Two-Factor Audio Verification System Using Biometric Watermarking and GAN-Based Watermarking

Abhinandan Sudharsan
CSE(AI&ML)
PES University
Bangalore, Karnataka, India
abhinandan.sudharsan29@gmail.com

Abhinav Bhargava
CSE(AI&ML)
PES University
Bangalore, Karnataka, India
abhinav4305@gmail.com

Anjali H Ramurs
CSE(AI&ML)
PES University
Bangalore, Karnataka, India
anjalihr05@gmail.com

Anurag Senapati
CSE(AI&ML)
PES University
Bangalore, Karnataka, India
anurags7475@gmail.com

***Abstract*—Background: The increasing sophistication of audio deepfakes and tampering techniques has made verifying the authenticity and integrity of voice recordings a significant challenge. Existing solutions often fail to provide robust assurance of an audio file's origin and unaltered content.**

Methods: We present a two-factor audio verification system that combines speaker recognition with hidden watermarking. The system extracts a unique voiceprint for each speaker and converts it into a small digital signature. This signature is then invisibly embedded into the audio using a neural network-based watermarking technique.

Results: Extensive testing on the LibriSpeech dataset demonstrates that our system achieves bit error rates reduced to zero after fine-tuning. The voiceprint-based watermark is consistently and accurately recovered, and the system reliably detects tampering and flags suspicious audio in practical scenarios.

Conclusions: By locking both who is speaking and where the audio came from, this approach strengthens audio authentication. These results affirm the system's reliability and scalability, and we plan to expand this work with more testing and real-world use cases to improve reliability.

Keywords—biometric watermarking, audio steganography, voiceprint authentication, audio forensics, generative adversarial networks

I. Background

Advances in deep learning have made it increasingly easy to create convincing synthetic speech, known as audio deepfakes. These fabricated audio recordings pose serious challenges by mimicking real voices so well that distinguishing authentic from manipulated speech becomes difficult. This threat undermines the reliability of audio in critical applications such as legal evidence, news media, and secure communications. Existing audio authentication methods often rely on watermarking or signal analysis alone, which are inadequate against sophisticated manipulations like voice conversion and seamless audio splicing.

This work aims to provide a reliable verification solution that leverages both biometric speaker recognition and advanced watermarking with neural networks. Biometric methods use unique features extracted from an individual's voice—called voiceprints—to verify identity. These voiceprints are generated by deep learning models trained on large datasets and can distinguish between different speakers with high accuracy. Despite this, biometric verification alone does not guarantee that the audio file has not been altered or fabricated, highlighting the need for additional security measures.

On the other hand, audio watermarking has evolved from basic data embedding to sophisticated, imperceptible techniques enabled by Generative Adversarial Networks (GANs). These GAN-based watermarking approaches invisibly embed data into audio signals, making tampering or unauthorized duplication detectable by extracting and analyzing these watermarks. However, watermarking alone does not confirm that the voice in the audio belongs to the authorized speaker, creating a gap in verification.

Our project addresses this gap by combining voice biometric hashes with GAN-based watermarking to create a two-factor verification system. This integration enhances the ability to confirm both the audio's source device and the speaker's identity. By embedding a cryptographic hash of the speaker's unique voiceprint into the audio, and then verifying this during playback, the system provides a robust approach to detect manipulated audio, deepfakes, and forged content. This hybrid approach aims to advance secure audio verification for real-world applications.

II. Methods

A. Model Overview:

The proposed system consists of two primary components: a biometric "Speaker Encoder" and an adversarial "Watermark GAN" framework. The Speaker Encoder is a convolutional neural network that extracts unique voice embeddings or voiceprints from speech audio. These embeddings capture speaker-specific characteristics and are hashed into a binary message. The Watermark GAN includes three 1D convolutional networks — the Generator, which embeds the secret hash invisibly into the audio; the Extractor, which recovers the embedded message; and the Discriminator, which discriminates between clean and watermarked audio to enforce watermark imperceptibility. Together, these modules form a two-factor authentication system where both the speaker's identity and the audio's origin can be simultaneously verified.This end-to-end pipeline has been empirically validated, with bit error rates reduced to zero after fine-tuning.

