Voice-Based Biometric System: One-Shot Learning for Unseen Speaker Generalization

Shreyas Nagoor[1], Garima Pandey[2]*, and Shashidhar G. Koolagudi[2]

[1] The National Institute of Engineering, Mysore, India
shreyasnagoor@gmail.com
[2] National Institute of Technology, Karnataka, Surathkal, India
garimapandey.217cs002@nitk.edu.in, koolagudi@nitk.edu.in

Abstract. Voice authentication is a crucial aspect of biometric security, particularly in sensitive applications such as financial transactions and secure communication. However, a major challenge lies in the need for scalable, secure, and interpretable solutions that perform well with minimal enrollment data, especially for unseen speakers. This study explores the use of one-shot learning for voice authentication by leveraging a custom Siamese network architecture. The network integrates ResNet-18 for feature extraction, GRU layers for sequential modeling, and an attention mechanism to refine embeddings for speaker verification. The model employs triplet margin loss to enable generalization to new, unseen speakers using just a single example per speaker, making it ideal for one-shot learning. When evaluated on the LibriSpeech dataset, the model demonstrates its ability to handle the inherent variability of real-world voice data. Extensive data augmentation is applied to simulate challenging acoustic conditions, further enhancing the model's robustness. These results underscore the potential of one-shot learning architectures to deliver secure, scalable, and efficient voice authentication solutions, particularly for high-stakes applications such as financial transactions and secure communication.

Keywords: Voice Authentication, Speaker Verification, Siamese Network, One-Shot Learning, Biometric Security

1 Introduction

Voice authentication, a specialized subset of speaker recognition, is a critical biometric technology used to verify an individual's identity based on unique vocal characteristics. Its non-intrusive nature, combined with its versatility across applications such as financial transactions, secure communications, and personalized digital assistants, has made it a prominent tool in modern security systems. Unlike other biometric modalities, such as fingerprint or facial recognition, voice authentication offers the advantage of remote and contactless verification, making it highly relevant in today's digital landscape. Voice authentication systems primarily perform two tasks: speaker verification and speaker identification.

Speaker verification is a 1:1 matching process that determines whether a given voice matches a claimed identity, ensuring secure access for an individual. In contrast, speaker identification involves a 1:N process to identify the speaker from a database of enrolled users. Both processes rely on sophisticated algorithms to analyze vocal features, capturing unique patterns of pitch, tone, rhythm, and pronunciation. Feature extraction plays a central role in voice authentication systems, enabling the differentiation between speakers. One of the most widely used techniques is Mel-Frequency Cepstral Coefficients (MFCCs), which effectively represent the spectral properties of speech by modeling the shape and dynamics of the vocal tract. MFCCs not only reduce the dimensionality of voice data but also provide robust representations that are less affected by environmental noise, making them a cornerstone of modern speaker verification systems.

The introduction of deep learning has revolutionized voice authentication by enabling systems to model complex temporal and spectral patterns in speech. Architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have demonstrated superior performance in learning hierarchical and sequential voice features. When trained on large datasets, these models significantly improve the accuracy and scalability of speaker recognition systems, even under challenging conditions. Despite these advancements, several challenges persist. Variations in voice due to factors such as emotional states, health conditions, and recording environments can degrade system performance. Additionally, spoofing attacks, such as replay attacks and synthetic voice impersonation, pose significant security risks. To address these issues, researchers have increasingly focused on integrating liveness detection mechanisms, which differentiate between genuine and spoofed voices, thereby enhancing the robustness of voice authentication systems.

In this context, this study aims to tackle these challenges by developing a custom Siamese network architecture for one-shot learning in voice authentication. Leveraging ResNet-18 for feature extraction, GRU layers for sequential modeling, and an attention mechanism for refining embeddings, the proposed model is designed to generalize effectively to unseen speakers with minimal enrollment data. Extensive data augmentation techniques are applied to simulate real-world conditions, ensuring robustness against variations in voice and environmental factors. In summary, voice authentication combines advanced feature extraction methods and deep learning architectures to provide secure and scalable solutions for biometric security. This research builds on these foundations to address real-world challenges, offering a framework for reliable voice authentication in high-stakes applications such as financial transactions and secure communication. The remaining paper is structured as follows: Section 2 reviews the related literature, Section 3 defines the problem statement and outlines the proposed methodology, Section 4 presents the results, and Section 5 concludes the paper.

