Voice Watermarking for Authentication and Copyright Protection Using Neural Models

Minh The Quang Tran
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
s3979562@rmit.edu.vn

Nhat Minh Nguyen
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
s3924871@rmit.edu.vn

Dat Man Nguyen
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
s3877932@rmit.edu.vn

Lee Jae Sung
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
s3977739@rmit.edu.vn

Huo-Chong Ling
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
huochong.ling@rmit.edu.vn

Linh Duc Tran
School of Science, Engineering and Technology
RMIT University
Ho Chi Minh city, Vietnam
linh.tranduc@rmit.edu.vn

***Abstract*—Voice information is now more vulnerable to misuse, ranging from traditional copyright theft to emerging threats in the era of artificial intelligence (AI), where voice cloning and deepfake synthesis can easily bypass conventional verification methods. The problem of voice verification and protection against copyright infringement can be addressed with this paper through the development of a neural watermarking model inspired by AudioSeal, an advanced deep learning watermarking system. We train and fine-tune AudioSeal for offline, pre-recorded speech, inserting watermarks that are both imperceptible as well as resilient against typical audio processing attacks. The system was developed for a proof-of-concept application to support end-to-end watermark embedding and detection within actual user workflows. Experimental outcomes demonstrate that our model attains high imperceptibility (PESQ ≈ 4.272, SI-SNR ≈ 39.335) with competitive robustness against compression, resampling, and noise attacks. These findings indicate the ability of watermarking based on deep learning to secure voice data for security-critical and copyright-related applications.**

Keywords—voice watermarking, audio authentication, robustness, imperceptibility, AI model

I. INTRODUCTION

The rapid development of digital audio data has boosted the demand for secure voice data protection and authentication. Digital watermarking as a whole has emerged as an effective tool for copyright protection, tampering detection, and security confirmation. However, voice data, as compared with text or image, comes with unique challenges in its temporal nature, perceptual sensitivity, and need for on-the-fly use. These issues are even more prevalent in the AI era as voice cloning and voice synthesis introduce huge risks of impersonation and unauthorized use.

Using synthetic audio in the wrong way, especially with deepfake technologies, has become a big security and trust issue. High-profile cases show how convincingly cloned voices can be used to commit fraud, impersonate someone else, and spread false information. For instance, in 2019, criminals used AI-generated voice to pretend to be a CEO and tricked a UK energy firm into sending about $243,000 to the bank account of a Hungarian supplier [1]. Deepfake audio clips have been used in politics and media to spread misinformation and cause false doubt and fear for the public. Recent surveys show that improvements in text-to-speech and voice cloning have made it possible to copy not only the words but also the tone, accent, and way of speaking of targeted people with high accuracy [2].

This paper contributes to addressing these challenges by designing, implementing, and evaluating a neural watermarking system specialized for voice audio. Specifically, we adapt AudioSeal, a deep learning watermarking model originally proposed for general audio, and retrain it on speech-focused datasets to improve its imperceptibility-robustness balance in voice-specific contexts.

The contributions of this paper are as follows:

1. Model adaptation: We retrain AudioSeal with speech-focused datasets to extend its capabilities to voice authentication.
2. Prototype application: We develop VoiceMark, a web-based system demonstrating practical deployment of watermark embedding and detection.
3. Evaluation: We compare our retrained model against classical watermarking methods, assessing imperceptibility, robustness, and detection accuracy under real-world audio edits.

The remainder of the paper is structured as follows: Section II reviews related work on classical and deep learning-based watermarking. Section III details the system design and

methodology, including dataset preparation, model training. Section IV describes the experimental setup and evaluation metrics. Section V presents results and discussion, while Section VI concludes with key findings and outlines directions for future research.

