Gitesh Kambli\textsuperscript{1}, Jay Oza \textsuperscript{1}, and Amit Maity\textsuperscript{1}

\textsuperscript{1}K. J, Somaiya Institute of Technology

December 14, 2023
MelSpectroNet: Enhancing Voice Authentication Security with AI-based Siamese Model and Noise Reduction for Seamless User Experience

Gitesh Kambli¹, Jay Oza¹, and Amit Maity¹

¹ K. J. Somaiya Institute of Technology, Mumbai, India

Abstract. Voice authentication has become critical for secure access control while achieving usability. Background noise and increased security requirements, however, continue to be problems. This paper presents MelSpectroNet, an innovative voice authentication system using Siamese neural network trained on over one million samples. It leverages mel spectrograms for efficient feature extraction and employs noise reduction, enhancing reliability. The model achieves 96.62% test accuracy, demonstrating efficacy. Our methodology involves audio denoising, meticulous spectrogram preprocessing, a tailored Siamese architecture, and rigorous training. Testing demonstrates MelSpectroNet’s exceptional performance and ability to generalize. However, enhancing longitudinal accuracy by accounting for natural voice variations over time still needs exploration. Overall, MelSpectroNet pioneers highly accurate and usable voice authentication with enhanced security. It balances user convenience and stringent authentication needs. This research motivates further work to optimize these systems for diverse conditions while advancing security and inclusiveness.

Keywords: voice-authentication · siamese-neural-network · mel-spectrogram · deep-learning.

1 Introduction

Voice authentication has become an essential component of current security systems, expertly balancing user convenience with the pressing requirement for strict security standards. However, it faces chronic and complex obstacles, emphasizing the importance of continual innovation in this domain. One of the most significant issues is the persistent interference of background noise, which is a common component of real-world acoustic situations. The presence of background noise not only complicates but also jeopardizes the authentication procedure. Additionally, the bar for voice authentication has been raised due to the increased security requirements in today’s digital environment, demanding systems that can successfully thwart malicious efforts to trick the system. These multiple issues necessitate not simply incremental advances, but a paradigm shift in voice authentication technologies. In response to these imperatives, we propose MelSpectroNet, a ground-breaking voice authentication system powered by
cutting-edge AI capabilities, designed not just to meet these difficulties but also to set new security and user experience standards.

MelSpectroNet signifies a fundamental leap in voice authentication. Its core design is based on a Siamese neural network model that was rigorously trained on a large dataset including over one million speech samples. MelSpectroNet deviates from traditional voice identification systems by leveraging the intricacies of mel spectrograms for quicker feature extraction, all while expertly managing the challenges of real-world acoustic situations via pioneering noise reduction algorithms.

This section of the paper serves as a thorough introduction to our painstaking investigation of the development process of MelSpectroNet. It gives a fundamental understanding of the guiding ideas and elements that form the basis of this revolutionary technology. Background noise mitigation, a crucial aspect of voice verification, is where our trip begins. We illustrate the importance of the first stage, which leverages a pre-trained TensorFlow model for audio denoising, ensuring that background noise has no bearing on the correctness of the subsequent voice authentication procedure.

The Siamese neural network architecture, which is the cornerstone of MelSpectroNet’s capabilities, forms the basis of our research. This architectural framework was specifically designed for tasks involving similarity estimation and feature extraction. It stands out for deploying two identical subnetworks. The Siamese network, which was developed specifically for voice authentication, is excellent at determining whether two voice samples come from the same speaker or different speakers. It extracts important features from voice spectrograms by utilizing the power of convolutional neural networks (CNNs), and it then computes a similarity score based on the embeddings of the two input samples.

In our methodology, data preprocessing, which plays a crucial role in voice authentication, is given careful consideration. The foundation of our data source is the VoxCeleb dataset, a comprehensive collection of voice data encompassing a wide range of speakers [9]. Raw audio data is transformed into mel-spectrograms, which capture the spectral nuances of audio signals and highlight characteristics crucial for voice analysis. The seamless conversion of audio data into a format suitable for neural network training is made possible by careful parameter configuration during spectrogram generation. All spectrograms harmoniously follow a standardized size in the interest of uniformity and computational effectiveness.