B. Dataset Collection and Preprocessing:

Training the Speaker Encoder leverages the LibriSpeech dataset, particularly the "train-clean-100" subset, containing thousands of audio recordings from diverse speakers. Audio files are converted to mel-spectrogram representations to serve as model inputs, capturing frequency and temporal audio features. For the GAN models, the MUSAN noise dataset is incorporated to introduce realistic background noise, improving watermark robustness. Random one-second audio clips are extracted and normalized, making them suitable for adversarial training. Shorter clips are padded, while longer ones are truncated to a fixed length.

C. Training Procedures:

The Speaker Encoder is trained under a supervised classification regime, optimizing cross-entropy loss to enforce speaker discrimination through sufficiently distinct embeddings. Training utilizes the Adam optimizer and incorporates strategies such as batch normalization and adaptive pooling for stability and input length flexibility. The Watermark GAN training optimizes three loss components: message recovery accuracy via mean squared error, audio fidelity via L1 loss comparing original and watermarked audio, and adversarial loss via binary cross-entropy with the Discriminator's predictions. Training proceeds iteratively over multiple epochs with alternating updates to the Generator-Extractor pair and the Discriminator. Model selection was guided by convergence criteria prioritizing minimal bit error rates and robust verification accuracy.

D. Watermark Embedding and Enrollment:

In enrollment, audio is passed through the Speaker Encoder to obtain a voiceprint embedding, from which a 64-bit secret hash is derived by thresholding components. This hash seed is then embedded into the original audio waveform by the Generator, producing a watermarked "sealed" file. This file carries both the biometric hash and the origin authenticity embedded imperceptibly, ready for transmission or storage. The system's efficacy in embedding and recovering these watermarks with zero bit errors was confirmed through extensive fine-tuning and testing.

E. Verification and Message Extraction:

Verification begins with the Discriminator evaluating the given audio for watermark presence, determining whether the file is authentic or a potential forgery. If watermarked, the Extractor recovers the embedded hash message. The audio is concurrently passed through the Speaker Encoder to produce a new voiceprint hash. These two hashes are compared bitwise, and their agreement within a small error threshold constitutes successful verification. Divergence above this threshold signals potential tampering, such as deepfake synthesis or voice alteration.

Divergence above this threshold signals potential tampering, such as deepfake synthesis or voice alteration.

F. System Integration and Workflow:

The entire system is integrated into an end-to-end pipeline orchestrating enrollment and verification. The design streamlines audio input processing, feature extraction, message embedding/recovery, and final authentication decision-making. The enrollment stage securely binds the speaker's biometric identity to the audio content, while the verification stage robustly confirms both identity and origin. Input and output formats are standardized for ease of use, and the system is designed for efficiency to support real-time or near-real-time applications. Performance profiling measured a throughput of approximately 295 samples per second, with an average batch processing time under 150 milliseconds, indicating the system's readiness for scalable deployment.

III. Results

The proposed two-factor audio verification system was evaluated on key aspects of biometric speaker recognition and watermark embedding robustness. Preliminary experiments with the Speaker Encoder model demonstrated the network's capacity to generate distinctive voice embeddings that enable clear separation between speakers. Using similarity scores derived from cosine distances, the system reliably discriminated between same-speaker and different-speaker audio samples drawn from the LibriSpeech test-clean dataset, indicating promising biometric verification performance.

The GAN-based watermarking framework showed effective embedding and recovery of 64-bit secret messages derived from voiceprint hashes. Under clean audio conditions, watermark extraction exhibited low bit error rates, reflecting the generator's ability to produce imperceptible and recoverable watermarks. To enhance resilience against environmental noise, the system incorporated MUSAN noise dataset augmentations during training, preparing the Extractor network for common audio distortions encountered in real-world scenarios.

The Integration of these components in the verification pipeline demonstrated their complementary strengths. The Discriminator accurately identified watermarked audio, enabling reliable extraction and comparison of hashes for authentication. Bit error thresholds were fine tuned to detect tampering, enabling detection of altered or deepfake audio with high confidence.

During training on the LibriSpeech corpus, the system's bit error rate (BER) was reduced from 20.31% at step 0 to 0.00% after 140 steps, reflecting great watermark recovery. Verification yielded a similarity score of 0.8303, surpassing the 0.7 threshold for authentication. Overall the throughput metrics demonstrated processing of approximately 295 samples per second with average batch times of 108 milliseconds.

