ***This research presents a comprehensive deep learning based framework for automatic forensic speaker recognition, designed to address the challenges commonly encountered in speaker identification tasks. The proposed system employs Convolutional Neural Networks (CNNs) trained on mel spectrogram representations of speech signals to capture distinctive vocal attributes unique to each individual. A dataset composed of speech recordings from 20 speakers, equally divided between male and female participants, was preprocessed into 3-second mel spectrogram segments to ensure consistent analysis and robust feature extraction.***

The CNN model was optimized to perform speaker classification and to generate discriminative speaker embedding's that effectively represent vocal identity. Performance evaluation demonstrated high classification accuracy and strong generalization across varying acoustic conditions. To assess the forensic reliability of the learned representations, the embedding distributions were analyzed using t-SNE visualization techniques. The resulting plots revealed well-defined speaker clusters with low intra-speaker variation, confirming the model's capability to differentiate between distinct voices.

Overall, the outcome highlight the effectiveness and interpretability of the proposed CNN-based approach for forensic speaker recognition. This framework holds significant potential for real world forensic applications, including voice evidence authentication, suspect identification, and speaker profiling. Furthermore, it provides a foundation for future research in developing more resilient and transparent models for forensic speech processing.

Keywords: Forensic Speaker Recognition, Deep Learning, CNN, Mel Spectrogram, Speaker Embedding, Inter-Speaker Variability, Intra-Speaker Variability, t-SNE.

I. INTRODUCTION

Speaker recognition has emerged as a vital component in forensic investigations, particularly in cases involving voice-based evidence such as phone calls, surveillance recordings, or intercepted communications. Identifying or verifying individuals based on their speech characteristics can significantly aid law enforcement in criminal profiling, suspect identification, and authentication tasks. The ability to accurately attribute a voice sample to a specific individual has far reaching implications in both judicial and intelligence situations [1], [2] [3].

However, the task of speaker recognition in forensic contexts is fraught with challenges. A major concern is intra-speaker variability differences in the same speaker's voice due to emotional state, health condition, or speaking style. Similarly, **i**nter-speaker variability, such as overlapping vocal features among different speakers, poses classification difficulties. Additional complications arise from real-world audio data being noisy, distorted, or truncated, particularly in covert recordings or phone interceptions, where utterances may be short and of poor quality[4],[5],[6].

To address these challenges, this study leverages Convolutional Neural Networks (CNNs) a class of deep learning models known for their effectiveness in learning spatial hierarchies from 2D inputs. Combined with mel spectrograms**,** which offer a time-frequency representation of audio aligned with human auditory perception, CNNs can learn discriminative features essential for robust speaker recognition. This makes them particularly well-suited for forensic applications where both accuracy and interpretability are crucial [7], [8] [9].

The design and implementation of a CNN-based speaker recognition framework that learns meaningful embedding's from mel spectrograms and classifies speakers is presented in this research. In addition to classification accuracy, we evaluate the model's performance by carefully examining inter and intra speaker variability in the learnt feature space. In order to improve interpretability, Speaker embedding's are displayed using t-distributed stochastic neighbor embedding (t-SNE), which clearly shows speaker separability and clustering behavior.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue XII Dec 2025- Available at www.ijraset.com

II. LITERATURE REVIEW

Speaker recognition has long been a focus of research in the fields of speech processing and forensic science. Over the years, methodologies have evolved from classical statistical models to advanced deep learning frameworks. This section describes significant advancements in the field and places the present work in the context of this changing environment.

A. Classical speaker recognition methods

Traditional speaker recognition systems relied heavily on statistical models such as Gaussian Mixture Models with Universal Background Models (GMM-UBM) and i-vector frameworks. The GMM-UBM approach models the distribution of spectral features and estimates speaker-specific parameters through adaptation techniques. Later, i-vectors became the dominant technique, providing a compact, low-dimensional representation of a speaker's characteristics derived from acoustic feature statistics. While effective in controlled settings, these methods are limited in capturing non-linear and high-dimensional variations, especially under forensic conditions such as noisy environments or short utterances [10], [11], [12].