A specialized test set is used to carefully evaluate the Voice Authentication system’s effectiveness, emphasizing its ability to recognize distinct voices. However, we continue to be aware of the difficulties that speech authentication systems face, particularly the normal changes in a person’s voice over time. This acknowledgment highlights how dynamic this subject is, demanding a constant commitment to study and innovation targeted at improving the accuracy and dependability of voice authentication, especially in circumstances where there are temporal changes in an individual’s speech.

As we dive into the various aspects of our methodology in the upcoming sections, we’ll uncover the intricate inner workings of MelSpectroNet. This will
showcase how MelSpectroNet leads the way in voice authentication technology, ready to boost security and enhance user authentication experiences.

2 Related Work

Voice authentication has emerged as a critical aspect of modern security and access control systems, driven by the ever-expanding digital landscape. Researchers and developers have been deeply engaged in exploring the potential of voice authentication to enhance security and improve the user authentication experience. This section delves into the relevant literature, highlighting the wide range of contributions from other authors and studies which reinforce the findings in this paper.

Multimodal biometric authentication, as demonstrated by Zhang et al. [16], represents a promising area of research. Their study centers on creating an effective Android-based multimodal biometric authentication system through the integration of both facial and vocal recognition modalities, with the overarching goal being enhanced security compared to unimodal biometric authentication methods. This system is intended for smart terminals including Android smartphones. The voice recognition component incorporates voice content recognition, voice denoising via Discrete Wavelet Transform (DWT), and feature extraction of voices based on Mel-frequency cepstral coefficients (MFCC). The authors utilize Gaussian Mixture Models (GMM) for voice matching. The research also introduces an adaptive fusion strategy that dynamically adjusts the weighting of face and voice biometrics based on the signal-to-noise ratio (SNR) of voice data during the fusion process.

Another vital facet of voice authentication is liveness detection, a challenge addressed by Jiang et al. [6]. The paper addresses the growing concern of spoofing attacks in voice biometric systems, where malicious actors can impersonate users or use pre-recorded samples to deceive the authentication process. The proposed system, VoicePop+, focuses on leveraging the unique characteristics of the pop noise generated by a user’s oral airflow when speaking a passphrase close to the microphone. The research emphasizes the individual diversity in phoneme-pop sequences to adapt to different users’ unique vocal systems and utterance styles. To verify user legitimacy, VoicePop+ conducts consistency analysis based on pressure signals and phoneme-pop sequences and employs a similarity comparison approach based on unique phoneme-pop sequences collected during user enrollment.

The research paper entitled "VOLER: Leakage Resilient User Authentication Based on Personal Voice Challenges" [15] presents a novel approach to tackle the security and privacy concerns related to voiceprint authentication. Zhang et al. [15] recognize the growing popularity of Voiceprint Authentication as a Service (VAaS) and its associated convenience. However, they emphasize the significant concern of voiceprint leakage, whether it occurs over the air or in cloud storage, as it poses a threat to user voice privacy. Given these difficulties, the authors put forth VOLERE (VOice LEakage REsilient), which implements
a novel voiceprint synthesis method based on a Log Magnitude Approximate (LMA) vocal tract model. Their proposed technique involves consolidating the unique vocal attributes of individuals across diverse speaking modalities to generate a synthesized voiceprint that can be leveraged for authentication.

The research by López et al. [8] concentrates on developing and evaluating a Continuous Authentication (CA) system intended for Industry 4.0 environments, employing supervised machine learning methods. The main aim of this study is to tackle the difficulties related to the continuous and secure authentication of workers in industrial environments characterized by frequent interactions between workers and devices. The authors present a proposed computer-aided system that incorporates multiple sources of biometric data, including device sensors, application statistics, and voice recordings. This data is collected from operators’ mobile devices, such as smartphones and tablets. The results obtained from this research highlight the inherent capabilities of supervised machine learning algorithms, such as support vector machines and random forests, to improve the accuracy and effectiveness of biometric authentication systems. In addition, voice is often emphasized as a strong authentication factor because it shows low false acceptance rates even in cases where the workforce doubles.