While detailed quantitative metrics such as Equal Error Rates and False Acceptance/Reject Rates await further experimentation, these early results establish a robust proof-of-concept validating the system architecture. The combined biometric and watermarking factors provide a layered defense, increasing difficulty for adversaries to forge authenticated audio without detection. Future work will extend evaluation to diverse acoustic conditions and adversarial scenarios to reinforce these findings.

IV. Discussion

The integration of biometric speaker verification with GAN-based audio watermarking opens a new avenue for enhancing audio authentication. By combining voiceprint-based identity verification and origin watermark embedding, the system addresses the inherent weaknesses of single-factor methods, providing a layered security approach. This dual mechanism significantly raises the difficulty for adversaries attempting to produce audio forgeries that can simultaneously bypass biometric and origin verification.

The Speaker Encoder's ability to generate distinct and discriminative embeddings confirms the viability of biometric voiceprints as reliable personal identifiers. When combined with a GAN-driven watermarking scheme that securely embeds these identifiers into the audio, the system essentially creates a robust "biometric lock" on the audio content. This approach holds substantial promise for critical applications such as forensic audio analysis, secure communications, and copyright protection, where verifying the authenticity and source of audio is crucial.

However, practical deployment challenges remain. Environmental noise, audio compression, and transmission artifacts can affect watermark invisibility and voiceprint consistency. While the training process incorporates noise augmentation to mitigate these factors, comprehensive testing across more diverse acoustic conditions is necessary. Furthermore, optimizing the balance between watermark strength and audio quality is crucial to maintaining both security and usability.

The modular design of the system provides flexibility for future enhancements, including the integration of additional biometric modalities or more sophisticated discriminator architectures to improve tampering detection sensitivity. Real-time processing remains a key requirement for practical use, motivating efforts to optimize model size and inference speed without sacrificing accuracy.

In conclusion, this two-factor verification framework offers a robust foundation for trustworthy audio authentication, addressing current and emerging threats from synthetic audio generation. Ongoing work will focus on extended evaluation, enhanced robustness, and deployment strategies to fully realize its practical potential and support secure voice-based systems in real-world settings.

V. Conclusion

This paper introduces a new method for verifying audio recordings by combining biometric speaker recognition with GAN-based watermarking. It demonstrates how deep learning can be used not only to extract unique speaker voiceprints but also to securely embed and recover these identifiers as imperceptible watermarks within the audio itself. These advances narrow the gap between verifying the speaker's identity and confirming the audio's authenticity, addressing critical issues arising from the rise of sophisticated audio deepfakes and tampering.

This opens up important opportunities for enhancing trust in digital audio across sectors including security, forensics, media distribution, and communications. By linking biometric and cryptographic verification, the system offers a robust dual-layer defense against forgery, raising the bar for adversaries seeking to spoof or manipulate audio content.

Our empirical results validate the feasibility and effectiveness of this approach,with fine-tuned models exhibiting zero bit error rates in watermark recovery. By linking biometric and cryptographic verification, the system offers a robust dual-layer defense against forgery, substantially raising the bar for adversaries attempting to spoof or manipulate audio content.

This framework paves the way for more advanced multi-factor authentication schemes that may incorporate additional biometric modalities or context-aware detection to further enhance audio security. Optimizing the balance between watermark imperceptibility and robustness remains an area of ongoing work, along with expanding the system's scalability and deployment readiness.

Ultimately, this work lays a sound foundation for AI-assisted audio verification technologies that safeguard the integrity and authenticity of voice content amid increasingly complex digital threats. This approach holds transformative potential for maintaining trust in voice-based communication and protecting against emerging synthetic media, marking a significant step toward more secure and reliable audio systems for modern applications and evolving adversarial challenges.

Acknowledgments

We thank Dr. Bhaskarjyoti Das (PES University) for his valuable guidance throughout the course, which greatly influenced this work. We also express our deep gratitude to our project mentor Namita Achyuthan for her continuous support, insightful feedback, and dedicated assistance that significantly contributed to the progress and success of this project.

References

[1] Deepfake Detection in Call Recordings: A Deep Learning Solution for Voice Authentication

[2] Deepfake Audio Detection via MFCC Features Using Machine Learning

[3] Watermarking Techniques for Content Integrity Verification, Tamper Detection and Forensics in Synthetic Media