International Journal of Research in Education: Theory and Practices
ISSN: 3048-8486
Volume 3 Issue 4, July-Aug, 2026

A Comparative Machine Learning Framework for Speaker Identification Using Audio Biometric Features

B.Pradeep[1], Vijaya Mallamari[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

Abstract. Speaker identification has become an essential component of modern biometric authentication systems because voice characteristics provide a convenient and non-invasive method for verifying human identity. The rapid development automatic speaker identification. Audio recordings from the VoxCeleb dataset are preprocessed to eliminate unwanted noise and silence before extracting Mel Spectrogram features that effectively represent speech characteristics. The extracted features are used to train and evaluate each classification model using standard performance measures such as accuracy and F1-score. Experimental observations indicate that the SVM classifier delivers the highest recognition performance among the evaluated models, while CNN also achieves competitive results. In contrast, the LSTM model records comparatively lower performance because of the sequential complexity of the available dataset. The study demonstrates the importance of selecting an appropriate learning algorithm for speaker recognition applications and provides valuable insights for developing reliable and efficient voice-based biometric systems.

keywords: Speaker Identification, Audio Biometrics, Machine Learning, Deep Learning, Support Vector Machine, Convolutional Neural Network, Long Short-Term Memory, Mel Spectrogram, Speech Recognition, VoxCeleb Dataset.

I. Introduction

Voice has emerged as one of the most dependable biometric characteristics for authenticating individuals because every person possesses unique vocal attributes that are difficult to imitate accurately. With the growing demand for intelligent authentication systems, speaker identification has gained considerable attention in applications such as secure access control, banking services, digital assistants, surveillance systems, forensic investigations, and smart home technologies. Unlike conventional biometric methods such as fingerprints or facial recognition, voice authentication offers a contact-

free and user-friendly solution that can be implemented remotely without requiring specialized hardware.

Recent advancements in artificial intelligence have transformed speech processing by introducing meaningful representations from speech signals and achieve higher recognition accuracy. Feature extraction methods such as Mel Spectrograms have become highly effective because they capture the frequency characteristics of human speech in a manner similar to human auditory perception.

This research focuses on possesses unique learning capabilities for handling speech information. SVM performs efficient classification in high-dimensional feature spaces, CNN extracts spatial representations from spectrogram images, while LSTM learns temporal dependencies from sequential audio data. Their comparative analysis provides a comprehensive understanding of their strengths and limitations in speaker identification tasks.

The VoxCeleb dataset is selected as the experimental benchmark because it contains speech recordings collected under realistic environmental conditions from numerous speakers. ration to improve feature quality. Performance evaluation is carried out using Accuracy and F1-score, allowing an objective comparison among the selected algorithms.

The primary objective of this work is to determine the most suitable machine learning model for reliable speaker identification while maintaining computational efficiency. The findings contribute to the development of robust voice biometric systems capable of supporting future intelligent authentication applications.

II. Literature Survey

Automatic speaker identification has experienced rapid growth due to continuous advancements in machine learning and deep learning technologies. Early research primarily relied on statistical modeling), which extracted handcrafted speech features for identifying speakers. Although these techniques achieved reasonable performance, their effectiveness decreased under noisy and real-world recording conditions.

The introduction of Support Vector Machine (SVM) classifiers significantly improved speaker recognition accuracy by providing effective decision boundaries in high-dimensional feature spaces. SVM-based systems demonstrated excellent classification capability when combined with discriminative speech representations such as . Their ability to generalize well with limited training data made them highly suitable for biometric authentication tasks.

speech analysis because spectrogram representations can be interpreted similarly to images. CNN architectures automatically learn hierarchical feature representations from speech signals without requiring extensive manual feature engineering. Several

researchers reported substantial improvements in recognition accuracy by utilizing multiple convolutional layers for extracting discriminative voice characteristics.

Long Short-Term Memory (LSTM) networks further expanded speech recognition research by modeling temporal relationships within sequential audio signals. Unlike conventional neural networks, LSTM retains long-term contextual information through specialized memory cells, enabling effective processing of continuous speech sequences. However, LSTM performance often depends on the availability of sufficiently large and balanced datasets, making model training computationally demanding.

sequence modeling to improve recognition robustness. Furthermore, transfer learning and attention-based mechanisms have demonstrated promising results by utilizing pretrained speech models for various biometric applications.

Although considerable progress has been achieved, challenges including speaker variability, environmental noise, recording quality, multilingual speech, and spoofing attacks continue to affect recognition performance. Consequently, selecting an appropriate classification algorithm remains an important research problem. Motivated by these observations, this study performs a comparative evaluation of SVM, CNN, and LSTM models using the VoxCeleb dataset to determine the most effective approach for speaker identification.

III. Research Methodology

The proposed speaker identification framework follows a systematic sequence of operations beginning with audio acquisition and ending with performance evaluation. The methodology is designed to preserve the original speech characteristics while extracting discriminative features that enable accurate speaker classification.

Initially, the VoxCeleb dataset is employed as the primary source of speech recordings. The dataset contains audio samples collected from numerous speakers under realistic recording environments, providing sufficient diversity

Before feature extraction, all audio recordings undergo preprocessing to enhance signal quality. Silent intervals and unnecessary background noise are removed to eliminate irrelevant information that could negatively influence classification accuracy. Audio normalization is then applied to maintain consistent amplitude levels across all recordings. These preprocessing operations improve the quality of the extracted speech features while reducing variations caused by recording conditions.

Following preprocessing, Mel Spectrograms are generated from every audio recording. Thedomain, after which Mel filter banks transform frequency components into the perceptual Mel scale. This representation effectively captures the spectral properties of speech and serves as the primary input for machine learning algorithms.

The extracted Mel Spectrogram features are organized together with their corresponding speaker labels. Data is randomly shuffled before model training to avoid sampling bias and improve model generalization. Metadata describing speaker identities and feature information is also generated to facilitate efficient dataset management.

Three different classification models are trained independently using identical feature representations. The Support Vector Machine performs supervised classification by constructing an optimal hyperplane that separates speaker classes with maximum margin. The Convolutional Neural Network automatically extracts hierarchical spatial representations from Mel Spectrogram images using convolutional and pooling operations before performing classification through processes sequential speech information by learning temporal dependencies through memory cells capable of retaining long-term contextual information.

After training, each model is evaluated using the testing dataset. to assess prediction capability and class balance. The comparative analysis identifies the most suitable model for speaker identification while maintaining consistency with the experimental setup and evaluation procedure described in the original research. The obtained results produces comparatively lower recognition accuracy under the selected experimental conditions.

IV. Existing System

Traditional speaker identification systems primarilrecognize individuals based on their voice characteristics. These systems generally begin by collecting speech recordings from publicly available datasets such as VoxCeleb, followed by preprocessing operations including silence removal, audio normalization, and feature extraction. Mel Spectrograms are commonly generated from the processed speech signals because they effectively represent the frequency distribution of human voice. The extracted features are then supplied to classification While SVM often provides strong classification performance and CNN effectively learns spatial patterns from spectrogram images, LSTM is capable of modeling sequential speech information for speaker recognition. These approaches have significantly improved voice-based biometric authentication and are widely used in applications including access control, forensic analysis, voice assistants, and security systems.

Despite these advancements, existing speaker identification methods still encounter several practical challenges. Recognition accuracy can decrease when speech recordings contain background noise, channel distortions, or variations in speaking style and pronunciation. limited datasets, leading to reduced performance during testing. Furthermore, differences in language, accent, recording devices, and environmental conditions may negatively affect model generalization. These limitations highlight the need for selecting appropriate feature extraction methods and robust classification algorithms that can achieve higher recognition accuracy while maintaining computational efficiency in real-world speaker identification applications.

V. Proposed System

The proposed system presents an efficient speaker identification framework that integrates advanced audio preprocessing with multiple machine learning models to achieve reliable voice-based biometric recognition. Initially, speech recordings are collected from the VoxCeleb dataset, which contains audio samples from numerous speakers recorded under diverse real-world conditions. The raw audio signals are first subjected to preprocessing operations, including silence removal and amplitude normalization, to eliminate unnecessary noise and improve the overall quality of the speech data. After preprocessing, each audio sample is transformed into a Mel Spectrogram using the Short-Time Fourier Transform (STFT), enabling the extraction of meaningful frequency-domain features that accurately represent the unique vocal characteristics of each speaker. These feature representations are then organized and labeled to create structured training and testing datasets suitable for supervised learning.

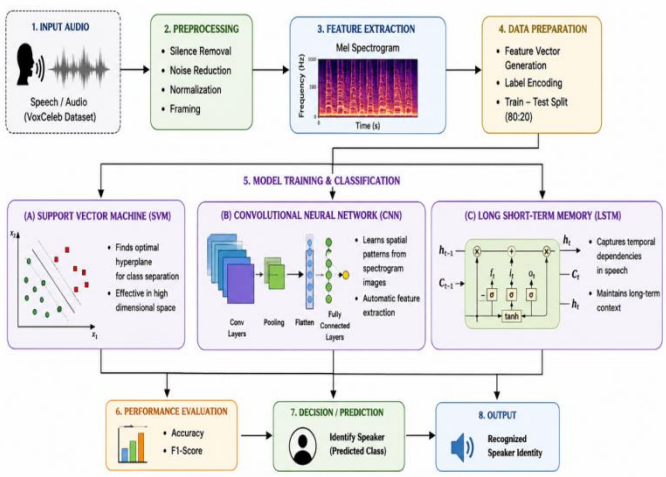
The extracted Mel Spectrogram features are subsequently provided as input to three classification models: SVM classifier separates speaker classes by constructing an optimal decision boundary in the feature space, providing strong classification capability even with high-dimensional data. The CNN model automatically learns hierarchical feature representations from spectrogram images through convolution and pooling operations, enabling effective recognition of complex speech patterns. The LSTM model captures temporal dependencies within sequential speech signals by utilizing memory cells that preserve contextual information over time. Each model is trained and evaluated using identical datasets to ensure a fair comparison of performance.

Following model training, the system predicts the identity of unknown speakers and evaluates the prediction quality using Accuracy and F1-score. The experimental results demonstrate that the Support Vector Machine achieves the highest overall recognition performance, while the CNN model also provides competitive classification accuracy. Although the LSTM network effectively models sequential speech information, its performance is comparatively lower under the selected experimental settings. By combining robust preprocessing, discriminative Mel Spectrogram feature extraction, and comparative evaluation of multiple learning algorithms, the proposed framework delivers a dependable and computationally efficient solution for speaker identification. This system can be effectively applied in various real-world domains, including biometric authentication, secure access control, forensic investigations, smart devices, voice-controlled applications, and intelligent security systems.

VI. System Architecture

The proposed speaker identification system follows a structured sequence of operations that transforms raw speech recordings into accurate speaker predictions using machine learning techniques. Initially, speech samples are collected from the VoxCeleb dataset, which contains voice recordings from multiple speakers recorded under different environmental conditions. These audio samples serve as the primary input to the

system. Before feature extraction, the recordings undergo preprocessing operations such as silence removal, background noise reduction, audio normalization, and framing. These preprocessing steps improve the overall quality of the speech signals by eliminating irrelevant information and ensuring that all audio samples have a consistent format suitable for analysis.



After preprocessing, the refined Spectrogram provides a visual representation of speech frequencies based on the human auditory system and effectively captures unique vocal characteristics required for speaker recognition. The extracted features are then organized into feature vectors, and corresponding speaker labels are assigned. The complete dataset is divided into training and testing subsets using an 80:20 split, allowing the models to learn from one portion of the data while evaluating their performance on previously unseen samples.

The prepared feature vectors are supplied independently to three classification models: Support Vector Machine (SVM), Convolutional Neural NetwoSVM classifier identifies the optimal decision boundary that separates different speaker classes with maximum accuracy. The CNN model automatically extracts high-level spatial patterns from Mel Spectrogram images through multiple convolutional and pooling layers, enabling efficient speaker classification. The LSTM network processes sequential speech information by learning temporal dependencies and preserving contextual information using memory cells, making it suitable for analyzing speech sequences. Each model is trained using identical datasets to ensure a fair and reliable performance comparison.

Once the training phase is completed, the trained models classify unknown speech samples by predicting the corresponding speaker identity. The effectiveness of each model is assessed using evaluation metrics such as Accuracy and F1-score, which measure classification correctness and overall model reliability. Based on the compar-

ative analysis, the Support Vector Machine demonstrates the highest recognition performance, followed by comparatively lower accuracy under the selected experimental conditions. suitable for practical applications including biometric authentication, access control, forensic investigations, intelligent voice assistants, and secure speech-based verification systems.

VII. Conclusion

This research presents a comprehensive comparative analysis of machine learning and deep learning techniques for speaker identification using audio biometric data. The proposed framework employs systematic audio preprocessing, Mel Spectrogram feature extraction, and three classification models— recognize speakers from voice recordings. The experimental evaluation highest recognition accuracy and F1-score among the selected models, while the CNN model also provides strong and reliable performance. Although the LSTM network is capable of learning temporal speech patterns, its performance is comparatively lower under the current experimental setup. The results indicate that selecting an appropriate feature representation together with a suitable classification algorithm plays a crucial role in achieving accurate speaker recognition. Overall, voice-based biometric authentication. The developed framework can be effectively applied in various real-world applications, including secure access control, digital identity verification, forensic investigations, intelligent voice assistants, banking authentication, and other speech-enabled security systems. Furthermore, the findings of this study provide a valuable foundation for future research aimed at improving speaker recognition accuracy through larger datasets, advanced deep learning architectures, transfer learning techniques, and robust noise-resistant feature extraction methods while maintaining high reliability in practical deployment scenarios.

References

1. N. Ohi, M. A. Hamid, M. F. Mridha, and M. "Deep Speaker Recognition: Process, Progress, and Challenges," by M. Monowar, IEEE Access, vol. 9, 89619–89643, 2021, doi: 10.1109/ACCESS.2021.3090109.
2. P. R. and Deepa. Khilar, "A Report on Voice Recognition System: Deep Neural Network Techniques, Methodologies, and Challenges," IEEE International Conference on Innovations in Power and Advanced Computing Technologies (i-PACT), Proceedings 2021, pp. 1–5, doi: 10.1109/i-PACT52855.2021.9697005.
3. "Deep Learning-Based Speaker Identification and Verification," N. R., R. Tanguturu, J. R. C., and A. K. M., IEEE International Conference on Signal and Information Processing (IConSIP), 2022, pp. 1–6; Proceedings 10.1109/ICoNSIP49665.2022.10007520.
4. "Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques," in Proceedings, N. M. Habibullah, N. Prachi, F. M. Nahiyan, and R. Khan. 10.1109/IRTM54583.2022.9791766 IEEE International Conference on Interdisciplinary Research in Technology and Management (IRTM), 2022, pp. 1–6.

5. S. A. Ghosh, N. Sarkar, A. K. Laskar, and R. "Speaker Recognition through Deep Learning Techniques: A Comprehensive Review and Research Challenges," H. Kashyap, Periodica Polytechnica Electrical Engineering and Computer Science, vol. 67, no. 3, 2023, pp. 300–336.
6. M. Y. and Ravanelli Bengio, "SincNet for Speaker Recognition from Raw Waveform," IEEE Signal Processing Letters, vol. 28, pp. 1020–1024, 2021.
7. "VoxCeleb2: Deep Speaker Recognition," J. A. Nagrani, S. Chung, and A. Zisserman, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, 2022, pp. 2984–2997.
8. "X-Vectors: Sturdy DNN Embeddings for Speaker Identification," D. Snyder, D. Povey, G. Sell, D. Garcia-Romero, and S. Khudanpur, IEEE Transactions on Audio, Speech, and Language Processing, vol. 30, 242-254, 2022.
9. H. Zeinali, L. Burget, and J. Cernocky, "A Thorough Analysis of Deep Learning for Text-Independent Speaker Validation," IEEE Access, vol. 10, 2022, pp. 18762–18780.
10. "Developments in Deep Neural Network-Based Speaker Verification," M. McLaren, D. Castán, L. Ferrer, and A. Lawson, IEEE Signal Processing Magazine, vol. 39, no. 5, 2022, pp. 62–76.
11. "ESPnet: End-to-End Speech Processing Toolkit," Nishitoba, IEEE Access, vol. 10, 22044–22060, 2022, S. Watanabe, T. Hori, T. Hayashi, S. Karita, and J.
12. "Recent Developments in Deep Learning-Based Speaker Recognition," Neural Computing and Applications, vol. 35, no. 4, 2023, pp. 2791–2810, Y. Qian and K. Liu. Yu.
13. "State-of-the-Art Speaker Recognition Using Deep Embedding Models," Dehak, Computer Speech & Language, vol. 81, Art. No. 101531, 2023; J. Villalba, N. Chen, G. Sell, D. Snyder, and N. [14] "Speaker Verification and Recognition: Current Advances and Prospects," Information Fusion, vol. 92, 2023, T. Kinnunen and H. Li.
14. X. Wang, Y. Wang, and J. Li, "Robust Speaker Identification through Attention-Based Deep Learning Framework," Expert Systems with Applications, vol. 230, Art. No. 120540, 2024.
15. Z. Zhang, L. Chen, and H. Zhao, “Hybrid CNN–LSTM Architecture for Automatic Speaker Recognition,” Applied Soft Computing, vol. 146, Art. no. 110654, 2024.
16. A. Kumar, R. Sharma, and P. Singh, “Deep Learning Approaches for Audio Biometrics: A Survey,” IEEE Access, vol. 12, pp. 21544–21570, 2024.
17. M. Ferrag, L. Maglaras, H. Janicke, and A. Ahmim, “Artificial Intelligence Techniques for Secure Biometric Authentication: A Survey,” Journal of Information Security and Applications, vol. 86, Art. no. 104115, 2025.
18. S. Li, Y. Zhou, and H. Wang, “Transformer-Based Speaker Recognition with Self-Supervised Learning,” Pattern Recognition, vol. 156, Art. no. 110875, 2025.
19. R. Gupta, V. Sharma, and N. Mehta, “Machine Learning and Deep Learning Models for Robust Speaker Identification: A Comparative Analysis,” Engineering Applications of Artificial Intelligence, vol. 136, Art. no. 109214, 2025.