Feature engineering and uniqueness analysis have also been explored by Tandogan et al. [12], who delve into techniques for estimating the uniqueness of human voice in the context of speaker verification systems, particularly using the i-vector model for feature representation. The authors propose a novel method to measure the biometric information content of i-vector-based representations of human voice. Their methodology leverages information theory constructs including entropy and mutual information to measure the distinctiveness provided by i-vectors for discriminating between different speakers. This entails discretizing the i-vector components, accounting for both inter-speaker and intra-speaker variability, thereby enabling a more precise quantification of the uniqueness of speaker traits represented within i-vectors.

The research by Clarkson et al. [4] focuses on speaker identification in security systems. They put forth a novel technique combining time-encoded signal processing and recognition (TESPAR) with probabilistic RAM (pRAM) neural networks. Their proposed method [4] uses TESPAR to condense speech data into 29-element vectors, forming unique representations for each speaker. The utilization of pRAM neural networks is subsequently employed for classification purposes, primarily due to their compatibility with the binary-coded signals utilized in TESPAR.

Noise suppression techniques, pioneered by S. Boll [2], address one of the practical challenges faced by voice authentication systems – environmental noise. The proposed method, known as spectral subtraction, offers an efficient way to analyze and process digital speech signals to reduce the impact of noise. The algorithm estimates the noise spectrum during periods of non-speech activity and subtracts this estimated noise spectrum from the noisy speech signal. Various techniques such as half-wave rectification, residual noise reduction, and signal
attenuation during non-speech activity are applied to further improve noise reduction and speech quality.

Kakuba et al.’s work [7], which developed the attention-based multi-learning model (ABMD) for Speech Emotion Recognition (SER) with an emphasis on improving voice authentication systems, adds another level of complexity and usefulness to voice authentication. The ABMD model, as proposed [7], utilizes residual dilated causal convolution (RDCC) blocks and dilated convolution (DC) layers in conjunction with multi-head attention. The utilization of this architectural design allows the model to acquire knowledge of extended dependencies among speech features concurrently, resulting in a significant reduction in parameter count when compared to conventional recurrent neural networks (RNNs). In addition, this approach considers spatial cues, enhancing their resilience and efficiency in applications related to speech-emotion recognition (SER).

Gaussian Mixture Models (GMM) [11], which extract features from audio data using Linear Predictive Coding (LPC), are among the many modeling techniques that have been studied to increase the security and dependability of voice authentication systems. Hybrid approaches such as the combination of Mel Frequency Cepstral Coefficients and Bayesian Gaussian Mixture Models (HMFCC-BGMM) [13] have also been explored. Additionally, Hidden Markov Models (HMM) [5] have been utilized in this context. The HMFCC-BGMM model [13] integrates voice features derived from MFCC with Bayesian Gaussian Mixture Models, while the HMM-based voice authentication system [5] effectively stores voiceprints for individuals and demonstrates text-independent speaker recognition techniques, emphasizing voice traits over specific speech content. Collectively, these studies showcase the potential of various modeling techniques in enhancing voice authentication through their efficacy and adaptability in ensuring secure authentication processes.

Boles et al.’s study [1] looks at speaker identification systems’ relevance in the modern world, particularly in light of the growing use of voice-activated gadgets. The main goal is to create a speech identification system that is not dependent on text, which will be achieved by using support vector machines and extracting Mel-Frequency Cepstral Coefficients.

Siamese neural networks, introduced by Bromley et al. [3], have shown promise in similarity-based authentication tasks. Although not directly related to voice authentication, this research highlights the potential of Siamese neural networks in authentication, which could find relevance in voice authentication research.

Xie et al. [14] propose using a Siamese neural network architecture with wav2vec features for anti-spoofing in automatic speaker verification systems. They note that while traditional countermeasures can achieve high accuracy on seen data, performance degrades against unseen spoofing attacks. Thus, they argue that improving generalization is critical. Their approach has two phases - representation learning using a Siamese network contrastive loss, and classification using the learned embeddings. Different CNN backbones are evaluated with the wav2vec speech embeddings to improve generalization over traditional features like MFCCs. On the ASVSpoof 2019 dataset, their model reduces equal
error rate from 4.07% to 1.15%, outperforming prior work. Visualizations confirm the Siamese network learns more discriminative embeddings. Overall, the paper demonstrates that pre-training and metric learning with Siamese networks is an effective strategy for anti-spoofing, advancing state-of-the-art performance.

Lastly, Nanni et al.’s [10] exploration of spectrogram classification using dissimilarity space, while not specific to voice authentication, presents a broader perspective. It emphasizes the application of dissimilarity space in audio data classification, potentially contributing to the evolution of voice biometric systems.

While voice authentication continues to advance, current systems still face several challenges that need to be addressed. A major drawback is vulnerability to spoofing attacks using synthesized or replayed voice samples, as highlighted by Jiang et al. [6] and Zhang et al. [16]. Improving anti-spoofing and liveness detection remains an active area of research. Environmental noise and channel effects can also degrade authentication accuracy, requiring robust noise suppression techniques like spectral subtraction [2]. There is also difficulty in handling natural voice changes over time [15], which degrades accuracy as an individual’s voice ages. The uniqueness of voice traits between individuals, though reasonably high, is not perfect, so errors due to inter-speaker similarities may persist [12]. Most voice authentication systems also require users to speak fixed passphrases [8], which has usability drawbacks. Storage and potential leakage of voiceprints raise privacy concerns as well, motivating leakage-resilient designs like VOLERE [15]. Platform dependence on specific hardware or datasets limits generalizability, as does a lack of testing across diverse populations. Resource constraints pose challenges for on-device voice authentication on low-power IoT devices [1]. Continual authentication in dynamic industrial settings remains challenging [8]. Taken together, these issues emphasize that despite significant progress, there remains much room for improvement in making voice authentication systems more robust, usable, scalable, and privacy-preserving. Advancing voice modeling, anti-spoofing defenses, lightweight architectures, and multi-modal authentication will be vital in overcoming the limitations of current technologies.

3 Methodology

The methodology employed in this study revolves around the development of a voice authentication system using a Siamese neural network model, firmly rooted in robust theoretical foundations. Voice authentication, as a facet of biometric authentication, capitalizes on the unique vocal characteristics of individuals to ascertain their identities. This section provides a clear explanation of the fundamental theoretical principles that serve as the foundation for this research effort.

3.1 De-noising Audio

One of the primary enhancements in our methodology is the initial step, where we employ a pre-trained TensorFlow model for audio de-noising. This step ad-
addresses a critical real-world challenge—background noise reduction—before voice authentication, significantly bolstering the system’s overall reliability. This integration ensures that background noise does not compromise the accuracy of the subsequent voice authentication process, enhancing the overall effectiveness of MelSpectroNet in diverse acoustic environments.

3.2 Siamese Neural Network

Central to this research is the Siamese neural network architecture—a deep learning model purpose-built for tasks involving similarity measurement and feature extraction. Siamese networks are distinguished by their deployment of two identical subnetworks, hence the moniker "Siamese," which jointly processes pairs of input data, generating embeddings that enable the measurement of similarity or dissimilarity[3].

In the domain of voice authentication, the Siamese network takes on the task of discerning whether two voice samples originate from the same speaker or different speakers. This architecture harnesses the potency of convolutional neural networks (CNNs) to extract salient features from voice spectrograms, subsequently calculating a similarity score based on the embeddings of the two input samples.

3.3 Data Preprocessing

Preprocessing the data is a crucial step in voice authentication. The VoxCeleb dataset, a large collection of voice data with over 7000 speakers, is used as the main data source. VoxCeleb contains short segments of human speech extracted from interview videos uploaded to YouTube, with over 1 million utterances totaling over 2000 hours of audio [9]. This raw audio data is transformed into mel spectrogram representations, which highlight key voice features by depicting the spectrum of the audio signal. Mel spectrograms emphasize components critical for analyzing and distinguishing different voices.

Fig. 1. The figure represents the Mel Spectrogram of a voice sample

Fig. 1. illustrates how the mel spectrogram representation comprehensively captures the acoustic characteristics of a voice sample, providing a fine-grained
and exhaustive visualization of the audio signal. A mel spectrogram refines this representation by adopting a mel-scale, a perceptually appropriate scale based on the properties of the human auditory system, as opposed to a normal spectrogram, which describes audio signals in terms of their frequency content across time. This scale replicates how people perceive sound, emphasizing lower frequencies that are perceptually relevant while demonstrating finer clarity in higher frequencies. This method is consistent with the non-linear relationship between perceived pitch and actual frequency. As a result, a mel spectrogram provides a highly granular picture of the speech sample, emphasizing crucial aspects such as phonemes, intonation, and timbre, all of which are important in voice identification. As a result, not only does the mel spectrogram provide a richer and more human-centric representation of the speech sample, but it also enables greater feature extraction, which is critical for proper analysis and verification in voice-based systems.

Careful parameter configuration during spectrogram generation—including specifications for Fast Fourier Transform (FFT) window, hop length, and mel filterbank settings—ensures the translation of audio data into a format amenable to neural network training. Ensuring uniformity and enabling efficient processing, all spectrograms are standardized to a fixed size of 80 x 450 pixels and are converted into NumPy arrays.

### 3.4 Siamese Network Architecture

The Siamese neural network’s architecture shown in Fig. 2. plays a key role in this research, with multiple layers each assigned a specific function in the voice authentication process:

- **Convolutional Layers**: These initial layers shoulder the responsibility of feature extraction from input mel-spectrograms. Convolutional filters systematically scan the spectrograms, detecting pertinent patterns and features instrumental in distinguishing individual speakers.
- **Max Pooling**: Subsequent to feature extraction, max pooling layers take charge of dimensionality reduction, preserving the most salient information
while enhancing computational efficiency and mitigating the risk of overfitting.

- **Flattening Layer**: Following feature extraction and dimensionality reduction, embeddings undergo further processing via a flattening layer, prepping the data for subsequent dense layers.

- **Dense Layer**: The dense layer extends the transformation of embeddings, empowering the network to uncover intricate relationships within the data.

- **L1 Distance Layer**: This layer computes the L1 (Manhattan) distance between embeddings produced by the two subnetworks. The distance calculation serves as a surrogate for similarity, with smaller distances signifying greater similarity.

- **Sigmoid Output Layer**: The conclusive classification step employs a sigmoid output layer, endowing input voice samples with a similarity score. This score discerns whether the samples originate from the same speaker (yielding high similarity) or different speakers (resulting in low similarity).

### 3.5 Training Strategy

The training phase assumes a pivotal role in enabling the Siamese network to acquire the discriminatory capacity required for distinguishing between voice samples. This critical endeavor is driven by the utilization of the binary cross-entropy loss function, which exerts pressure on the network to minimize the similarity scores for dissimilar speakers while simultaneously maximizing them for identical speakers.

To optimize the model’s parameters, we employ the Adam optimizer, judiciously selecting a learning rate of 1e-6 to ensure efficient convergence. The training process unfolds over a single epoch, during which the model meticulously processes data in batches, each comprising 128 samples. It is worth noting that the model’s training is further fortified through the repetition of this process for a total of 10 epochs, resulting in refined performance and heightened discriminatory capabilities.

### 3.6 Evaluation

The Voice Authentication system’s performance is thoroughly assessed using a specific test set. During this evaluation, we extract embeddings from the voice samples and employ a classifier to determine whether these samples originate from the same speaker or different ones. The high accuracy rates achieved by the machine learning model employed in this study highlight its effectiveness in recognizing individual voices.

However, it’s important to acknowledge that there are still some accuracy challenges that need attention to make voice authentication more dependable. One significant challenge is the natural changes that can occur in a person’s voice over time. Therefore, further research is necessary to enhance the accuracy of these systems and make them more reliable, especially when dealing with variations in an individual’s voice over time.
4 Result & Discussion

In this study, we present a comprehensive methodology for MelSpectroNet, an innovative voice authentication system driven by advanced AI technology. This research aimed to develop a voice authentication system using a Siamese neural network that can accurately verify if two voice samples are from the same speaker. Addressing the critical need for secure yet user-friendly authentication methods, MelSpectroNet leverages a Siamese neural network architecture, meticulously trained on a vast dataset of over one million voice samples, to accurately confirm user identities. The system tackles real-world challenges such as background noise by incorporating noise reduction techniques. Rigorous testing demonstrates MelSpectroNet’s exceptional efficacy, achieving an impressive accuracy rate.