II. Related work

Audio watermarking techniques can be broadly divided into classical signal processing and deep learning-based approaches. Among the classical approaches, Singular Value Decomposition (SVD)-based watermarking [3] – [6] has been widely studied due to its robustness against noise and compression. In these methods, audio frames or transformed coefficients are decomposed into singular values and vectors, and watermark bits are embedded by modifying selected singular values. Because singular values represent intrinsic structure, such perturbations remain stable against many distortions. Early work by Özer et al. [3] applied SVD directly to the Short-Time Fourier Transform (STFT) spectrogram, embedding watermarks by adaptively modifying singular values, which proved robust to filtering and compression while preserving imperceptibility. More recent advances include SVD combined with frequency transforms, such as DCT-SVD [4] and DWT-SVD approaches [5], and adaptive frameworks that adjust embedding strength to maximize inaudibility while maintaining resilience [6]. Liu et al. [6] proposed a DWT-SVD scheme with differential embedding, which dynamically tunes parameters per frame, achieving strong resistance to noise and re-encoding.

The Short-Time Fourier Transform (STFT) is also a popular time-frequency domain for audio watermarking since it provides a joint time–frequency representation aligned with human auditory perception. By embedding bits into selected STFT coefficients, watermarking systems can exploit psychoacoustic masking while distributing redundancy across time frames. Jayarani et al. [7] proposed a zero-watermarking approach based on STFT, extracting signatures without altering the host audio. More recently, Liu et al. [8] introduced Timbre Watermarking, which embeds a secret bitstring in the STFT spectrogram of speech and survives state-of-the-art voice cloning pipelines. Gan et al. [9] extended this with SyncGuard, a deep learning watermark embedded across STFT frames to resist desynchronization attacks such as cropping and time-scaling.

Spread-spectrum watermarking is another traditional method that aims to make things more stable. This method uses pseudo-random sequences to spread a narrow-band watermark signal across wide frequency bins, which makes it resistant to filtering and compression [10]. The embedded signal is hard to get rid of because it spreads energy over a wide area, which would ruin the host audio. But this strength comes at the cost of payload capacity, since each bit needs to be embedded in many samples.

Invariant-feature methods build on classical methods by using signal features that don't change when they are distorted. SVD has been a popular choice because changing the singular values of audio frames usually keeps them strong against common signal processing [3] – [6]. Zhao et al. recently came up with the Frequency Singular Value Coefficient (FSVC), which encodes watermark bits in the ratio of singular values between two segments. FSVC is resistant to desynchronization attacks like time-scale modification or cropping, while earlier SVD-based schemes were not [11].

While these classical techniques achieve varying degrees of imperceptibility and robustness, they are often constrained by limited payload capacity, vulnerability to desynchronization, or high computational cost. To overcome these limitations, recent research has shifted toward a more modern approach: in the past five years, deep learning-based watermarking has emerged as the state of the art. Pavlović et al. [12] demonstrated a Deep Neural Network (DNN) embedder - detector achieving <1% Bit Error Rate (BER) on speech with high imperceptibility (Perceptual Evaluation of Speech Quality (PESQ) ≈ 4.33, Signal-to-Noise Ratio (SNR) > 38 dB). Singh et al. [13] proposed SilentCipher, which further improved imperceptibility by incorporating psychoacoustic masking and pseudo-differentiable compression layers. Timbre watermarking [8], presented at NDSS 2024, specifically targets voice cloning attacks, embedding repeated cues in the frequency domain that persist through cloning pipelines.

AudioSeal [14], presented at ICML 2024, represents a milestone, introducing a generator–detector architecture with perceptual masking and localized detection. It enables segment-level watermark identification and achieves detection speeds two orders of magnitude faster than prior neural methods [14]. However, AudioSeal's pretrained models are trained on general audio, and its robustness against adaptive Artificial Intelligence (AI)-driven manipulations (e.g., cloned voices) remains underexplored.

To address this gap, our project builds directly on AudioSeal, retraining and fine-tuning it for offline, prerecorded voice watermarking. By tailoring datasets and evaluations to voice-specific challenges - including cloned and accented speech - we extend AudioSeal's imperceptibility-robustness balance into new contexts. This positions our work as a step towards specialized, voice-focused watermarking that is inaudible and practical for real world deployment.

III. Methodology / system design

A. Model Methodology

AudioSeal's architecture [14] is based on the EnCodec [15] and the framework is a deep learning-based system built on a jointly trained generator and detector architecture. It is designed to be a proactive solution for AI-generated audio detection by embedding a watermark directly into the audio at the time of its creation.