B. Deep Learning-Based Methods

Recent advancements in deep learning have led to the development of more powerful speaker recognition models. One such model is the x-vector framework, which extracts speaker embedding's using Time Delay Neural Networks (TDNNs) and is highly effective for speaker verification tasks. Additionally, Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have been employed to capture spatial and temporal patterns in speech. CNNs, in particular, are well-suited for processing 2D time-frequency representations like spectrograms and have shown promising results in various speaker identification tasks[13],[14][15].

C. Use of Mel Spectrograms and Embedding Spaces

Mel spectrograms have gained popularity as input features due to their alignment with the human auditory system. They provide a perceptually relevant representation of speech, preserving both temporal and frequency information. When used in conjunction with CNNs, mel spectrograms enable the extraction of discriminative speaker embedding's, which can be projected into high-dimensional spaces for classification or comparison. These embedding's offer interpretable and reusable representations of speaker identity, making them suitable for forensic analysis [16], [17].

D. Speaker Variability in Forensic Studies

Forensic speaker recognition introduces unique challenges such as intra-speaker variability (variation within the same speaker) and inter-speaker similarity (vocal resemblance across different speakers). Prior studies have highlighted the impact of speaking style, emotional state, channel distortion, and environmental noise on recognition accuracy. While traditional systems struggle to adapt to these variations, deep learning-based embedding's have demonstrated better resilience. However, there is still limited work on visualizing and quantifying speaker variability in the embedding space, particularly from a forensic perspective [18], [19], [20].

E. Research Gap and Motivation

Although CNN-based models and mel spectrograms have shown potential in general speaker recognition, their application in forensic contexts remains under explored, especially with regard to inter and intra-speaker variability analysis. Most existing studies focus on improving accuracy, with less emphasis on the interpretability of embedding's or their forensic relevance[21][22].This study addresses this gap by developing a CNN-based framework that not only performs speaker classification but also provides visual and quantitative insights into speaker variability using t-SNE and embedding distance metrics. This approach enhances the forensic applicability of deep learning models by combining high performance with interpretability.

III. METHODOLOGY

A. Database Specification

The database specifications for the speech corpus used in this forensic speaker recognition study are detailed as follows:

- Sampling Frequency: 16 kHz
- Recording Environment: Studio setting
- Microphone Distance: 10 cm from the speaker

- Temperature: 25°C
- Channels: Single channel (mono)
- Participants: 10 male and 10 female speakers, aged between 20 and 50 years
- Speech Samples: A total of 1,000 speech samples collected from these 20 speakers, with each speaker recording five sentences in both read and continuous modes, repeated ten times (five in each mode)
- Sentence Content: Five Marathi sentences composed of idioms and proverbs, designed to capture speech variability
- Data Collection Setup: Recordings made using a Computerized Speech Lab (CSL) system during consistent afternoon sessions to minimize environmental variability

This structured approach ensures uniformity and control over the recording conditions, which is crucial for modelling speaker characteristics reliably. The below table shows the sentences used for experiment.

Table 1 Marathi sentences used for database

Sr.no	Sentences
1.	रिकाम्या पेपर ला जाहिरातीचा आधार
2.	लवकर उठे लवकर निजे त्यास आरोग्य संपत्ती लाभे
3.	ऊस गोड लागला म्हणून मुळासकट खाऊ नये
4.	एक तीळ सात जणांनी वाटून खाल्ला
5.	हिरे मोती हे मौल्यवान आहे

B. Pre-processing

Mel spectrogram extraction using the librosa library is used to for pre-processing. Each audio file is loaded at a fixed sample rate (16,000 Hz), converted to a mel spectrogram with 40 mel bands, and then normalized to decibel units. The resulting spectrograms are either padded or truncated to a fixed number of frames (100), ensuring uniform input size for the CNN model. This technique emphasizes the perceptually relevant frequency features which are effective for speaker recognition tasks.