2 Literature Survey

The field of voice authentication has witnessed significant advancements over the years, transitioning from traditional techniques to modern deep learning approaches. Early methods predominantly relied on handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), which capture the spectral characteristics of speech signals [1]. These features were commonly used in conjunction with statistical models like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). GMM-based models, as demonstrated by Suvarna et al. [2], utilized Gaussian distributions to represent speakers and employed the Expectation-Maximization (EM) algorithm for classification. Despite their foundational role, these methods struggled with scalability and robustness to environmental variations.

The advent of machine learning introduced algorithms such as K-Nearest Neighbors (KNN) and Artificial Neural Networks (ANNs) for speaker recognition. KNN, known for its simplicity, offered shorter training times, making it preferable in specific scenarios. However, ANN-based methods demonstrated superior performance when combined with MFCCs and Linear Predictive Coding (LPC) features, particularly in text-independent settings.

The introduction of deep learning marked a turning point in voice authentication. Convolutional Neural Networks (CNNs) emerged as effective tools for extracting local spectral features from speech. Jalil et al. [3] demonstrated the robustness of CNNs by using Mel-spectrogram inputs to classify speakers in noisy environments. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks further improved performance by capturing temporal dependencies in speech signals. El-Moneim et al. [4] achieved high accuracy in text-independent speaker recognition tasks by employing LSTM-RNN architectures with MFCC and log-spectrum features as inputs.

Recent advancements have focused on hybrid and ensemble techniques to improve accuracy and robustness. For instance, hybrid feature extraction methods combining Discrete Wavelet Transform (DWT) and Principal Component Analysis (PCA) have been employed to reduce dimensionality while retaining essential information. These techniques, as explored by Feras et al., enhanced model performance by addressing variability in speaker signals. Similarly, vector quantization techniques provided concise representations of MFCC features, reducing computational costs.

Deep learning-based architectures have also explored novel mechanisms such as attention layers, which focus on the most critical parts of the speech signal. Attention mechanisms enhance embedding quality by emphasizing relevant time steps, leading to improved verification and identification accuracy. Studies like those by Bai et al. [5] have highlighted the growing reliance on attention-driven architectures in speaker identification tasks.

Benchmark datasets such as LibriSpeech and VoxCeleb have become standard for evaluating speaker identification models. LibriSpeech, with its diverse speaker demographics and high-quality audio, provides a controlled environment for testing [6]. In contrast, VoxCeleb introduces real-world challenges, such as

background noise and device variations, making it an essential benchmark for evaluating robustness [7]. Despite advancements, speaker identification systems still face challenges in generalizing to diverse speakers and handling environmental variability. Differences in accents, languages, demographics, noise levels, and room acoustics often degrade performance, while computationally intensive architectures hinder real-time deployment.

The proposed Siamese network architecture effectively addresses these challenges. By leveraging ResNet-18 for robust feature extraction and GRU layers for sequential modeling, the model generates efficient and discriminative speaker embeddings. The use of triplet margin loss enables one-shot learning, facilitating generalization to unseen speakers with minimal enrollment data. Environmental robustness is achieved through extensive data augmentation, including Gaussian noise addition and pitch shifting, which simulate diverse acoustic conditions. Additionally, the architecture is optimized for computational efficiency. When evaluated on the LibriSpeech dataset, the model demonstrates high accuracy and scalability, advancing the state of speaker identification systems.

3 Methodology

3.1 Problem Statement

Voice authentication has become a crucial aspect of biometric security, particularly in applications such as financial transactions, secure communications, and personalized digital services. The core challenge lies in accurately verifying a speaker's identity based on unique vocal characteristics, especially when only limited enrollment data is available and the system encounters entirely unseen speakers. Additionally, real-world scenarios introduce complexities such as background noise, variability in speaker attributes, and environmental distortions, all of which pose significant challenges to effective voice authentication.

Conventional approaches to voice authentication often struggle to generalize to speakers not encountered during training, particularly in conditions with minimal data availability. To address these challenges, this work introduces a Siamese network-based architecture specifically designed for one-shot learning, enabling efficient speaker verification with minimal enrollment samples. The proposed framework integrates ResNet-18 for robust feature extraction, GRU layers to model temporal dependencies, and an attention mechanism to enhance the discriminative power of the generated speaker embeddings.

To ensure robustness against diverse acoustic environments, data augmentation techniques such as Gaussian noise, time stretching, and pitch shifting are applied. These augmentations not only simulate real-world acoustic variability but also improve the model's ability to handle variations in speaker characteristics and environmental conditions. The architecture is designed to be scalable and adaptable, making it suitable for deployment in practical scenarios involving dynamic and unpredictable audio conditions.