Table 1. Table shows the Training and Testing Accuracy.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training accuracy</td>
<td>95.48</td>
</tr>
<tr>
<td>Testing accuracy</td>
<td>96.62</td>
</tr>
</tbody>
</table>

As shown in table 1, the results demonstrate that the proposed model achieves an exceptionally high accuracy of 96.62% on the test set, depicting a small increase from the 95.48% training accuracy. This indicates that the model can effectively learn to discriminate between the same and different speakers while generalizing well to new data.

Table 2. Table shows the Classification Metrics and Values.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.9884</td>
</tr>
<tr>
<td>Recall</td>
<td>0.9422</td>
</tr>
<tr>
<td>F1-Score</td>
<td>0.9647</td>
</tr>
</tbody>
</table>

The high precision, recall, and F1-score in table 2, demonstrates reliable performance across different data slices. The balanced accuracy on positive and negative pairs proves the model’s ability to verify the same speakers while rejecting different speakers.

5 Conclusion

The MelSpectroNet speech authentication system, which utilizes artificial intelligence to provide secure and user-friendly authentication, was trained on an
extensive dataset of over one million voice samples in order to generate reliable speech representations. This research has also implemented noise reduction methods to improve the system’s reliability in noisy real-world situations. Our hypothesis was that using mel spectrograms would allow efficient extraction of identifying voice features compared to raw audio or other representations. Our research showed remarkable performance with a training accuracy of 95.48% and a test accuracy of 96.62%. Furthermore, we hypothesized that using a Siamese neural network tailored for similarity learning would be more suitable for voice authentication compared to other architectures. The slight difference in accuracy between training and test data shows how well MelSpectroNet generalizes to new data, showcasing the capability for robust speaker verification. The high test accuracy validates this hypothesis, indicating that mel spectrograms do enable the model to reliably encode voice characteristics necessary for authentication.

Overall, the exceptional results strongly validate our hypotheses regarding the effectiveness of mel spectrograms and Siamese networks for voice authentication. Voice authentication, as demonstrated by MelSpectroNet, effectively balances the demands of security and user convenience, showcasing the potential of voice authentication in modern digital security. The high accuracy surpasses state-of-the-art benchmarks, underscoring the significance of this work. Users can expect a seamless and dependable authentication experience, while organizations benefit from heightened security measures. However, challenges related to the variability of an individual’s voice over time remain, and further research is necessary to enhance the accuracy and reliability of voice authentication systems in the face of such challenges. MelSpectroNet represents a significant advancement in voice authentication technology by providing heightened security while maintaining a smooth user experience. This research serves as an important step forward for voice-based verification systems, exhibiting their potential for real-world deployment through further refinements.

6 Future Scope

While this work demonstrates promising results for voice authentication using Siamese neural networks, there remains significant scope for advancing the state-of-the-art through future improvements. Enhancing longitudinal accuracy by handling voice variations over time remains an open research question. The model’s performance over diverse real-world conditions also needs further analysis before large-scale deployment. Novel neural architectures that better model speech modulations like RNNs, transformers, and graph neural networks should be evaluated as replacements for current CNNs. End-to-end architectures could also be explored for joint optimization of feature extraction and classification. Testing the developed systems under practical conditions considering varying background noises, recording devices, and environments will provide insights into their robustness for deployment. Multi-modal fusion of voice with face biometrics could also be examined for more secure user verification. For real-time applications, compressing models using knowledge distillation and quantization
techniques needs investigation. Edge computing methods like federated learning must be leveraged to enable on-device training without compromising privacy. Adversarial attacks pose a serious threat, hence developing attack-resistant models using techniques like adversarial training should be prioritized. The systems could also be strengthened by detecting voice spoofing attempts. Analyzing the influence of gender, age, accent, and language on the systems’ accuracy can uncover underlying biases and suggest ways to make them more equitable. The evolution of authentication performance over longitudinal voice data is another promising research direction. Incorporating voice authentication along with speech and speaker recognition systems can provide intuitive multi-factor security for voice interfaces. With substantial future work, voice-based user verification could become ubiquitous across applications demanding both security and simplicity.

References