1) Datasets and preprocessing

For training and evaluation, the English subset of the VoxPopuli corpus was used [16], which comprises 543 hours of speech at 16 kHz, and contributed by approximately 1,300 unique speakers. The dataset was divided into training, validation, and test splits following the official VoxPopuli partition.

2) Watermark Embedding Process (Generator)

The Generator [14] is the component responsible for embedding the watermark. It takes an audio signal, input is denoted as s . The generator then predicts a watermark waveform δ of the same size. The watermarked audio, s_w, is

created by simply adding the watermark to the original audio, following the equation:

$$s_w = s + \delta \quad (1)$$

The generator can optionally encode a secret message of 16 bits into the watermark, providing 65,536 possible choices.

3) Watermark Extraction Process (Detector)

The Detector [14] analyzes the watermarked audio, sw. It outputs a probability score between 0 and 1, for each time step. This probability indicates the likelihood of a watermark being present at that specific sample. The detector can also extract the secret 16-bit message if one is embedded in the watermark.

4) Training and Optimization

AudioSeal's generator and detector are trained simultaneously in a joint optimization process [14]. This co-training ensures that the generator creates a watermark that is both imperceptible to human ears and highly detectable by the detector.

- Imperceptibility: A key optimization is a novel perceptual loss function inspired by auditory masking. This custom loudness-based loss minimizes the audible difference between the original and watermarked audio, ensuring minimal signal alteration.
- Localized detection: The model is trained with a specific augmentation strategy where random segments of watermarked audio are replaced with silence or non-watermarked audio from the same batch.
- Robustness: To maximize its resilience, the training includes a regimen of audio distortions, such as compression, noise addition, resampling, and others. This "resilient by design" approach ensures the watermark remains detectable even after manipulation.

5) Technical Environment

AudioSeal requires Python version 3.8 or higher and PyTorch version 1.13.0 or higher. Other necessary libraries include omegaconf, julius, ffmpeg and numpy. The opensource code is available on GitHub [17].

6) Training details

Our model was trained from scratch at a 16 kHz sampling rate. The training run for 100 epochs using the Adam optimizer, batch size 12 (limited by the 40 GB GPU memory of a single NVIDIA A100), update interval 2,000 steps, a sample pool of 35,000 training samples for each epoch was randomly selected from the dataset (num_sample (train) = 35,000) and learning rate at 4×10^{-5}. To encourage stable convergence, training was conducted in two phases. During the first 70 epochs, only the localization and detection losses were weighted, while perceptual losses were disabled. This allowed the model to prioritize accurate watermark localization and detection. Moreover, we fine-tuned the model from epoch 70 to 100, during which the learning rate was reduced to 1×10^{-5}. The checkpoint obtained after the initial 70 epochs was primarily optimized for the core functions of watermark embedding and detection with speed and precision. In the fine-tuning stage, the model was further trained to enhance imperceptibility and robustness by gradually introducing perceptual losses and increasing them to their final weights: adversarial loss (adv) = 4.0, feature-matching loss (feat) = 4.0, L1 reconstruction loss (l1) = 0.1, multi-scale spectrogram loss (msspec) = 2.0, and time-frequency loudness ratio loss (tf_loudnessratio) = 10.0. All other augmentations and hyperparameters followed the default settings provided in the AudioCraft GitHub [18].

B. Prototype application

To make our watermarking model feasible, we built a prototype web application, VoiceMark, that integrated the embedding and detection modules into a complete system for end user. The application contains a React/Next.ts frontend and a Node.js backend interacting with the watermarking engine via REST APIs. The backend handles the audio watermark embedding/detection procedures, as well as management of files and user data. The frontend provides a user-friendly interface to allow users of all background easily access the tool.

The above system-level deployment is not the main contribution of the paper but demonstrates the realizability of using our watermarking scheme in a real environment. It also facilitates end-to-end testing with real user interaction, i.e., the watermarking pipeline is not merely theoretically correct but practically feasible as well.

IV. Experimental Setup

A. Evaluation dataset

The dataset used for evaluating the models' performance consisted of 100 unseen audio samples, each one minute in length: 50 music samples from the Free Music Archive [19] and 50 speech samples from English public-domain LibriVox [20]. Each sample was then watermarked using our trained AudioSeal model and the baseline method (hybrid SVD with STFT). To assess audio quality, we compared the 100 watermarked audio samples against their corresponding non-watermarked original using the established comparative metrics. For detection evaluation, we applied various edits to the watermarked samples to simulate common attacks and then measured detection performance using the chosen metrics.

B. Audio quality assessment

We evaluate the audio quality of our trained AudioSeal model against the baseline hybrid method of SVD and STFT using two metrics: Scale-Invariant Signal-to-Noise Ratio (SISNR) [21] and Perceptual Evaluation of Speech Quality (PESQ) [22] with 100 watermarked samples and 100 original samples (non-watermarked).

SI-SNR is a fidelity metric that measures how much distortion is introduced into an audio signal after processing [21]. The output is expressed in decibels (dB) and the range is unbound. A score of 0 dB means the distortion energy is equal to the signal energy, indicating poor quality. Positive values between 10 to 20 dB indicate good quality, and values above 20 dB are considered very high quality. In general, higher SI-SNR corresponds to better preservation of the original audio.

PESQ is an objective perceptual quality metric standardized by ITU-T [22]. It estimates how a human listener would rate the quality of audio after degradation by watermarking, compression, or other edits. PESQ produces scores between –0.5 and 4.5, with higher values indicating better perceptual quality. A score closes to 4.5 reflects transparent audio quality, while scores near 1.0 or lower reflect heavily degraded audio.

C. Detection robustness evaluation

This section discusses the robustness evaluation of our trained AudioSeal model and the baseline hybrid SVD with STFT method against various common audio distortions as follows:

- MP3 Compression (32 kbps): The audio samples were compressed to a low bitrate of 32 kbps, simulating lossy distribution over bandwidth limited channels.
- Resampling (32 kHz): Each sample was resampled from 16 kHz to 32 kHz and back to 16 kHz.
- Speed Change (1.25×): The playback speed of each sample was increased by a factor of 1.25.
- Additive White Gaussian Noise ($\sigma = 0.05$): Random Gaussian noise was added directly to the waveform with a fixed deviation of 0.05.
- Additive Pink Noise ($\sigma = 0.01$): Pink noise with a standard deviation of 0.01 was added to the waveform.

Evaluation performance was quantified using the following evaluation metrics:

- Accuracy (Acc): The average of detection success on positives (TPR) and success on negatives (1FPR). Ranges from 0.0 to 1.0, with higher values indicating better overall detection accuracy.
- True Positive Rate (TPR): Measures the watermarked audio correctly identified as containing a watermark. Ranges from 0.0 to 1.0, with higher values indicating fewer missed detections.
- False Positive Rate (FPR): Measures the proportion of non-watermarked audio incorrectly classified as watermarked. Ranges from 0.0 to 1.0, with lower values indicating fewer false detection.
- Bit Error Rate (BER): Measures the watermark bits incorrectly decoded. Values range from 0.0 to 1.0, with lower values indicating higher decoding fidelity.

V. Results and Discussion

This section presents and analyzes the experimental results obtained from evaluating our custom AudioSeal model against the baseline hybrid SVD with STFT method.

Table I shows the comparison between AudioSeal versus SVD & STFT in term of audio quality using SI-SNR and PESQ.

TABLE I. Audio Quality Assessment

Methods	*SI-SNR*	*PESQ*
AudioSeal	40.208	4.354
SVD & STFT	35.225	4.162

For audio quality assessment, Table I shows that our custom AudioSeal model outperforms the baseline SVD with STFT method in terms of SI-SNR and PESQ. This shows that the watermark embedding process brings very minimal perceptual distortion. Fig.1 and Fig.2 further support this finding by comparing the spectrograms of non-watermarked and watermarked sample audio. The overall spectral structure is preserved, with only subtle, imperceptible changes which is the watermark itself, which can barely be seen as a horizontal line in between 4096 to 8192 Hz, confirming that the watermark does not noticeably degrade audio quality.

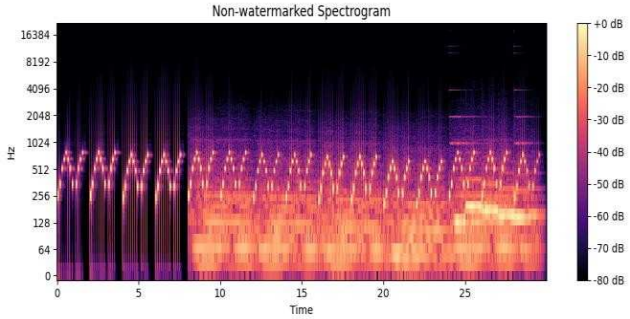

Fig. 1. Spectrogram of non-watermarked audio, used as a baseline for comparison with the watermarked version.

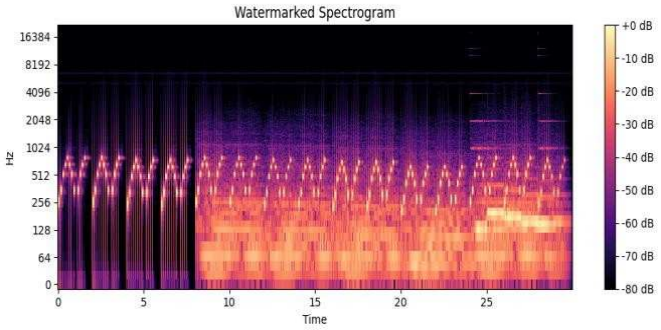

Fig. 2. Spectrogram of watermarked audio, showing the preserved structure even after watermarking.

Table II shows the detection performance of AudioSeal model under common audio edits and attacks, including MP3 compression, resampling, speed change, white noise, and pink noise.

TABLE II. Detection Performance of Custom Model Under Common Edits and Attacks

	AudioSeal		
Edit/Attacks	***Acc.***	***TPR/FPR***	***BER***
MP3	0.94	0.89/0.00	0.109
Resampling	0.96	0.92/0.00	0.093

	AudioSeal		
Edit/Attacks	***Acc.***	***TPR/FPR***	***BER***
Speed	0.56	0.22/0.10	0.282
White Noise	0.91	0.92/0.10	0.088
Pink Noise	0.97	0.95/0.00	0.089

Table III shows the detection performance of the baseline SVD with STFT approach under the same set of audio edits and attacks.

TABLE III. DETECTION PERFORMANCE OF SVD WITH STFT UNDER COMMON EDITS AND ATTACKS

	SVD with STFT		
Edit/Attacks	***Acc.***	***TPR/FPR***	***BER***
MP3	0.76	0.62/0.10	0.228
Resampling	0.77	0.60/0.11	0.236
Speed	0.70	0.50/0.12	0.247
White Noise	0.70	0.55/0.14	0.268
Pink Noise	0.73	0.58/0.12	0.249

For robustness assessment, Table II and III shows that the custom model consistently performs better than most common audio edits and attacks. The only exception is the Speed edit, where the baseline SVD/STFT achieves up to 0.70 accuracy compared to only 0.56 for AudioSeal. Figs. 3–7 support these results by visualizing the effect of an attack through spectrograms. In Figs. 3–5, the aforementioned watermark can still be seen and perfectly preserved after the attacks. Note that in Figs. 6–7, even though the watermark cannot be seen with this form of visualization due to the nature of the attack, the result measured in Table II still prove the robustness of the watermark against them.

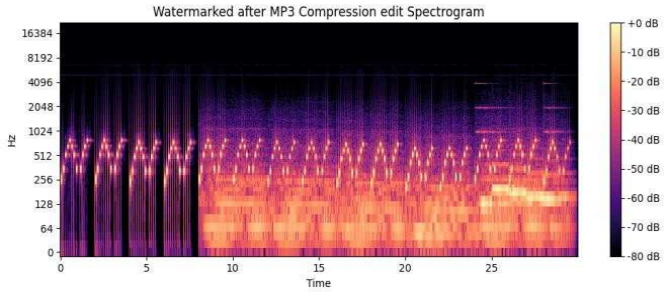

Fig. 3. Spectrogram of the watermarked audio after a MP3 compression attack, showing the preserved watermark.