C. Feature Extraction

The features considered for deep learning in this forensic speaker recognition approach are primarily mel spectrograms. These are time-frequency representations of speech signals that align closely with human auditory perception, capturing both spectral and temporal information essential for distinguishing different speakers [23]. The mel spectrograms serve as 2D input images that are processed by the CNN to learn discriminative features or embedding's that encode speaker-specific vocal characteristics. Additionally, once trained, the model generates speaker embedding's in a high-dimensional space, which can be visualized and analyzed for intra- and inter-speaker variability, further supporting forensic identification[24].

This figure 1 presents a detailed comparison of mel spectrograms for four different speakers (Sandip Thorat, Vijay Dangar, Simran Maniyar, and Karuna Kirwale) as they read the same sentence, visualized using their speech recordings. Every subplot is a mel spectrogram that shows the spoken sentence's frequency content with time for a single speaker. Time (in seconds) is displayed on the x-axis, while frequency (in Hz) is displayed on the y-axis. The shorter spectrograms of Simran Maniyar and Vijay Dangar indicate that either they read the line more quickly. One of the main indicators in forensic speaker identification is the density and height of formant bands, which have characteristics unique to each speaker.

In figure 2 each subplot visualizes acoustic features using frequency (Hz, y-axis), time (seconds, x-axis), and signal amplitude/intensity shown as color. In mel spectrograms, brighter colors (yellow-orange) represent higher energy, while darker colors (purple-black) indicate lower energy. The black region in Sandip Thorat's subplot on the right side indicates that this utterance is shorter than the full horizontal plotting window, so the remaining time axis is effectively silence or zero-padding. A similar effect is visible in Simran Maniyar's panel, where the sentence appears to end earlier, followed by a darker region with substantially reduced energy. In contrast, Vijay Dangar's and Karuna Kirwale's spectrograms fill more of the time axis with active speech, suggesting slightly longer or more continuous realizations of the same sentence, possibly with fewer or shorter silent pauses.

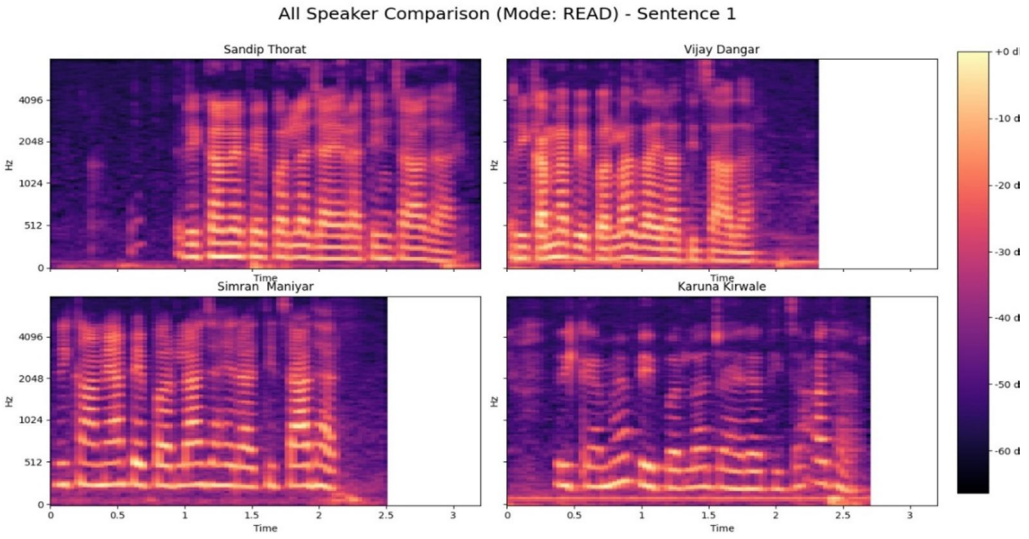


Figure 1. Mel Spectrogram of Read Mode

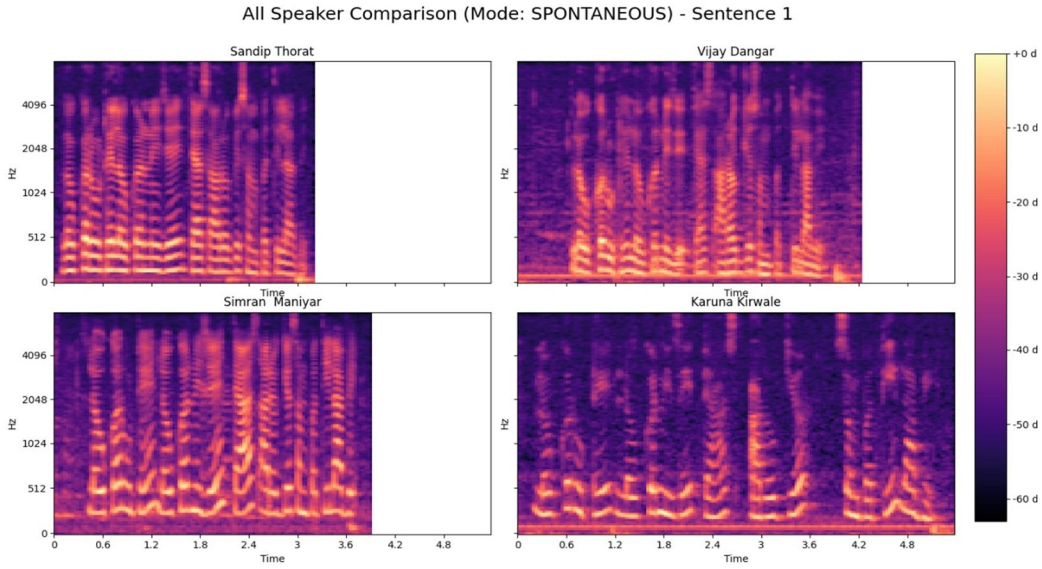


Figure 2. Mel Spectrogram of Spontaneous Mode

D. Model Architecture

A Convolutional Neural Network (CNN) was designed to learn discriminative features from the mel spectrograms. The architecture consists of a feature extractor and a classifier head. This speaker identification network uses a deep convolutional architecture to process (1, 40, 100) mel spectrogram inputs, passing them through three Conv2d blocks with increasing filter sizes (32, 64, 128) and max pooling, followed by flattening and a fully connected embedding layer to generate 128-dimensional speaker vectors[25],[26]. A final linear classifier outputs logits for 40 distinct speakers. The model was trained using Adam optimizer (learning rate 0.001), with Cross Entropy loss, batch size 8, and for 15 epochs on an 80/20 train-test split, following contemporary standards for effective speaker identification tasks using deep learning[27] [28].

IV. EXPERIMENTS AND RESULTS

A. Speaker Classification Accuracy

After 15 epochs of training, the convolutional neural network (CNN) model was evaluated on an independent test dataset to assess its speaker recognition capabilities. The model exhibited strong performance, demonstrating its ability to accurately identify speakers based on their voice samples.

Specifically, the evaluation results showed an accuracy of 97.50%, a precision of 97.61%, a recall of 97.50%, and an F1-score of 97.48%. These consistently high metrics indicate that the CNN effectively learned robust and discriminative features from the mel spectrogram inputs, supporting its suitability for forensic analysis and reliable speaker identification tasks.

B. Inter vs Intra Speaker Analysis

A comprehensive analysis of both within-speaker (intra-speaker) and between-speaker (inter-speaker) variability in the learned embedding space to evaluate the forensic robustness of the proposed CNN-based speaker recognition system. This analysis is vital in forensic applications to ensure the system can reliably distinguish different speakers even when dealing with short or degraded speech samples. This figure 3 presents a 2-D t-SNE projection of utterance-level embedding's for four speakers. Each point corresponds to one utterance; proximity reflects similarity in the high-dimensional embedding space, while the two t-SNE components have no direct physical interpretation. Marker shapes index speakers (circles: Sandip Thorat; squares: Vijay Dangar; triangles: Simran Maniyar; crosses: Karuna Kirwale), and color encodes speaking mode (blue: read; red: spontaneous). Points form compact, well-separated speaker clusters, indicating that the embedding's are strongly speaker-discriminative, with read and spontaneous tokens forming nearby sub-clusters that reflect style-induced variability.

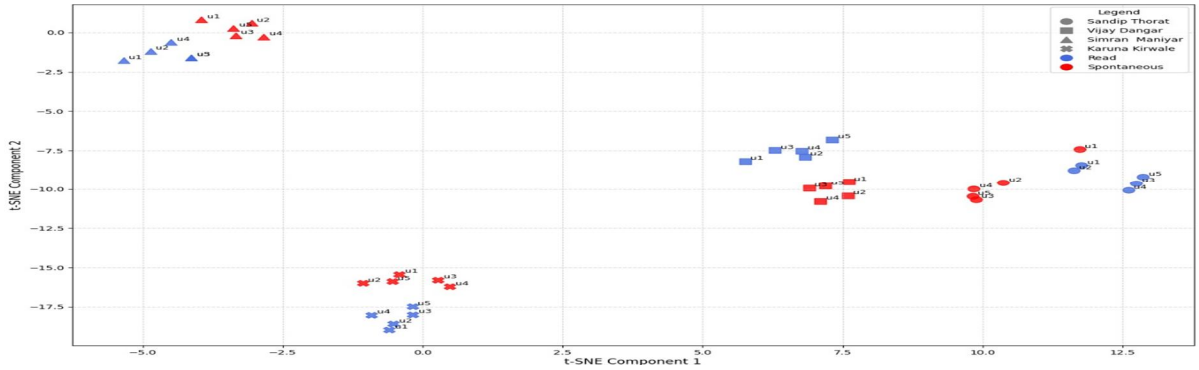


Figure 3 t-SNE visualization for inter and intra speaker variation

1) Intra Speaker Variability

The t-SNE figure 3 demonstrates low intra-speaker variability, as data points corresponding to the same speaker form tight and cohesive clusters regardless of whether the samples are recorded in read or spontaneous mode. Within each cluster, the individual utterances for a speaker are closely grouped together, indicating that the CNN model produces consistent and robust embedding's for that speaker across different speaking styles and sessions. This minimal spread within speaker clusters reflects the model's ability to effectively minimize intra-speaker variability, which is crucial for reliable forensic speaker identification

2) Inter Speaker Variability

The t-SNE figure 3 provides clear evidence of high inter-speaker variability, as reflected by the well-separated clusters corresponding to different speakers. Each cluster contains samples uniquely belonging to one individual, and there is minimal or no overlap between the embedding's of different speakers. This distinct separation indicates that the CNN model is highly effective at extracting and learning features that differentiate one speaker's vocal patterns from another. High inter-speaker variability is essential for forensic speaker recognition, as it ensures reliable discrimination between speakers and robust identification performance.

C. Embedding Visualization with t-SNE

To gain deeper insights into the model's learned speaker representations, embedding visualization was performed using t-Stochastic Neighbor embedding (t-SNE). This method helps to project high-dimensional speaker embedding's into a 2D space, enabling visual inspection of speaker clustering and separability.

The embedding's were extracted from the penultimate layer of the trained CNN model by setting the return_embedding=True flag during inference. Each audio sample was passed through the network, and the resulting 128-dimensional embedding's were collected for all speakers. These were then reduced to two dimensions using t-SNE with a perplexity of 30 and a random state of 42 for reproducibility.

The resulting 2D scatter plot revealed distinct clusters corresponding to individual speakers**,** with minimal overlap. This observation supports the claim that the CNN learned meaningful speaker-specific features, where intra-speaker embedding's were tightly grouped and inter-speaker embedding's were well-separated. Such visualization provides interpretability to the deep learning model, which is particularly valuable in forensic applications where understanding and explaining decisions is critical [29], [30].

D. Confusion Matrix Analysis

The confusion matrix demonstrates that the CNN model achieves highly reliable speaker identification, with most test segments correctly classified along the diagonal and only a few off-diagonal errors. Although each speaker contributes only ten test audio recordings, the cell values exceed ten because every audio file is divided into multiple 3-second **mel-spectrogram segments**, and each segment is treated as an independent test instance; thus, a single speaker often contributes 20–30 test segments, which naturally increases the counts in each cell. The diagonal cells appear darker, indicating a high number of correctly classified segments, while the few lighter off-diagonal cells represent minor misclassifications. These limited errors occur mainly among same-gender speakers, reflecting natural acoustic similarity within gender groups, whereas cross-gender confusion is almost absent, and indicating strong separation of male and female vocal characteristics. Overall, the confusion matrix confirms the model's strong discriminative power, well-separated speaker embedding's, and dependable performance suitable for forensic applications.

The most significant feature is the diagonal line of dark blue squares running from the top-left to the bottom-right; these represent correct predictions where the model's guess matched the actual speaker, such as correctly identifying Aniket Sonkawade 14 times or Pratibha Bhise 15 times. The model is extremely accurate, as evidenced by the fact that almost all values lie on this diagonal. There are only two visible errors in the entire grid: one instance where the speaker Sandip Thorat was misclassified as Pawan Kamble, and one where Shital Wadhai was mistakenly identified as Karuna Kirwale. Aside from these two isolated mistakes, the empty white space in the rest of the grid confirms that the system achieved near-perfect precision across all 20 speakers.

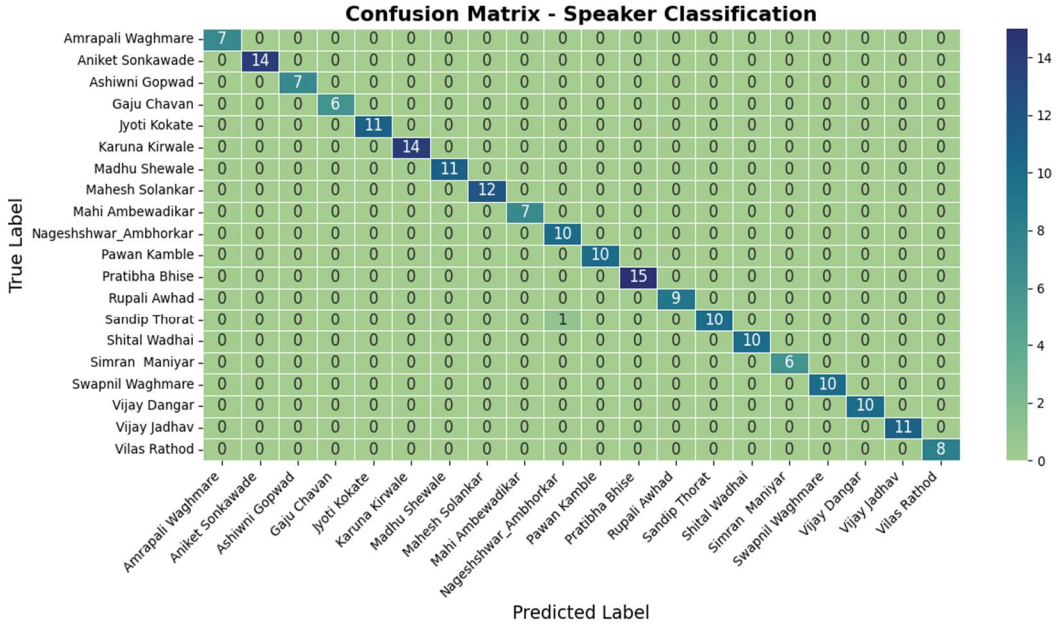


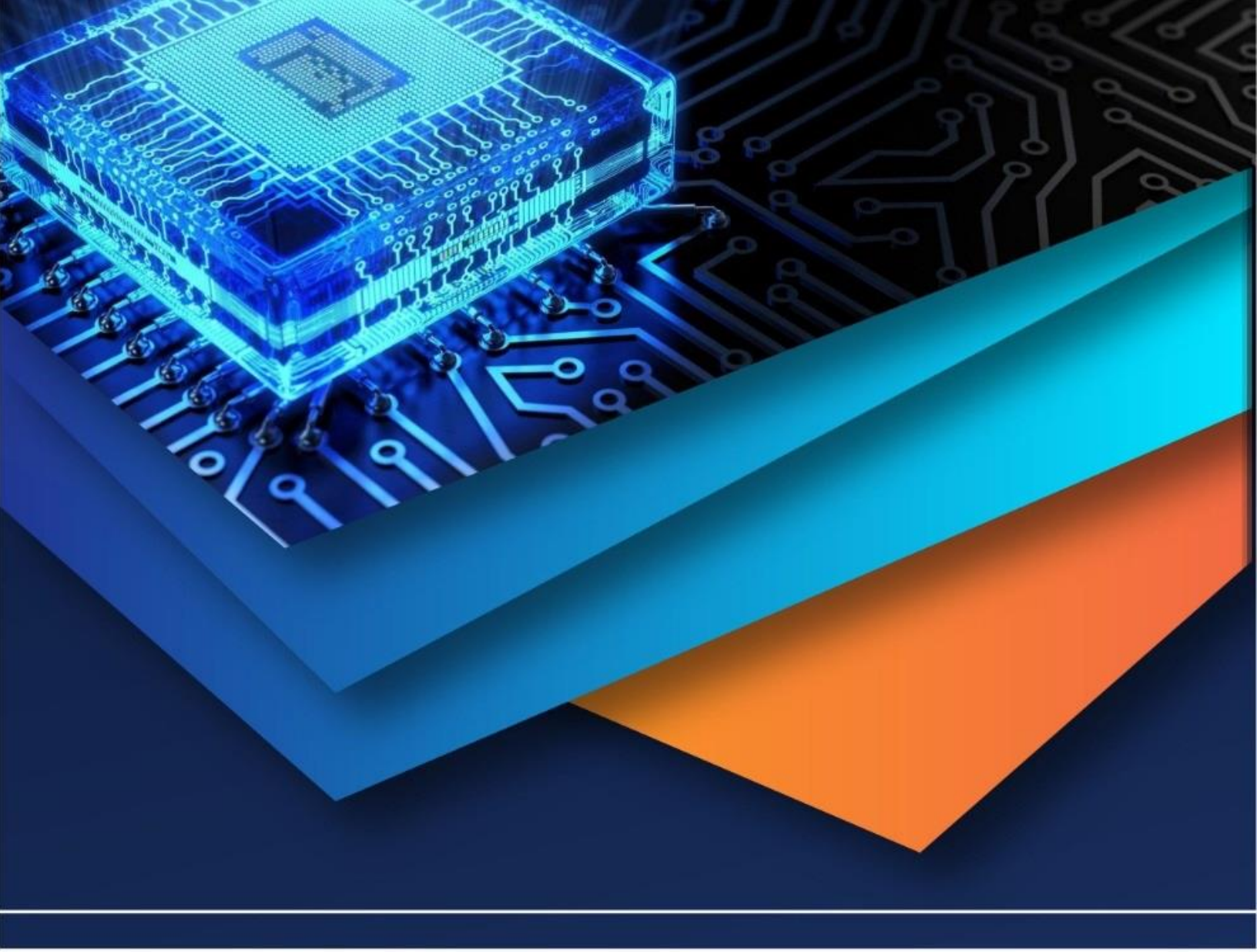
Figure 4 Confusion matrix

V. CONCLUSION

This research presents a deep learning-based framework for forensic speaker recognition, utilizing CNNs and mel spectrograms to effectively classify speakers and extract meaningful speaker embedding's. The system demonstrates high classification accuracy and robust performance in representing individual vocal characteristics, making it a promising approach for forensic applications. A key strength of the framework is its ability to capture and visualize intra- and inter-speaker variability through embedding analysis and t-SNE projection. The results confirm the model's potential to support speaker identification and differentiation, which are essential in forensic investigations.

REFERENCES

[1] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N Kingsbury, B. (2012), Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
[2] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329-5333). IEEE.
[3] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. In INTERSPEECH 2017 (pp. 2616-2620).
[4] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).
[5] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10(1-3), 19-41.
[6] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788-798.
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016) "Deep residual learning for image recognition". In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[8] O'Shaughnessy, D. (2000). Speech communications: human and machine. IEEE Press.
[9] Fougeron Cecile., (2022), Intra-speaker phonetic variation in read speech: comparison with inter-speaker variability in a controlled population. DOI: 10.21437/Interspeech.2022-10965.
[10] Supaporn Bunrit., (2019), Text-Independent Speaker Identification Using Deep Learning Model of Convolution Neural Network. International journal of Machine Learning and computing, vol.9, No. 2, 143-148.
[11] Volker Dellwo., Adrian Leemann., Marie-Jose Kolly., (2014), The Recognition of read and spontaneous speech in local vernacular: The Case of Zurich German. Journal of Phonetics
[12] Andrej Drygajlo, "Automatic Seaker Recognition for Forensic Case Assessment and Interpretation", Forensic Speaker Recognition 2012.
[13] Tharmarajah Thiruvaran, Eliathamby Ambikairajah, Jukien Epps, "FM Features for Automatic Forensic Speaker Recognition", ISCA 2008
[14] L.A.Khan, M.S.Bai G, Amr M. Youssef, "Speaker Recognition from Encrypted VoIP Communications", ELSEVIER 2009
[15] Enrico Marchetto, Federico Avanzini, Federico Flego," An Automatic Speaker Recognition System for Intelligence Applications", EURASIP 2009
[16] S. Malik, Fayyaz A. Afsar, "Wavelet Transform Based Automatic Speaker Recognition", IEEE 2009
[17] Joseph P.Campbell, Wade Shen, "Forensic Speaker Recognition", 2009
[18] Tomi Kinnunen, Haizhou Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", 2009
[19] Vibha Tiwari, "MFCC and its applications in speaker recognition", 2010
[20] M.G.Sumithra, K. Thanuskodi and A.Helen Jenifer Archana, "A New Speaker Recognition System with Combined Feature Extraction Techniques", Journal of Computer Science 2011
[21] Miranti Indar Mandasari, Mitchell Mclaren, David A. Van Leuven, "The Effect of Noise on Modern Automatic Speaker Recognition Systems", IEEE 2012
[22] Andrej Drygajlo, "Automatic Speaker Recognition for Forensic Case Assessment and Interpretation", 2012
[23] Homayoon Beigi, "Speaker recognition: Advancements and Challenges", 2012
[24] Parul, R.B. Dubey, "Automatic Speaker Recognition System", 2012
[25] Karthik Selvan, Aju Joseph, Anish Babu K.K., "Speaker Recognition System for Security Application", IEEE Recent Advances in Intelligent Computational System 2013
[26] Najiya Abdualrahiman , Ranju K.V., " Text Dependent Speaker Recognition",2013.
[27] Omid Ghahabi, Javieer Hernando, "Deep Belief Networks for I- Vector Based Speaker Recognition", IEEE 2014
[28] Ehsan Varini, Xin Lei, Erik Mcdermott, Ignacio Lopez Moreno, Javier Gonzalez-Dominuez, "Deep Neural Network for Small Footprint Text-Dependent Speaker Verification", IEEE International Conference on Acoustic, Speech and Signal Processing 2014
[29] Freds Richardson, "Deep Neural Network Approaches To Speaker and Language Recognition", IEEE 2015
[30] Miss.Sarika S. Admuth, Mrs.Shubhada Ghugardare," Survey Paper on Automatic Speaker Recognition Systems", 2015



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY