Challenges in Protection against Deepfakes in Speech

Priyanshi Pal[1]*, Lauri Juvela*[2]*, Isabel Trancoso*[1]*, Alberto Abad*[1]

[1]INESC-ID, Instituto Superior Técnico, Portugal
[2]Department of Information and Communications Engineering, Aalto University, Finland
{priyanshipal53, isabel.trancoso, alberto.abad}@inesc-id.pt, lauri.juvela@aalto.fi

Abstract

Recent advances in Voice Cloning and Text-to-Speech (TTS) have enabled high-fidelity voice synthesis, making zero-shot adaptation feasible. These developments improve accessibility through personalized speech synthesis, cross-lingual voice transfer, and scalable applications such as automated dubbing. However, they also introduce risks, as realistic deepfake audio can be used for impersonation, misinformation, reputational harm, and financial fraud. In non-adversarial settings, transparency is ensured via watermarking or metadata tagging, with regulations such as Article 50 of the European Union's AI Act requiring disclosure and machine-readable labeling [1]. However, in adversarial scenarios, detection methods offer limited effectiveness due to the diminishing gap between synthetic and real speech and their dependence on continual adaptation to advances in synthesis systems. This has led to increased focus on proactive defenses that modify audio at the source by embedding perturbations to support detection and to degrade the performance of voice cloning systems [2].

Building on this, the consideration of two problem settings emerge: (i) improving detection of synthesized content in non-adversarial environments, and (ii) developing defensive strategies that can interfere with or reduce the effectiveness of voice cloning systems in adversarial use cases. Hence, the objective of this work is to examine the challenges associated with such proactive perturbation methods, while also providing an overview of the current research landscape and highlighting potential future directions. Prior review papers such as [3], [4], and [5] have advanced the understanding of voice cloning and detection methods. However, comparatively less attention has been given to proactive measures in existing surveys.

To facilitate a structured discussion, we adopt an adversarial framing. Voice cloning models and protection removal techniques are referred to as the *attacker*, while systems that inject proactive perturbations are termed the *defender*. The attacker seeks to replicate a target speaker's voice by generating high-fidelity speech characterized by strong speaker similarity, intelligibility, natural prosody, and data efficiency [3]. Additionally, attackers may aim to bypass automatic speaker verification (ASV) systems. The defender-side threat model is summarised in Table 1, detailing capabilities, objectives, robustness to transformations, and attacker access assumptions, including model knowledge and inference settings. In contrast to attacker, protective perturbations introduced by the defender aim to disrupt downstream synthesis by inducing (i) identity obfuscation (IO), either targeted (T), where speaker attributes are manipulated toward a chosen individual [6, 13, 26] or group [10, 16], or untargeted (UT) [17, 23, 24, 25], where identity information is broadly degraded, (ii) degradation in speech quality [17] and naturalness, (iii) reduced intelligibility [23], and (iv) decreased performance of speaker verification [24, 25]or identification systems [28]. Both attacker and defender operate under varying assumptions of data access and model knowledge, which directly constrain their capabilities. These settings are typically categorized as Black-Box, Grey-Box, and White-Box, corresponding to no, partial, or full knowledge of model internals, respectively. In Black-Box settings, attackers often rely on surrogate models (transfer-attack based approaches) [29], or in some cases, ensembles of encoders, leveraging the assumption that speaker representations are learned in a similar manner across models. Capability is further influenced by factors such as support for zero-shot (ZS), one-shot (OS) or fine-tuned (FT) voice conversion (VC)/TTS, and real-time (RT) constraints.

In devising protective perturbations, ensuring their robustness remains a key challenge. In particular, maintaining effectiveness under signal transformations encountered during transmission, such as resampling, compression, and noise, can be difficult, alongside robustness to adaptive removal attacks including neural compression and diffusion-based denoising. Although many methods can be resilient to standard transmission distortions, they often do not account for neural purification processes that can remove embedded perturbations. This is also observed in audio watermarking, where schemes frequently fail under modern neural processing pipelines [30, 31]. Another challenge arises from purification-based attacks. Recent diffusion-based and codec latent-space purification methods can remove perturbations while preserving speaker identity and perceptual quality. This makes it difficult to design defenses that remain effective under targeted purification, especially in adaptive or white-box settings. Taken together, these challenges suggest that designing perturbation or watermarking methods that remain detectable through cloning [32] and adaptive removal attacks [33, 34], while also preserving utility remains a non-trivial and relatively under-explored problem.

Lastly, the challenge of achieving cross-architectural transferability for protective perturbations remains a significant hurdle. Since improving cloning performance depends heavily on enhancing speaker representation [3], the diverse methods by which identity is encoded across various architectures pose a difficult problem for defenders. This is particularly evident in systems that do not rely on traditionally extractable speaker embeddings, such as those based on diffusion models or neural-codec language models. In these cases, the speaker's identity is often intertwined with the generative process itself rather than being isolated in a fixed vector. Furthermore, recent benchmarking has found that deepfakes produced by these two specific architectures, i.e., codec-based and diffusion-based, are among the hardest to detect [35], underscoring the need for protection mechanisms that generalize across fundamentally different synthesis pipelines.

Index Terms: speech privacy, watermarking, deepfakes, perturbation based protection methods

1. Acknowledgements

This work was funded by the European Union's Horizon Europe research and innovation programme grant No 101168193 (DOI: https://doi.org/10.3030/101168193) and by national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025).

2. References

[1] E. Parliament and C. of the European Union, "REGULATION (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008,

Table 1: *Taxonomy of proactive defense methods*

Defender	**Capability**	**Objective**	**Robustness**	**Attacker Access**		
	IO type	Disruption attribute	Transformations Description	Inference	Knowledge	Model Examples
VSMask [6]	T (+RT)	speaker embeddings	Denoise, Resamp, Quant, Filter-BP/RIR	ZS/OS-TTS	White-Box	AdaIN-VC [7], SV2TTS [8], AutoVC [9]
Voice Cloak [10]	T	speaker representation	MP3/AAC, Filter-LP, Gaussian Noise	ZS/FS-VC	Grey-Box	DiffVC (Diffusion) [11], DDDM-VC (Diffusion) [12]
Attack VC [13]	T	speaker embeddings	–	ZS-VC	Grey-Box	AutoVC, AdaIN-VC
Anti-Fake [14]	T, UT	speaker embeddings	–	ZS-TTS/VC	Black-Box	SV2TTS, YourTTS (VAE) [15]
Gao et al. [16]	T	speaker embeddings	–	ZS-TTS	Black-Box	QuickVC (VC), FreeVC (VC), TriAANVC (VC)
SafeSpeech [17]	UT (+RT)	Quality, Speaker Similarity	Denoise-DEMUCS, Resamp, Quant, MP3	FT/ZS-TTS	Black-Box	VITS (VAE) [18], StyleTTS 2 (GAN) [19], F5TTS (Flow) [20], FishSpeech (LLM) [21], GlowTTS (Flow) [22]
E2E-VGuard [23]	T, UT	Intelligibility, Timbre	Denoise-DNN, Resamp, Quant, MP3	FT/ZS-TTS	Black-Box	CosyVoice (LLM), F5-TTS (Flow), StyleTTS2 (GAN), VITS (VAE)
FreeTalk [24]	UT	ASV accuracy	Noise, Time-shift, Vol-scale, Smoothing	ZS-TTS/VC	Black-Box	Tacotron2 (RNN), Glow-TTS (Flow), YourTTS (VAE), Speedy-Speech (CNN)
VocalCrypt [25]	UT (+RT)	ASV accuracy	Denoise-std, MP3, Compression	ZS-TTS/VC	Black-Box	ElevenLabs (API), StyleTTS2 (GAN), XTTSv2 (LLM), SEED-VC (Flow Matching)
RoVo [26]	T (+RT)	speaker embeddings	Denoise-DeepFilter, MP-SENet, Spec-Mask	ZS-TTS/VC	White-Box	YourTTS (VAE), AdaptVC (Autoencoder), TorToise (Transformer) [27]

(EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)," Jun. 2024, https://artificialintelligenceact.eu/article/50/. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive Detection of Voice Cloning with Localized Watermarking," Jun. 2024, arXiv:2401.17264 [cs]. [Online]. Available: http://arxiv.org/abs/2401.17264

[3] H. Azzuni and A. E. Saddik, "Voice Cloning: Comprehensive Survey," May 2025, arXiv:2505.00579 [cs]. [Online]. Available: http://arxiv.org/abs/2505.00579

[4] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A Survey on Speech Deepfake Detection," *ACM Computing Surveys*, vol. 57, no. 7, pp. 1–38, Jul. 2025, arXiv:2404.13914 [cs]. [Online]. Available: http://arxiv.org/abs/2404.13914

[5] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," *Sensors*, vol. 25, no. 7, p. 1989, Mar. 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/7/1989

[6] Y. Wang, H. Guo, G. Wang, B. Chen, and Q. Yan, "VSMask: Defending Against Voice Synthesis Attack via Real-Time Predictive Perturbation," in *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, May 2023, pp. 239–250, arXiv:2305.05736 [cs]. [Online]. Available: http://arxiv.org/abs/2305.05736

[7] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," Aug. 2019, arXiv:1904.05742 [cs]. [Online]. Available: http://arxiv.org/abs/1904.05742

[8] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," Jan. 2019, arXiv:1806.04558 [cs]. [Online]. Available: http://arxiv.org/abs/1806.04558

[9] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5210–5219. [Online]. Available: https://proceedings.mlr.press/v97/qian19c.html

[10] Q. Hu, J. Wu, W. Lu, and X. Luo, "VoiceCloak: A Multi-Dimensional Defense Framework against Unauthorized Diffusion-based Voice Cloning," Dec. 2025, arXiv:2505.12332 [cs]. [Online]. Available: http://arxiv.org/abs/2505.12332

[11] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme," Aug. 2022, arXiv:2109.13821 [cs]. [Online]. Available: http://arxiv.org/abs/2109.13821

[12] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "DDDM-VC: Decoupled Denoising Diffusion Models with Disentangled Representation and Prior Mixup for Verified Robust Voice Conversion," May 2023, arXiv:2305.15816 [eess]. [Online]. Available: http://arxiv.org/abs/2305.15816

[13] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending Your Voice: Adversarial Attack on Voice Conversion," 2020, version Number: 3. [Online]. Available: https://arxiv.org/abs/2005.08781

[14] Z. Yu, S. Zhai, and N. Zhang, "AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. Copenhagen Denmark: ACM, Nov. 2023, pp. 460–474. [Online]. Available: https://dl.acm.org/doi/10.1145/3576915.3623209

[15] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 2709–2720. [Online]. Available: https://proceedings.mlr.press/v162/casanova22a.html

[16] J. Gao, H. Li, Z. Zhang, and Z. Wu, "Black-Box Adversarial Defense Against Voice Conversion Using Latent Space Perturbation," in *ICASSP 2025 - 2025 IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad, India: IEEE, Apr. 2025, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/10890469/

[17] Z. Zhang, D. Wang, Q. Yang, P. Huang, J. Pu, Y. Cao, K. Ye, J. Hao, and Y. Yang, "SafeSpeech: Robust and Universal Voice Protection Against Malicious Speech Synthesis," Apr. 2025, arXiv:2504.09839 [cs]. [Online]. Available: http://arxiv.org/abs/2504.09839

[18] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," Jun. 2021, arXiv:2106.06103 [cs]. [Online]. Available: http://arxiv.org/abs/2106.06103

[19] Y. A. Li, C. Han, V. S. Raghavan, G. Mischler, and N. Mesgarani, "StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models," Nov. 2023, arXiv:2306.07691 [eess]. [Online]. Available: http://arxiv.org/abs/2306.07691

[20] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. JianZhao, K. Yu, and X. Chen, "F5-TTS: A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 6255–6271. [Online]. Available: https://aclanthology.org/2025.acl-long.313

[21] S. Liao, Y. Wang, T. Li, Y. Cheng, R. Zhang, R. Zhou, and Y. Xing, "Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis," Nov. 2024, arXiv:2411.01156 [cs]. [Online]. Available: http://arxiv.org/abs/2411.01156

[22] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: a generative flow for text-to-speech via monotonic alignment search," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Curran Associates Inc., 2020.

[23] Z. Zhang, D. Wang, Y. Mi, Z. Wu, J. Gao, Y. Cao, K. Ye, M. Xue, and J. Hao, "E2E-VGuard: Adversarial Prevention for Production LLM-based End-To-End Speech Synthesis," Nov. 2025, arXiv:2511.07099 [cs]. [Online]. Available: http://arxiv.org/abs/2511.07099

[24] Y. Pu, Z. Feng, C. Zhou, J. Chen, C. Hu, H. Hu, and S. Ji, "FreeTalk:A plug-and-play and black-box defense against speech synthesis attacks," Aug. 2025, arXiv:2509.00561 [cs]. [Online]. Available: http://arxiv.org/abs/2509.00561

[25] Q. Fei, W. Hou, X. Hai, and X. Liu, "VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect," Feb. 2025, arXiv:2502.10329 [cs]. [Online]. Available: http://arxiv.org/abs/2502.10329

[26] S. Kim, S. Park, D. Kim, J. Lee, and D. Choi, "RoVo: Robust Voice Protection Against Unauthorized Speech Synthesis with Embedding-Level Perturbations," 2025, version Number: 1. [Online]. Available: https://arxiv.org/abs/2505.12686

[27] J. Betker, "Better speech synthesis through scaling," May 2023, arXiv:2305.07243 [cs]. [Online]. Available: http://arxiv.org/abs/2305.07243

[28] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso, "FoolHD: Fooling Speaker Identification by Highly Imperceptible Adversarial Disturbances," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Toronto, ON, Canada: IEEE, Jun. 2021, pp. 6159–6163. [Online]. Available: https://ieeexplore.ieee.org/document/9413760/

[29] G. Wang, C. Zhou, Y. Wang, B. Chen, H. Guo, and Q. Yan, "Beyond Boundaries: A Comprehensive Survey of Transferable Attacks on AI Systems," May 2025, arXiv:2311.11796 [cs]. [Online]. Available: http://arxiv.org/abs/2311.11796

[30] Y. Özer, W. Choi, J. Serrà, M. K. Singh, W.-H. Liao, and Y. Mitsufuji, "A Comprehensive Real-World Assessment of Audio Watermarking Algorithms: Will They Survive Neural Codecs?" in *Interspeech 2025*. ISCA, Aug. 2025, pp. 5113–5117. [Online]. Available: https://www.isca-archive.org/interspeech_2025/ozer25_interspeech.html

[31] P. O'Reilly, Z. Jin, J. Su, and B. Pardo, "Deep Audio Watermarks are Shallow: Limitations of Post-Hoc Watermarking Techniques for Speech," Apr. 2025, arXiv:2504.10782 [cs]. [Online]. Available: http://arxiv.org/abs/2504.10782

[32] Y. Özer, W. Ge, Z. Zhang, X. Wang, and J. Yamagishi, "Self Voice Conversion as an Attack against Neural Audio Watermarking," Jan. 2026, arXiv:2601.20432 [cs]. [Online]. Available: http://arxiv.org/abs/2601.20432

[33] M. Abbasihafshejani, A. N. Sakib, and M. Jadliwala, "VocalBridge: Latent Diffusion-Bridge Purification for Defeating Perturbation-Based Voiceprint Defenses," Jan. 2026, arXiv:2601.02444 [cs]. [Online]. Available: http://arxiv.org/abs/2601.02444

[34] W. Fan, K. Chen, C. Liu, W. Zhang, and N. Yu, "De-AntiFake: Rethinking the Protective Perturbations Against Voice Cloning Attacks," Jul. 2025, arXiv:2507.02606 [cs]. [Online]. Available: http://arxiv.org/abs/2507.02606

[35] X. Mao, K. Li, C. Baird, E. X. Tao, and D. Lin, "Benchmarking Fake Voice Detection in the Fake Voice Generation Arms Race," Oct. 2025, arXiv:2510.06544 [cs]. [Online]. Available: http://arxiv.org/abs/2510.06544