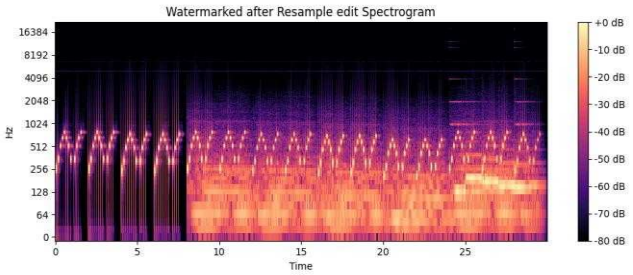

Fig. 4. Spectrogram of the watermarked audio after resampling, showing the preserved watermark.

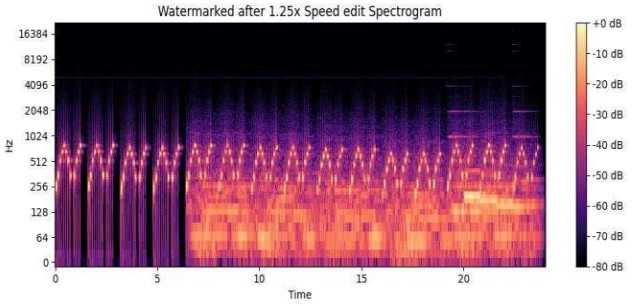

Fig. 5. Spectrogram of the watermarked audio after a speed edit, showing the preserved watermark.

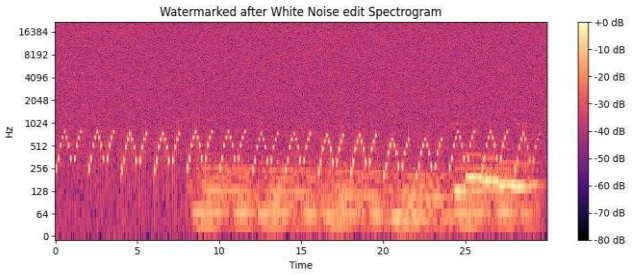

Fig. 6. Spectrogram of the watermarked audio after a white noise edit.

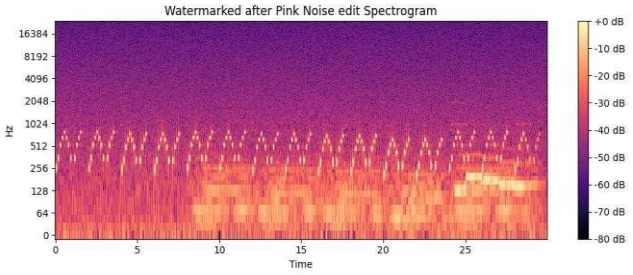

Fig. 7. Spectrogram of the watermarked audio after a pink noise edit.

For the trade-off, we prioritized the AudioSeal model's core performance: embedding an imperceptible watermark into the audio waveform and ensuring reliable detection, extraction, and bit decoding from the watermarked audio. Robustness was considered secondary and was fine-tuned only after optimizing these primary aspects.

VI. CONCLUSION AND FUTURE WORK

Our results show that classical watermarking methods generally underperform compared to neural network-based approaches such as AudioSeal. Although our implementation of AudioSeal was not trained to the fullest extent recommended by

its developers and in the AudioSeal paper [14], this was primarily due to hardware and time constraints. Our experiments were limited to a single NVIDIA A100 GPU with 40 GB of memory and 250 GB of storage, which restricted the size of the dataset we could process and required extensive pilot runs to identify hyperparameters and training settings suitable for our hardware. Combined with a timeframe of only four months for both training and research, these factors prevented us from training the model to its full capacity. Nevertheless, the model still outperformed classical techniques. This highlights the promise of deep learning-based watermarking for enhancing voice data protection.

While our implementation demonstrates that AudioSeal can be adapted under limited hardware and time constraints, future work should focus on scaling training to more closely match the setup in the AudioSeal paper [14]. Specifically, although our model was trained for 100 epochs, the detector network continued to show improvements beyond this point, suggesting that extended training (120 epochs or more) could further enhance performance. Additionally, the hyperparameter num_sample (train) was fixed at 35,000 in our experiments (compared to the default of 500,000). Since this parameter controls the number of training samples drawn per epoch, systematically exploring different values may lead to better trade-offs between training speed and model performance. Finally, leveraging larger computational resources (e.g., multi-GPU setups) would allow longer training schedules, larger datasets, and closer adherence to the AudioSeal setting. These directions would help unlock the full potential of the model in terms of imperceptibility and robustness.