This study aims to bridge the gap between research and practical deployment by addressing key challenges in voice authentication, including generalization to

unseen speakers, robustness under variable conditions, and efficient learning with minimal enrollment data.

3.2 Dataset

In this study, the widely used LibriSpeech dataset serves as the foundation for training and evaluating the voice authentication model. Specifically, the train-clean-100 subset is utilized, consisting of approximately 100 hours of high-quality audio recordings from 251 speakers. This subset ensures diversity in speaker demographics, including variations in age, gender, and accents, making it well-suited for voice authentication tasks[**?**].

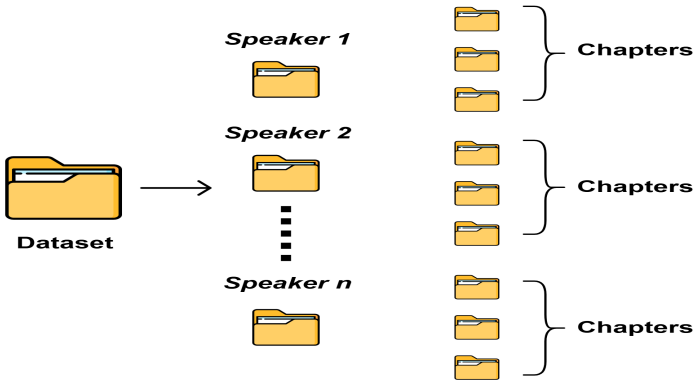

Fig. 1. Structure of the Librispeech Dataset

Data Preprocessing: To ensure uniformity in audio input, all speech recordings were standardized to a sampling rate of 16 kHz, aligning with the model's requirements[8]. Following resampling, the audio waveforms were transformed into Mel-Frequency Cepstral Coefficients (MFCCs) to effectively capture the spectral and dynamic characteristics of the vocal tract. MFCCs serve as a compact and informative feature representation, aiding in speaker differentiation.

To further improve the model's generalization capabilities under real-world conditions, a series of data augmentation techniques were applied. Gaussian noise was added to simulate environments with background disturbances, enhancing robustness to noise. Pitch shifting was employed to account for variations in tone and speaker characteristics, facilitating better generalization across different speaking styles. Additionally, time stretching was used to model changes in speaking rates, ensuring the system remains invariant to temporal distortions in speech. These augmentations play a crucial role in enhancing the model's resilience to acoustic variability, allowing it to perform well across diverse scenarios.

Dataset Splits: For training and evaluation, the dataset was split into 80% training and 20% validation subsets. The training set includes data from 80% of the speakers, providing sufficient diversity for the model to learn robust and discriminative embeddings. The validation set, comprising data from the remaining 20% of speakers, ensures that the model's performance is evaluated on entirely unseen identities, thereby testing its generalization capabilities.

Triplet Dataset Generation: To train the Siamese network, a custom Triplet-Dataset class was implemented to generate triplets of audio samples. Each triplet consists of an anchor sample, a positive sample from the same speaker, and a negative sample from a different speaker. This triplet generation strategy enables the model to minimize the embedding distance between samples of the same speaker while maximizing the distance between samples from different speakers.

The class ensures that all samples are resampled to 16 kHz, transformed into MFCCs, and padded to a fixed sequence length for batch processing. Data augmentation techniques, such as noise addition, pitch shifting, and time stretching, are applied during training to simulate real-world variability.

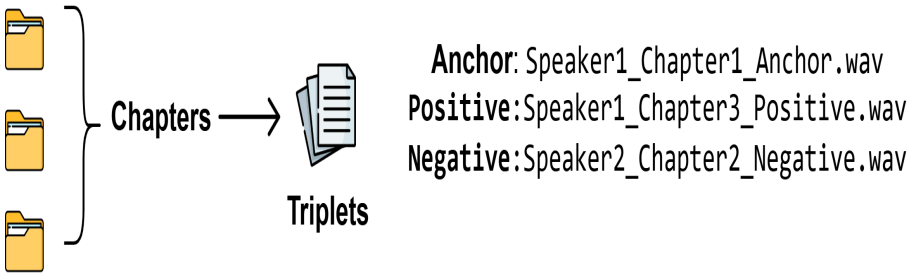

Fig. 2. Triplet Dataset Generation for Speaker Authentication

3.3 Feature Extraction

The raw audio signals are preprocessed to generate robust features for speaker verification. The feature extraction pipeline utilizes Mel-Frequency Cepstral Coefficients to capture the spectral characteristics of speech. Additionally, data augmentation and normalization are applied to enhance model robustness[9].

MFCC Extraction: The raw waveform is first resampled to a standard sampling rate of 16,000 Hz. MFCCs are then extracted using the `torchaudio.transforms.MFCC` function, with 40 coefficients per frame. The key parameters include:

- `n_mfcc=40`: Extracts 40 MFCC coefficients per frame.

– `n_fft=512`, `n_mels=80`, `hop_length=160`: Standard settings for the Fourier transform and mel filterbank.

This transformation captures the speaker's vocal tract characteristics in a compact form.

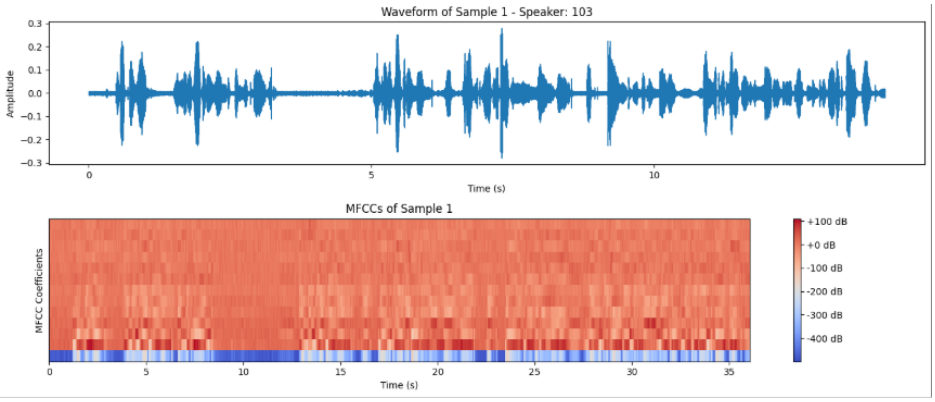

Fig. 3. Waveform and Mel-Frequency Cepstral Coefficients (MFCCs) of Sample 1 from Speaker 103.

Normalization: To ensure consistency across samples, the MFCCs are normalized by subtracting the mean and dividing by the standard deviation of each coefficient over time. This process ensures that each coefficient has a zero mean and unit variance, preventing any dimension from dominating due to scale differences.

Data Augmentation: To enhance the model's robustness, augmentation techniques such as Gaussian noise addition, time stretching, and pitch shifting are applied[10]. These augmentations simulate real-world variations in speech, such as background noise and speaking speed, helping the model generalize better to unseen data.

Padding: To handle variable-length inputs, MFCC sequences are either padded or truncated to a fixed length. If a sequence exceeds the maximum length, it is truncated; otherwise, it is padded with zeros. This ensures that all inputs to the model maintain a consistent length. This feature extraction pipeline transforms raw audio into normalized, augmented, and padded MFCC features, which are then used for speaker authentication. The resulting features are consistent, robust to real-world conditions, and well-suited for deep learning models.

3.4 Model Architecture for Voice Authentication

Overview: The proposed voice authentication system is built on a Siamese network architecture designed for one-shot learning. The model incorporates ResNet-18 for feature extraction, GRU layers for sequential modeling, and an attention mechanism for refining speaker embeddings. The network uses triplet loss to minimize the distance between embeddings of the same speaker while maximizing the distance between embeddings of different speakers. This architecture generalizes well with minimal enrollment data, making it suitable for real-world voice authentication applications.

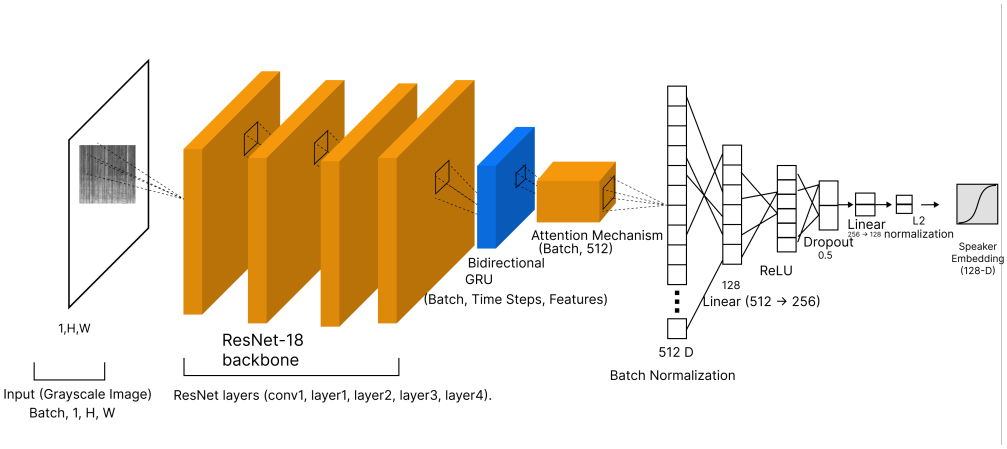

Fig. 4. Model architecture for the voice authentication system.

Siamese Network Framework: The architecture follows a Siamese network structure [11] with two identical branches that share weights. Given a pair of speech signals, the network computes embeddings for each input and compares them to determine if they belong to the same speaker. The training process optimizes the network to reduce the distance between embeddings of the same speaker while increasing the distance between embeddings of different speakers. ResNet-18 is employed for feature extraction from raw audio, while GRU layers capture sequential dependencies in speech. An attention mechanism further refines the final embeddings to enhance speaker identity representation. Triplet loss is used as the optimization criterion, leveraging anchor, positive, and negative sample comparisons to improve speaker discrimination.

Feature Extraction using ResNet-18: The first part of the network uses ResNet-18, a Convolutional Neural Network (CNN), for feature extraction from audio signals[12]. The audio input is transformed into a spectrogram or Mel-frequency cepstral coefficients (MFCCs), capturing both frequency and time-domain characteristics of the speech. The ResNet-18 model, with its residual

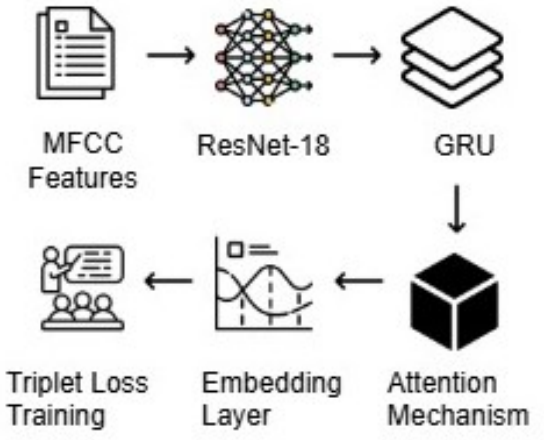

Fig. 5. Flow of the proposed speaker verification system, starting with MFCC feature extraction, followed by ResNet-18, GRU, and attention mechanism, optimized using triplet loss training.

connections, learns deep features from the input spectrogram without suffering from vanishing gradients. These layers extract hierarchical representations, capturing features from low-level to high-level. The network generates a feature vector that encapsulates the speaker's vocal characteristics, tone, pitch, and other speaker-specific patterns.

Sequential Modeling using GRU Layers: After the feature extraction phase, GRU (Gated Recurrent Units) layers model the sequential dependencies in the audio. Speech data has a temporal nature, making it crucial to capture these dependencies in order to understand the speaker's vocal patterns over time[13]. GRU layers process the sequence of features extracted by ResNet-18, modeling long-range dependencies that are essential for distinguishing speakers based on their speech patterns, rhythm, and intonation. GRU produces a sequence of embeddings, each corresponding to different time frames of the speech signal.

Attention Mechanism for Embedding Refinement: To enhance the representation learned by the network, an attention mechanism is applied to the embeddings produced by the GRU layers[14]. The attention mechanism focuses on the most relevant parts of the speech signal, such as distinctive features that best define a speaker's identity. Self-attention computes a set of attention weights for each element in the sequence, emphasizing the most informative parts of the

audio while ignoring noise or irrelevant features. The weighted sum of the embeddings is then computed, yielding a refined context vector that represents the speaker's identity in a more compact and discriminative manner.

Embedding Projection and Normalization: Once the embeddings are refined by the attention mechanism, they are passed through a fully connected projection layer to produce the final embedding used for speaker verification[15]. This layer maps the final embeddings to a fixed-dimensional space (e.g., 256 dimensions), ensuring that the embeddings of different speakers are well separated. The embeddings are then normalized to ensure consistency and prevent any dimension from dominating the embedding space due to scale differences.

Loss Function (Triplet Loss): The model is trained using triplet loss, which compares three samples: an anchor, a positive, and a negative sample[16]. The anchor and positive samples come from the same speaker, while the negative sample comes from a different speaker. The goal of triplet loss is to ensure that the anchor and positive samples are closer in the embedding space, while the anchor and negative samples are farther apart.

$$\mathcal{L} = \max(d(a,p) - d(a,n) + \alpha, 0) \tag{1}$$

Here, $d(a,p)$ is the distance between the anchor and positive samples, $d(a,n)$ is the distance between the anchor and negative samples, and α is the margin that enforces a minimum separation between positive and negative samples. By minimizing this loss, the model maps speech signals to a high-dimensional space where similar speakers are close together and dissimilar speakers are far apart.

Cosine Similarity for Verification: For speaker verification, the cosine similarity is computed between the embeddings of two input audio samples:

$$\text{cosine\_similarity}(x,y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{2}$$

Here, x and y are the embeddings of the two input samples, and $\|\cdot\|$ denotes the Euclidean norm. A higher cosine similarity indicates that the two samples belong to the same speaker, while a lower similarity indicates they are from different speakers[17].

The proposed architecture combines ResNet-18 for feature extraction, GRU layers for modeling sequential dependencies, and an attention mechanism for refining speaker embeddings. By using triplet loss during training and cosine similarity for verification, the model achieves high accuracy in speaker verification tasks with minimal enrollment data. The model generalizes to unseen speakers and handles variations in speech, making it robust for real-world voice authentication applications.

4 Results and Analysis

The proposed Siamese network for voice authentication was evaluated comprehensively to assess its effectiveness in distinguishing between speakers. The training process, which utilized the triplet margin loss function, exhibited consistent improvement across both training and validation phases. Over the course of 10 epochs, the training loss steadily decreased from an initial value of 0.7925 to 0.4405, while the validation loss reduced from 0.5425 to 0.3767. These findings indicate that the model successfully learned discriminative speaker features while maintaining strong generalization to unseen speakers. Figure 6 provides a graphical representation of the training and validation losses, showcasing the stability and convergence of the network.

Fig. 6. Training and Validation Loss Over Epochs

The evaluation of the speaker verification task was conducted using the validation set from the LibriSpeech dataset, which includes diverse acoustic conditions and speaker characteristics. The model achieved an area under the Receiver Operating Characteristic curve (ROC-AUC) of 0.85, demonstrating its robustness in distinguishing between same-speaker and different-speaker pairs. The Operating Characteristic (ROC) curve in Figure 7 underscores the efficacy of the proposed voice authentication model in distinguishing between legitimate users and impostors. With an Area Under the Curve (AUC) of 0.8517, the model demonstrates robust performance, significantly surpassing the baseline of a random classifier (AUC = 0.5). The steep initial rise of the curve at low false positive rates (FPR) reflects the system's high sensitivity, enabling it to correctly authenticate genuine users (high true positive rate, TPR) while minimizing false

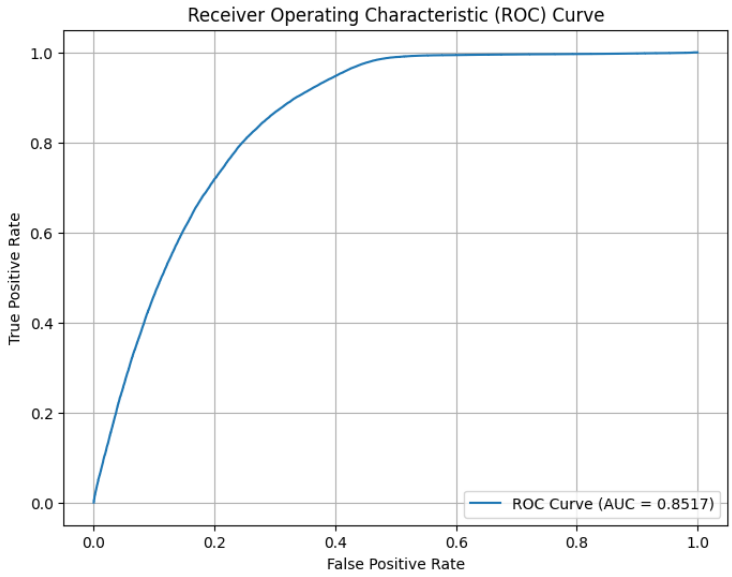

Fig. 7. Receiver Operating Characteristic (ROC) Curve

acceptances—a critical feature in security-sensitive applications such as voice-based access control or biometric authentication systems.

In voice authentication, minimizing false positives (incorrectly accepting impostors) is paramount to prevent unauthorized access, while maintaining high true positives ensures a seamless user experience by reducing false rejections of legitimate users. The AUC value indicates a balanced trade-off between these metrics, suggesting the model reliably differentiates between authentic and fraudulent voice samples. This performance aligns with industry standards for biometric systems, where an AUC above 0.8 is often considered competitive for real-world deployment.

To gain deeper insights into the model's performance, a confusion matrix was generated. The results, summarized in Figure 8, show that the system achieved an overall accuracy of 85.47%. This accuracy reflects the model's effectiveness in distinguishing between the "Same" and "Different" classes. Specifically, the model achieved a true positive rate (TPR) of 92.18% for the "Same" class and a true negative rate (TNR) of 77.48% for the "Different" class, indicating strong performance across both classes.

The results validate the model's overall performance, demonstrating its ability to accurately distinguish between the "Same" and "Different" classes. While the model exhibits strong performance, particularly in identifying the "Same" class with high precision (82.94%), there is still room for improvement in handling the "Different" class, as indicated by the lower precision (77.48%) and recall (77.48%). Future work will focus on optimizing the decision threshold to

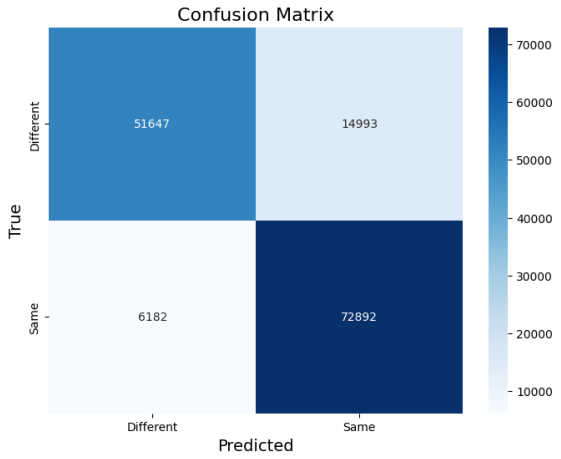

Fig. 8. Confusion Matrix Heatmap

further refine the model's ability to correctly classify both "Same" and "Different" instances.

Comparison of Speaker Verification Models:

The proposed Siamese network was benchmarked against leading speaker verification models using the LibriSpeech dataset, as shown in Table I. The Siamese network achieved an accuracy of 85.47% and an ROC-AUC score of 85.00%. Notably, it required only a single example (lasting a few seconds) for enrollment, showcasing its efficiency.

Table 1. Comparison of the proposed model with state-of-the-art speaker verification systems evaluated on the LibriSpeech dataset.

Model	**Enrollment Data Required**	**ROC-AUC (%)**	**Accuracy (%)**
i-Vector + PLDA	5–10 minutes per speaker	80.00	82.00
Deep CNN + Attention	2–5 minutes per speaker	86.50	88.00
Unsupervised Speech Representations	1–2 minutes per speaker	87.00	89.50
Proposed Siamese Network	**1 example (few seconds)**	**85.17**	**85.47**

In contrast, the i-Vector + PLDA model, which achieved an accuracy of 82.00% and an ROC-AUC of 80.00%, required significantly more enrollment

data, ranging from 5 to 10 minutes per speaker. Similarly, the Deep CNN with Attention model demonstrated a higher accuracy of 88.00% and an ROC-AUC score of 86.50%, but required 2 to 5 minutes of enrollment data per speaker. The Unsupervised Speech Representations approach performed the best in terms of accuracy, reaching 89.50% with an ROC-AUC of 87.00%, but still required 1 to 2 minutes of speaker data for enrollment.

This comparison highlights the efficiency of the Siamese network in minimizing the need for extensive enrollment data while maintaining competitive accuracy levels in speaker verification tasks on the LibriSpeech dataset, all with minimal computational power.

5 Conclusion

This work introduces a custom-designed Siamese network tailored for one-shot learning in voice authentication. By leveraging discriminative embedding learning, the proposed system effectively verifies speakers even with limited enrollment data. Through structured learning objectives and carefully applied data augmentation, the approach demonstrates resilience to varied acoustic conditions, making it a promising solution for voice authentication tasks.

However, the current system's performance could be further refined. Challenges such as extreme noise interference and complex distortions remain areas for improvement. Additionally, while computational efficiency is a notable strength, utilizing advanced processing capabilities and distributed computational frameworks could significantly enhance the system's scalability and robustness.

Future efforts can focus on integrating broader datasets to cover diverse acoustic and demographic contexts, as well as exploring multi-modal biometric solutions for enhanced security and adaptability. This study underscores the viability of a minimal-enrollment voice authentication framework and lays the groundwork for future innovations in secure and scalable biometric systems.

Declarations

- **Funding:** This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
- **Conflict of interest/Competing interests:** The authors declare that they have no conflict of interest.
- **Ethics approval:** Not Applicable
- **Consent to participate:** Not Applicable
- **Consent for publication:** Not Applicable
- **Availability of data and materials:** The datasets discussed in the manuscript are publicly available for research purposes.
- **Research Involving Human and/or Animals:** Not Applicable
- **Informed Consent:** Not Applicable

– **Authors' contributions:** Conceptualization: Shreyas Nagoor, Garima Pandey, Shashidhar G. Koolagudi; Methodology: Shreyas Nagoor, Garima Pandey; Formal analysis and investigation: Shreyas Nagoor, Garima Pandey; Writing - original draft preparation: Shreyas Nagoor, Garima Pandey; Writing - review and editing: Shreyas Nagoor, Garima Pandey, Shashidhar G. Koolagudi; Supervision: Shashidhar G. Koolagudi

References

1. Davis, S. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* **28**, 357–366 (1980).
2. Kumar, G. S., Raju, K. P., CPVNJ, M. R. & Satheesh, P. Speaker recognition using gmm. *International Journal of Engineering Science and Technology* **2**, 2428–2436 (2010).
3. Jalil, A. M., Hasan, F. S. & Alabbasi, H. A. *Speaker identification using convolutional neural network for clean and noisy speech samples*, 57–62 (IEEE, 2019).
4. El-Moneim, S. A. *et al.* Text-independent speaker recognition using lstm-rnn and speech enhancement. *Multimedia Tools and Applications* **79**, 24013–24028 (2020).
5. Bai, Z. & Zhang, X.-L. Speaker recognition based on deep learning: An overview. *Neural Networks* **140**, 65–99 (2021).
6. Panayotov, V., Chen, G., Povey, D. & Khudanpur, S. *Librispeech: an asr corpus based on public domain audio books*, 5206–5210 (IEEE, 2015).
7. Nagrani, A., Chung, J. S. & Zisserman, A. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).
8. Specified, N. Speaker identification using mfcc feature extraction and ann classification. *Springer* (2024). URL `https://link.springer.com/article/10.1007/s11277-024-11282-1`.
9. Liu, X., Sahidullah, M. & Kinnunen, T. Learnable mfccs for speaker verification. *arXiv preprint arXiv:2102.10322* (2021). URL `https://arxiv.org/abs/2102.10322`.
10. Zhou, Z., Chen, J., Wang, N., Li, L. & Wang, D. Adversarial data augmentation for robust speaker verification. *arXiv* (2024). URL `https://arxiv.org/abs/2402.02699`.
11. Hajavi, A. & Etemad, A. *Siamese capsule network for end-to-end speaker recognition in the wild*, 7203–7207 (IEEE, 2021).
12. Chen, Z., Wang, H., Yeh, C.-H. & Liu, X. Classify respiratory abnormality in lung sounds using stft and a fine-tuned resnet18 network (2022). URL `https://arxiv.org/abs/2208.13943`. `2208.13943`.
13. Erichson, N. B., Lim, S. H. & Mahoney, M. W. Gated recurrent neural networks with weighted time-delay feedback (2022). URL `https://arxiv.org/abs/2212.00228`. `2212.00228`.
14. Iqbal, M., Iqbal, M. & Iqbal, M. Resnet-18 with attention mechanism-bidirectional lstm hybrid approach for music genre classification. *International Journal of Computer Science* **6**, 45–56 (2025). URL `https://ijcs.net/ijcs/index.php/ijcs/article/view/4456`.
15. Seo, S. & Kim, J.-H. Self-attentive multi-layer aggregation with feature recalibration and normalization for end-to-end speaker verification system. *arXiv preprint arXiv:2007.13350* (2020). URL `https://arxiv.org/abs/2007.13350`.

16. Zhang, C. & Koishida, K. *End-to-end text-independent speaker verification with triplet loss on short utterances*, 1635–1639 (ISCA, 2017).
17. Li, C. *et al.* Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304* (2017). URL `https://arxiv.org/abs/1705.02304`.