Interspeech 2024
1-5 September 2024, Kos, Greece

SecureSpectra: Safeguarding Digital Identity from Deep Fake Threats via Intelligent Signatures

Oguzhan Baser[1], *Kaan Kale*[2], *Sandeep P. Chinchali*[1]

[1]Department of Electrical and Computer Engineering, The University of Texas at Austin, USA
[2]Department of Electrical and Electronics Engineering, Bogazici University, Turkey

oguzhanbaser@utexas.edu

Abstract

Advancements in DeepFake (DF) audio models pose a significant threat to voice authentication systems, leading to unauthorized access and the spread of misinformation. We introduce a defense mechanism, SecureSpectra, addressing DF threats by embedding orthogonal, irreversible signatures within audio. SecureSpectra leverages the inability of DF models to replicate high-frequency content, which we empirically identify across diverse datasets and DF models. Integrating differential privacy into the pipeline protects signatures from reverse engineering and strikes a delicate balance between enhanced security and minimal performance compromises. Our evaluations on Mozilla Common Voice, LibriSpeech, and VoxCeleb datasets showcase SecureSpectra's superior performance, outperforming recent works by up to 71% in detection accuracy. We open-source SecureSpectra to benefit the research community.

Index Terms: audio cloning, deepfake, voice spoofing, voiceprint, anti-spoofing, audio signature, differential privacy

1. Introduction

The escalating sophistication of DeepFake (DF) technologies is increasingly compromising the security of voice-authenticated applications. Recent DF models can clone voices from recordings as brief as 10-second recordings [1], amplifying the risks in critical domains, such as access to banking and medical records by voice [2]. Consequently, there is an urgent need to safeguard against voice misuse and to ensure the security of voice-based interactions. Notably, our work is partly motivated by DF attacks targeting political leaders [3, 4], which raises concerns similar to the Cambridge Analytica case [5] and leads to incompliance with GDPR [6].

Existing state-of-the-art methods use machine learning (ML) models to identify cloned audio [7]. However, these ML classifiers struggle against DF models. This is largely because advanced cloning schemes [1, 8, 9] rely on Generative Adversarial Networks (GANs), and GANs are optimized to deceive their internal ML classifiers (discriminators) to refine clones. Due to a lack of additional orthogonal information, traditional ML models are not effective in DF detection and upper-bounded by the discriminator performance. Our key innovation lies in enriching the original audio with orthogonal information, which we call *the signature*, to improve the DF detection performance. Also, this method applies to any type of DF model, such as diffusion models [10] and VAEs [11], in addition to GANs.

Specifically, as shown in Fig.1, our design[1] provisions an irreversible signal processing module (green) at the audio owner's end. This module enables authenticated users to sign the audio with their private key, generating signed audio that is resistant to the extraction of the exact signature without the private key yet remains verifiable. Our key technical insight, as we empirically demonstrate in Fig.3, lies in the observation that DF models encounter challenges in generating high-frequency (HF) signals due to overfitting to human speech. This phenomenon arises primarily from the irrelevant nature of HF signals, such as background bird noise, during training. DF models prioritize learning speech waveforms, neglecting uncorrelated high-frequency noise. Our approach is complementary to advances in generative models and instead pertains to authenticity-critical applications such as banking and election campaigns.

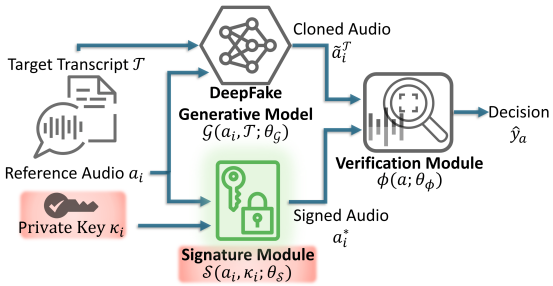

Figure 1: ***Digital Identity Secured Voice Authentication**: Imagine a public figure, Alice, who releases a speech of a_i. Malicious Eve (top row) employs a DF model $\mathcal{G}$, parameterized by $\theta_{\mathcal{G}}$, to mimic Alice's voice in a transcript $\mathcal{T}$, creating a clone $\tilde{a}_i^{\mathcal{T}}$. We aim to develop a verification module ϕ parameterized by θ_ϕ to decide $\hat{y}_a$ if an audio a comes from Alice. Our approach (bottom row) first gives Alice a private key κ_i. Then, our novel signature module (green) combines her voice a_i with her key κ_i to produce signed audio a_i^*. The signed audio closely resembles the original while being distinguishable from fake versions. The verifier ϕ can identify the signature in an audio without revealing it. If Eve attempts to use the signed audio a_i^* in her model, the generated clone $\tilde{a}_i^{\mathcal{T}}$ does not contain the signature. The key and the signature module are kept confidential (red) to prevent attacks.*

Literature Review: Various methods have been explored to ensure the speech authenticity, as in Fig.1. Anonymization is commonly employed to obscure speech identity, and it significantly degrades speech naturalness and intelligibility [12, 13]. Watermarking stands out for embedding information within audio signals, facilitating post-processing regeneration [14, 15, 16]. Despite its utility, the ease of the watermark extraction poses security threats. Besides, the objectives of anonymization (hiding user identity) and watermarking (maintaining data copy-

[1]https://github.com/UTAustin-SwarmLab/SecureSpectra

10.21437/Interspeech.2024-2349

right) are distinct from our DF countermeasure (securing user identity). Signal processing methods offer cost-effective signatures but suffer from reversibility, which allows the extraction of the original signature since both the original speech and its signed version are public [17, 18]. ML classifiers have emerged to discern between original and cloned signals [19, 20, 7, 21]. However, a generative DF model is trained to deceive its discriminators, analogous to these classifiers [22]. Hence, their performance is limited by the DF models' discriminators.

Principal Contributions: In light of prior work, our contributions are four-fold. First, we systematically analyze the impact of DF models on the frequency band, highlighting a significant energy reduction in the HF regime, as empirically demonstrated in Fig. 3. Second, we introduce a secure signature method leveraging this observation to detect unauthorized DF attacks in voice-authenticated systems. Our method, depicted in Fig.1, surpasses the performance of existing state-of-the-art ML-based detection mechanisms. Third, we explore the integration of Differential Privacy (DP) techniques to safeguard private key integrity, mitigating the risk of unauthorized access. Finally, we open-source our code[1], facilitating further research in this area.

2. Methodology

SecureSpectra comprises three key components: a generative model for DF attacks, a signature model, and a verification model to detect the signature's presence, as shown in Fig.1.

Adversary's Threat Model: We now describe our assumptions on the adversary's threat model and how each component in Fig.1 mitigates different types of attacks. First, the adversary can train a DF model to create a forged audio. However, our signature module (green) effectively thwarts such attempts. Second, a sophisticated adversary can steal the private signature model's weights. However, we combat this by making the users sign their audio with their private keys (red). Finally, authenticated adversaries with a signature model can attempt to reverse engineer private keys from the verifier model. To combat this, we add DP to private keys during training to effectively prevent such reverse engineering attacks. We now describe each of these threats and their defense in turn.

Generative DeepFake Model: Let $\{a_i\}_{i=1}^N$ be N unique samples drawn from the data distribution $\mathbf{D}$. An adversary trains a GAN on these samples to create synthetic audio $\tilde{a}_i^{\mathcal{T}}$ that resembles the original a_i and vocalizes the provided text $\mathcal{T}$ from a text corpus $\mathbf{T}$. We focus on GANs for notational convenience, though the concepts apply to any generative models, such as diffusion models [10] and VAEs [11]. A GAN comprises a Generator $\mathcal{G}$ and a Discriminator $\mathcal{D}$. Generator $\mathcal{G}(a_i, \mathcal{T}; \theta_{\mathcal{G}})$, parameterized by $\theta_{\mathcal{G}}$, takes a reference audio a_i and a transcript $\mathcal{T}$ and generates a cloned audio $\tilde{a}_i^{\mathcal{T}}$ corresponding to the transcript. Discriminator $\mathcal{D}(a; \theta_{\mathcal{D}})$, parameterized by $\theta_{\mathcal{D}}$, takes audio data a and predicts whether it is synthetic $\tilde{a}_i$, or not a_i. Formally, the problem is expressed as:

$$\min_{\theta_{\mathcal{G}}} \max_{\theta_{\mathcal{D}}} \mathbb{E}_{a\sim\mathbf{D}}[\log \mathcal{D}(a;\theta_{\mathcal{D}})] + \mathbb{E}_{a,\mathcal{T}\sim\mathbf{D},\mathbf{T}}[\log(1-\mathcal{D}(\mathcal{G}(a,\mathcal{T};\theta_{\mathcal{G}});\theta_{\mathcal{D}}))]. \quad (1)$$

The objective is to train the generator parameters $\theta_{\mathcal{G}}$ to produce synthetic audio that is indistinguishable from real audio by deceiving the discriminator, while the discriminator parameters $\theta_{\mathcal{D}}$ are trained to distinguish original samples from clones.

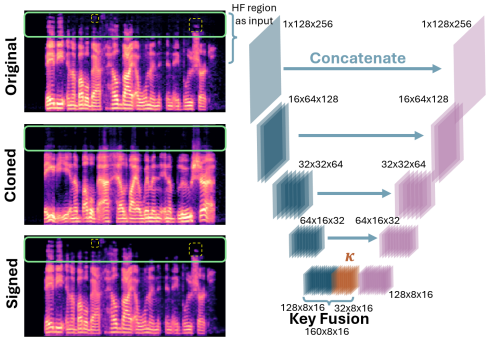

Figure 2: ***Key Observation for High-Frequency Regime:*** *By comparing the spectrograms of the original audio (top) and its cloned version (middle) for the same transcript, we observe a distinct absence of HF content (green) in the DF audio. This discrepancy arises from the bias of DF models toward mimicking user speech, which predominantly emphasizes lower-frequency regions. A U-net (right) signs the audio (bottom) with unrecognizable slight modifications (yellow) in the HF regime.*

Signature Model: Our main goal is to sign the audio a_i such that its clone $\tilde{a}_i^{\mathcal{T}} = \mathcal{G}(a_i^*, \mathcal{T}; \theta_{\mathcal{G}})$ lacks the signature. We introduce a signature module $\mathcal{S}$ that operates on an audio a_i and a private key κ_i to generate a signed version of the audio a_i^*. Each user has a distinct private key κ_i to sign the audio samples: $a_i^* = S(a_i, \kappa_i; \theta_{\mathcal{S}})$. The parameters $\theta_{\mathcal{S}}$ are optimized to minimize the ℓ_1 norm of the original and signed audio samples. The ℓ_1 norm encourages sparsity in the spectrum with minor changes, preserving HF quality [23]. The signature module's loss is formulated as:

$$\mathcal{L}_{\mathcal{S}} := \frac{1}{N}\sum_{i=1}^{N} ||a_i - S(a_i, \kappa_i; \theta_{\mathcal{S}})||_1. \quad (2)$$

We fuse the private key with the audio representation at the deepest layer of the U-Net architecture [24]. Such integration ensures that the private key is intricately woven into the utterance, thereby complicating any attempts at extraction without compromising the audio quality. This mechanism secures the keys and ensures the signature embedding is inherently resilient to reverse engineering. This resilience is further bolstered by the signature model's confidentiality and DP (Sec. 2). Deciphering the embedded signature without explicit knowledge of the signature model and private key becomes extremely challenging. This layered approach to security ensures the integrity of audio signatures, providing a robust defense against unauthorized access and manipulation.

Verification Model: Following the design of the signature model, our next logical step is to introduce a verification model ϕ designed to classify audio signals as signed or not. It is a *public* model and reveals "only the existence of the signature". The model ϕ outputs a binary prediction $\hat{y}$ for audio a_i through $\hat{y} = \phi(a_i; \theta_\phi)$, where θ_ϕ represents the parameters trained to minimize the verifier loss $\mathcal{L}_\phi$ before public release. For given unsigned audio with ground truth (a_i, y_i) and signed audio with ground truth (a_i^*, y_i^*), the verifier loss $\mathcal{L}_\phi$ is defined as:

$$\mathcal{L}_\phi := -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\phi(a_i;\theta_\phi)) + (1-y_i^*)\log(1-\phi(a_i^*;\theta_\phi)), \quad (3)$$

which is the extended binary cross-entropy loss for our setting.

Joint Training: After designing the models, we ensure that the signature and verification modules complement each other's functions. Achieving both hidden signatures and accurate verification requires the signature model to integrate verifier gradients during training. Thus, both models learn from each other by minimizing the joint loss shown as:

$$\min_{\theta_\phi, \theta_\mathcal{S}} \mathcal{L}_\mathcal{S} + \mathcal{L}_\phi. \tag{4}$$

The training begins with a forward pass through the signature model to compute its loss. Then, both the signed and original audio samples undergo a forward pass in the verification model to calculate the verifier loss. These two losses are then aggregated, as shown in Eq. 4, for optimization. Subsequently, the gradients propagate backward through both models, with particular attention to the verification model. To calibrate the potential instabilities arising from the verification module and affecting the signature module, the learning rate for the verifier model is set to one-tenth of that for the signature model.

Guarded Inference: Following the training phase, the verification model becomes publicly accessible for authenticity and signature checks, while the signature model remains confidential to prevent unauthorized signature replication. This approach safeguards against adversarial attempts to create counterfeit signatures because it requires both the signature model and private keys to generate valid signatures. Only authorized users with private keys can embed signatures, maintaining the process's security and privacy. This dichotomy effectively protects digital audio identities from unauthorized access or manipulation.

Differential Privacy (DP) on Private Keys: After the defense against external threats, we now address insider risks with DP. Authorized malicious users or exposure of the signature model allows adversaries to reverse engineer the system and deduce private keys used in model training [25]. To address this, we introduce precisely calibrated DP noise to the private keys during the signature model's training. This ensures that the model can embed signatures without incorporating identifiable information about individual keys into its parameters. This prevents adversaries from extracting private keys when they have the signature model. Formally, private keys κ are preserved by adding DP noise η sampled as:

$$\eta \sim \frac{1}{2b} \exp\left(-\frac{|\eta|}{b}\right), \tag{5}$$

where the noise scale b is the ratio $\Delta K/\epsilon$. Here, ΔK represents the global sensitivity, indicating the maximal cosine distance between any two private keys κ. ϵ denotes the privacy loss. It configures the trade-off between key protection and model accuracy and ensures adherence to the ϵ-DP standard [26]. This standard mandates that the alteration of a single element in a private key, resulting in a transition from κ to κ', should not significantly affect the probability distribution of the output of the model $\mathcal{S}$, as in $\Pr[\mathcal{S}(\kappa) \in R] \leq e^\epsilon \times \Pr[\mathcal{S}(\kappa') \in R]$, where R represents any arbitrary subset of outcomes. Each operation in the pipeline enhances the integrity of signed audio, addressing DF threats and privacy concerns in voice-based applications. Each signature is intricately tied to the utterance and private key, adding extra robustness.

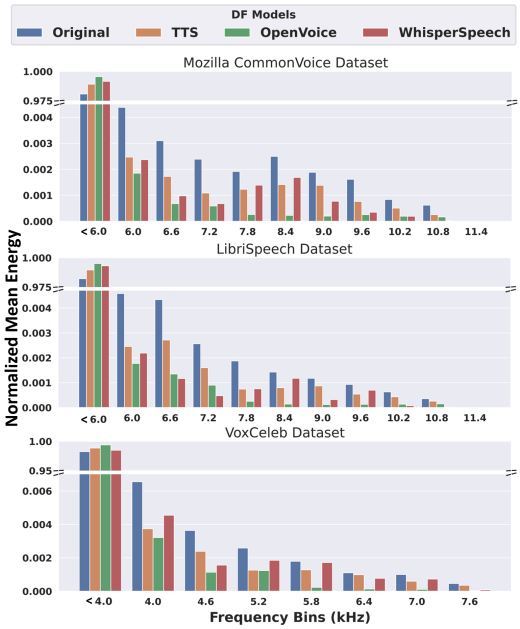

Figure 3: ***Spectral Analysis of Original and Cloned Audio:*** *We empirically analyzed the spectral content across the original audio recordings (blue) and their cloned counterparts (orange, green, red) generated by state-of-the-art DF models. The analysis encompasses CommonVoice, LibriSpeech, and VoxCeleb datasets, with all audio samples converted into spectrograms. The frequency spectrum was segmented into bins, each representing a 600 Hz bandwidth, where the energy content within each bin was averaged and plotted. The results, derived from testing on three known audio datasets with three advanced DF models, highlight a* ***discernible attenuation in the HF components in the DF-generated audio compared to the original ones****, indicating a characteristic shortfall of the DF models in replicating the HF energy profile of genuine audio recordings.*

3. Experimental Setup

In this section, we detail the experimental setup, outlining the datasets, models, hyperparameters, and evaluation metrics.

Datasets: To demonstrate the efficacy of our proposed pipeline, we utilize three widely recognized speech datasets: **Mozilla Common Voice** [27], **LibriSpeech** [28], and **VoxCeleb** [29]. The ASVspoof2021 [30] is unsuitable for our analysis as it lacks regenerable DFs from signed audio to compare its effects.

DF Models: We evaluated the resilience of SecureSpectra and the effects of DF models on spectrograms. We selected three state-of-the-art GAN models for this purpose: **Coqui.ai TTS** [1], **OpenVoice** [8], and **WhisperSpeech** [9]. These models were chosen for their ability to synthesize speech from text inputs using "reference audio," making them ideal for examining the impact of DF on the authenticity of signed audio. Using these models, we cloned audio from speech datasets to vocalize texts by Camus, Dickens, Orwell, and Thoreau. This process created cloned datasets that showcased not only the HF synthesis capabilities of the DF models but also our verification system's ability to detect audio authenticity.

Benchmarks: Our evaluation of SecureSpectra encompasses three distinct configurations: Verification Only, Signature + Verification, and Signature + Verification with DP noise. The **Verification Only** serves as a baseline, demonstrating detec-

tion performance without our signature embedding technique. It highlights the improvement done by our signature mechanism. The core configuration, **Signature + Verification**, demonstrates robustness against DF attacks, achieved through the joint training of the signature and verification modules. Introducing **DP into the Signature + Verification** configuration prevents reverse engineering threats and balances performance-privacy trade-offs in trustless settings. Furthermore, SecureSpectra's performance is compared with two state-of-the-art anti-spoofing solutions from INTERSPEECH 2023. The first method, **Whisper Based** [19], uses the Whisper [31] features of audio to detect DF attacks with an ML classifier. The second, **SASV2-Net** [20], utilizes a multi-stage training approach that transfers information from speaker verification in the VoxCeleb dataset to DF detection on the ASVspoof2019 dataset, showcasing orthogonal information leverage similar to our approach in some aspects.

Evaluation Metric: To compare the benchmarks, we assess the test accuracies of classification models on the recordings from 100 individuals separately. Each individual's test set comprises cloned and original samples with equal density. Also, we evaluate the benchmark Equal Error Rates (EERs) on each dataset.

Signatures: Each key κ, consisting of 32 binary digits, can be viewed as vectors. To gauge the global sensitivity ΔK of these keys, we compute the maximum cosine distance between them. Subsequently, we introduce DP noise η with $\epsilon = 30$. The noise is injected during training $(\kappa+\eta)$ to ensure that the model learns from a distribution of private keys rather than individual ones.

Model Architectures: The signature model, shown in Fig.2, employs a U-Net tailored for processing the HF spectrograms of audio. It consists of a 5-layer CNN encoder that projects the input. At the deepest layer of the encoder, the private key κ is concatenated with the latest feature map and convolved to maintain the original feature map size. This augmented feature map then undergoes decoding through four layers, each receiving a skip connection from the encoder at the corresponding level. For verification, a 7-layer CNN with kernel size 3 condenses the input features by doubling the channel size in each layer, ultimately leading to a fully connected layer that outputs a binary prediction on the presence of κ without disclosing any other information. More model details are available in our repository[1].

Hyperparameters: Both models are initialized using Xavier initialization [32], and the training is performed using the Adam optimizer [33] with an initial learning rate of 2e-4 for the signature model and 2e-5 for the verification model. Training is carried out to minimize total validation loss, with early stopping after 10 rounds of no improvement. The experiments were executed with 8 NVIDIA RTX A5000 GPUs (8x24 GB RAM).

4. Results

Our empirical analysis (Fig.3) demonstrates that DF models tend to overlook HF patterns while generating cloned speech. Furthermore, in Fig.4 and Table 1, we demonstrate that SecureSpectra performs better than recent work in the literature. Additionally, we show that adding DP noise reduces performance slightly with the benefit of a more secure pipeline.

How does orthogonal information affect DF detections?

Our analysis suggests that DF generators, designed to deceive their intrinsic GAN discriminators, similar to our Verification Only module, can be effectively countered by leveraging orthogonal information. This is evidenced by the performance improvements observed when transitioning from a conventional CNN detection module (green) to methods employing auxiliary information. Specifically, Whisper Based (blue) and SASV2-Net (orange) leverage transcription and speaker verification information, respectively. Unlike these approaches, which rely on information correlated with the audio's content, SecureSpectra embeds completely orthogonal information, the κ signature, into the audio (red and purple). This strategy yields a robust mechanism for DF detection across all user scenarios.

Limitations: SecureSpectra's scalability is linked to the length of private keys. As user numbers grow, the requisite key length increases, potentially complicating the model's training process. Also, our evaluation excludes the audio channel noise.

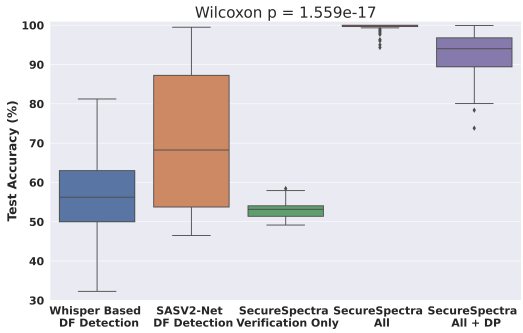

Figure 4: ***User-Level Performance Across Benchmarks:*** *We evaluate the DF detection benchmark test accuracies across 100 distinct users with 200 audio samples each (100 original, 100 cloned). The orange and blue box plots show the accuracies of the two recent works. The green box plot provides a baseline of our pipeline without signature embedding. The purple and red box plots show the performance of our approach with and without DP noise, respectively.* ***Our method, particularly with signature embedding, surpasses existing models****, enhancing verification-only accuracy by 81% and outperforming comparative works by 71% and 42%. DP noise adds additional security with a marginal decrease in accuracy by 4%.*

Table 1: ***The Benchmark % EERs (↓) on Individual Datasets:*** *SecureSpectra dramatically reduces EER across all datasets.*

Benchmarks	CV [27]	LS [28]	VC [29]
Whisper [19]	41.6	36.2	36.8
SASV-2 [20]	4.01	12.2	3.75
Verification Only	48.3	43.6	46.0
SecureSpectra	1.50	1.10	1.36
SecureSpectra + DP	2.96	2.74	2.83

5. Conclusion

This paper introduces SecureSpectra, a robust method for protecting audio from cloning attacks using HF signatures that DF models cannot accurately reproduce. Through empirical analysis, we show the inability of DF models to mimic HF signals, thereby significantly increasing detection accuracy by embedding signatures. We open-source SecureSpectra[1] to support ongoing research. As DF technologies advance, SecureSpectra offers robust protection for digital identity. Moving forward, we aim to enhance SecureSpectra by advancing the verification module through multitask learning with speaker verification. This approach ensures that users can only sign their personal audio, strengthening both robustness and security.

6. Acknowledgements

This material received support from the National Science Foundation under grant no.2148186 and is further supported by funding provided by federal agencies and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program. This article solely reflects the opinions and conclusions of its authors and does not represent the views of any sponsor.

7. References

[1] E. Gölge, "Coqui.ai," 2023, available: https://github.com/coqui-ai/TTS, [Accessed: 12-Jan-2024].

[2] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, p. 113368, 2023.

[3] M. Pawelec, "Deepfakes and democracy (theory): how synthetic audio-visual media for disinformation and hate speech threaten core democratic functions," *Digital society*, vol. 1, no. 2, p. 19, 2022.

[4] T. Kirchengast, "Deepfakes and image manipulation: criminalisation and control," *Information & Communications Technology Law*, vol. 29, no. 3, pp. 308–323, 2020.

[5] M. Pavlíková, B. Šenkýřová, and J. Drmola, "Propaganda and disinformation go online," *Challenging online propaganda and disinformation in the 21st century*, pp. 43–74, 2021.

[6] I. G. P. TEAM, *EU General Data Protection Regulation (GDPR), third edition: An Implementation and Compliance Guide*, 3rd ed. IT Governance Publishing, 2019, available: http://www.jstor.org/stable/j.ctv7fcwb9, [2024-02-01].

[7] R. Liu, J. Zhang, G. Gao, and H. Li, "Betray Oneself: A Novel Audio DeepFake Detection Model via Mono-to-Stereo Conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 3999–4003.

[8] Z. Qin, W. Zhao, X. Yu, and X. Sun, "Openvoice: Versatile instant voice cloning," *arXiv preprint arXiv:2312.01479*, 2023.

[9] J. P. Cłapa, "Whisperspeech," 2023, available: https://github.com/colab/WhisperSpeech, [Accessed: 12-Jan-2024].

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020, pp. 6840–6851.

[11] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[12] A. Cohen-Hadria, M. Cartwright, B. McFee, and J. P. Bello, "Voice anonymization in urban sound recordings," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2019, pp. 1–6.

[13] D.-Y. Wu, Y.-H. Chen, and H. yi Lee, "VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net Architecture," in *Proc. INTERSPEECH 2020*, 2020, pp. 4691–4695.

[14] A. Agaskar, "Practical Over-the-air Perceptual AcousticWatermarking," in *Proc. INTERSPEECH 2022*, 2022, pp. 714–718.

[15] H. Chen, B. Darvish, and F. Koushanfar, "SpecMark: A Spectral Watermarking Framework for IP Protection of Speech Recognition Systems," in *Proc. INTERSPEECH 2020*, 2020, pp. 2312–2316.

[16] X. Zhang, Y. Xu, R. Li, J. Yu, W. Li, Z. Xu, and J. Zhang, "V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection," *arXiv preprint arXiv:2404.16824*, 2024.

[17] M. Campi, G. W. Peters, N. Azzaoui, and T. Matsui, "Machine learning mitigants for speech based cyber risk," *IEEE Access*, vol. 9, pp. 136 831–136 860, 2021.

[18] Y. Guo and A. Tyagi, "Voice-based user-device physical unclonable functions for mobile device authentication," *Journal of Hardware and Systems Security*, vol. 1, pp. 18–37, 2017.

[19] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *Proc. INTERSPEECH 2023*, 2023, pp. 4009–4013.

[20] S. H. Mun, H. jin Shim, H. Tak, X. Wang, X. Liu, M. Sahidullah, M. Jeong, M. H. Han, M. Todisco, K. A. Lee, J. Yamagishi, N. Evans, T. Kinnunen, N. S. Kim, and J. weon Jung, "Towards Single Integrated Spoofing-aware Speaker Verification Embeddings," in *Proc. INTERSPEECH 2023*, 2023, pp. 3989–3993.

[21] X. Wang, B. Zeng, H. Suo, Y. Wan, and M. Li, "Robust audio anti-spoofing countermeasure with joint training of front-end and back-end models," in *Proc. INTERSPEECH*, 2023, pp. 4004–4008.

[22] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[23] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 11 2007.

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[25] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[26] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.

[27] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[30] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans *et al.*, "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ASVspoof 2021 Workshop-Automatic Speaker Verification and Spoofing Coutermeasures Challenge*, 2021.

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. [Online]. Available: https://proceedings.mlr.press/v9/glorot10a.html

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.