REFERENCES

[1] F. Muhly, E. Chizzonic, and P. Leo, 'AI-deepfake scams and the importance of a holistic communication security strategy', International Cybersecurity Law Review, pp. 1–9, 2025.

[2] S. S. Kolekar, D. J. Richter, M. I. Bappi, and K. Kim, 'Advancing AI voice synthesis: Integrating emotional expression in multi-speaker voice generation', in 2024 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2024, pp. 193–198.

[3] H. Özer, B. Sankur, and N. Memon, 'An SVD-based audio watermarking technique', in Proceedings of the 7th Workshop on Multimedia and Security, New York, NY, USA, 2005, pp. 51–56.

[4] A. Kanhe and A. Gnanasekaran, 'Robust image-in-audio watermarking technique based on DCT-SVD transform', EURASIP Journal on Audio, Speech, and Music Processing, vol. 2018, no. 1, p. 16, 2018.

[5] S. A. M. Rizvi and S. P. S. Chauhan, 'Robust Digital Audio Watermarking Based on SVD and Modified Firefly Algorithm', Journal of Information Security, vol. 9, no. 1, pp. 1–11, 2017.

[6] X. Liu, X. Li, C. Shi, X. Niu, and L. Xiong, 'A novel SVD-based adaptive robust audio watermarking algorithm', Multimedia Tools and Applications, vol. 83, no. 27, pp. 69443–69465, 2024.

[7] A. Electa Alice Jayarani, M. R. Bhatt, and D. D. Geetha, 'Zero Watermarking on Audio Based on STFT', in 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 2018, pp. 253–256.

[8] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, 'Detecting Voice Cloning Attacks via Timbre Watermarking', in Network and Distributed System Security Symposium, 2024.

[9] Z. Gan, X. Hu, S. Li, Z. Qian, and X. Zhang, 'SyncGuard: Robust Audio Watermarking Capable of Countering Desynchronization Attacks', arXiv [cs.CR]. 2025.

[10] D. Kirovski and H. Malvar, 'Robust spread-spectrum audio watermarking', 02 2001, vol. 3, pp. 1345–1348 vol.3.

[11] J. Zhao, T. Zong, Y. Xiang, L. Gao, W. Zhou, and G. Beliakov, 'Desynchronization Attacks Resilient Watermarking Method Based on Frequency Singular Value Coefficient Modification', IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 29, pp. 2282–2295, June 2021.

[12] K. Pavlović, S. Kovačević, I. Djurović, and A. Wojciechowski, 'Robust speech watermarking by a jointly trained embedder and detector using a DNN', Digital Signal Processing, vol. 122, p. 103381, 2022.

[13] M. K. Singh, N. Takahashi, W. Liao, and Y. Mitsufuji, 'SilentCipher: Deep Audio Watermarking', in Interspeech 2024, 2024, pp. 2235– 2239.

[14] R. San Roman, P. Fernandez, H. Elsahar, A. D´efossez, T. Furon, and T. Tran, 'Proactive Detection of Voice Cloning with Localized Watermarking', ICML, 2024.

[15] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, 'High Fidelity Neural Audio Compression', arXiv [eess.AS]. 2022.

[16] C. Wang et al., 'VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation', arXiv [cs.CL]. 2021.

[17] facebookresearch, "GitHub - facebookresearch/audioseal: Localized watermarking for AI-generated speech audios, with SOTA on robustness and very fast detector," GitHub, 2024. https://github.com/facebookresearch/audioseal

[18] "AudioCraft," GitHub, Nov. 27, 2023. https://github.com/facebookresearch/audiocraft

[19] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, 'FMA: A Dataset For Music Analysis', arXiv [cs.SD]. 2017.

[20] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, 'MLS: A Large-Scale Multilingual Dataset for Speech Research', in Interspeech 2020, 2020.

[21] Y. Luo and N. Mesgarani, 'TasNet: time-domain audio separation network for real-time, single-channel speech separation', arXiv [cs.SD]. 2018.

[22] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, 'Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs', in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing.