NaVo: Natural Voice Protection against Voice Cloning Attacks via Generative Universal Adversarial Audio

Seoyoung Park *, *Seungmin Kim* *, *Sohee Park*, *Dain Kim*, *Thien An Nguyen*, *Thien-Phuc Doan*, *Souhwan Jung* **, *Daeseon Choi* **

Soongsil University, Seoul, Republic of Korea

{seozxro, kimsm0802, sosohi, dikim420, anthiennng1995}@soongsil.ac.kr,
{phucdt, souhwanj, sunchoi}@ssu.ac.kr

Abstract

Proactive defense against voice cloning disrupts synthesis by injecting adversarial perturbations. However, prior methods rely on iterative optimization, often causing perceptual distortion and high latency that limit real-time deployment. We propose **NaVo**, a generative framework that produces natural-sounding universal adversarial audio. NaVo guides a latent diffusion–based text-to-audio model with fine-tuned low-rank adaptation modules tailored to gender and acoustic categories such as rain, music, and babble, enabling high-fidelity, speaker-independent adversarial audio generation. Unlike optimization-based approaches, NaVo supports real-time protection through simple audio mixing without gradient-based inference. It achieves a 76% defense success rate against a widely used commercial voice cloning system in a black-box setting. By jointly addressing universality, audio quality, and computational efficiency, NaVo offers a practical solution for proactive voice protection.

Index Terms: Proactive Defense, Voice Cloning Attack, Universal Adversarial Audio (UAA)

1. Introduction

The rapid advancement of deep learning-based speech synthesis has reached a point where voice cloning can replicate a target speaker's voice with high fidelity using only a few seconds of audio. While these technological advancements offer significant benefits, they have also introduced grave security risks, ranging from the generation of fake news using synthesized public figures to identity fraud via impersonation of private individuals [1, 2]. Currently, defenses against voice cloning are broadly categorized into *post-detection* [3] and *proactive defense* [4]. Post-detection methods attempt to detect whether a given audio is synthetically generated; however, they often suffer from poor generalization to unseen speech and, more fundamentally, can only react after the cloned audio has already been circulated, which is a significant limitation. Proactive defense [5–7], in contrast, injects subtle perturbations into the original speech signal to prevent the speaker encoders of voice cloning models from accurately extracting unique characteristics, thereby neutralizing cloning attempts at the source.

While existing proactive defenses offer strong security performance, they suffer from two limitations. First, the injected perturbations often manifest as high-frequency noise or waveform distortions, making the speech perceptually unnatural. This degradation in audio quality limits the practical use of protected speech in real-world applications, such as social media. Second, prior methods rely on iterative gradient-based optimization for each individual sample. This per-sample optimization causes prohibitive computational overhead, rendering these approaches unscalable for large-scale processing or real-time scenarios.

To overcome these limitations, we propose **NaVo** (**Na**tural **Vo**ice Protection) [8], a generative proactive defense framework that synthesizes natural ambient audio as Universal Adversarial Audio (UAA). The central idea is to leverage semantically meaningful acoustic environments, rather than imperceptible noise, to disrupt speaker identity extraction while maintaining listening quality. Concretely, we adapt a text-conditioned audio generation model to produce defense-oriented ambient sounds that can be seamlessly mixed with the original speech.

Crucially, NaVo is entirely speaker-independent; once trained with our proposed α-divergence-based distributional target loss, it provides immediate protection for unseen speakers without any additional retraining or per-sample optimization. Consequently, NaVo generates UAA through a single forward pass, ensuring real-time applicability while achieving both defensive effectiveness and perceptual naturalness by simply mixing the generated audio with the speaker's utterance.

To the best of our knowledge, this work is the first to employ natural acoustic environments as adversarial perturbations and provide speaker-independent protection in the audio domain, thereby resolving the fundamental trade-off between defense effectiveness and perceptual naturalness in proactive voice defense. Through extensive evaluations on state-of-the-art Text-to-Speech (TTS) models under white-box, black-box, and adaptive attack scenarios, we demonstrate the practical utility and scalability of NaVo, achieving significant defensive gains across all settings.

2. Proposed Method: NaVo

2.1. System Overview

NaVo adopts AudioLDM2 [9] as its backbone to generate natural-sounding Universal Adversarial Audio. As a latent diffusion-based text-to-audio model, AudioLDM2 synthesizes high-quality, semantically aligned audio from text prompts, enabling controllable generation across diverse acoustic styles. To repurpose this generation capability for voice cloning defense, NaVo fine-tunes the model's UNet via LoRA. The overall framework is illustrated in Figure 1. The system takes a text prompt p and the original speech x as inputs, where the final protected audio x_{prot} is defined as:

$$x_{prot} = x + \mathcal{G}(z|p; \Theta + \Delta\Theta) \tag{1}$$

where $\mathcal{G}$ denotes the generative model with pre-trained parameters Θ and the low-rank update weights $\Delta\Theta$ learned via LoRA. Unlike conventional proactive defenses that directly optimize adversarial perturbations in the waveform domain, NaVo

*These authors contributed equally and are listed alphabetically.
**indicates the corresponding author.

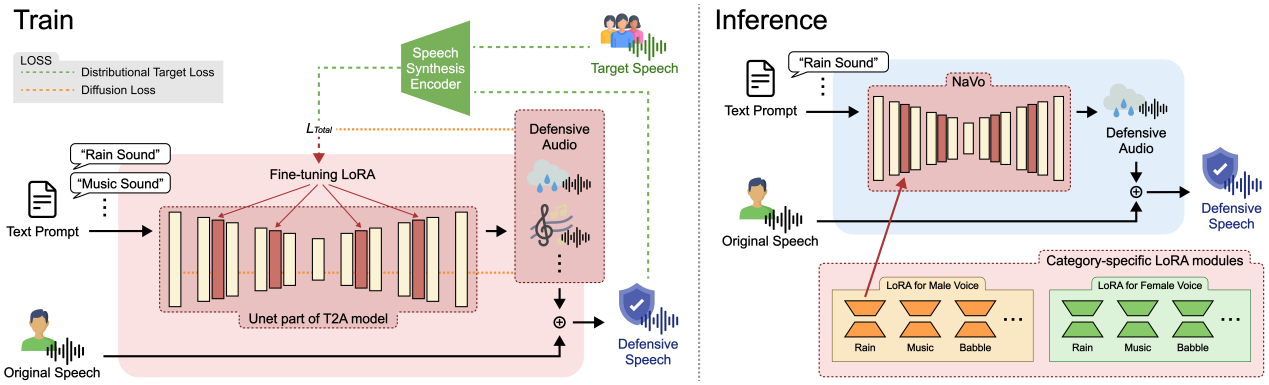

Figure 1: ***Overview of the NaVo framework.*** *(Left) During training, the base text-to-audio (T2A) model is fine-tuned via LoRA using the proposed distributional target loss. (Right) During inference, an appropriate category-specific LoRA module is applied to generate UAA, which is then simply mixed with the original speech to construct the final protected audio.*

generates semantically meaningful ambient sounds and mixes them with the original speech. This generative approach ensures robust identity protection while preserving the perceptual naturalness of the audio.

2.2. Modular LoRA Adaptation for Universal Defense

NaVo aims to generate Universal Adversarial Audio that generalizes across diverse speakers and acoustic categories. In practice, however, training a single model to simultaneously account for all speaker variations and diverse ambient sounds while maintaining optimal adversarial effectiveness is highly challenging from an optimization perspective. To address this limitation, NaVo constructs multiple LoRA [10] modules, each specifically optimized for distinct conditions.

Conventional proactive defenses typically optimize perturbations to shift the protected speech's embedding toward a specific target speaker. In contrast, NaVo employs a distribution-based approach to achieve a universal defense. We identified gender as the primary target attribute because it exhibits clear acoustic distinctness and, more importantly, forms distinct clusters within the speaker embedding space of voice cloning models. As demonstrated by the t-SNE visualization in Figure 2, the speaker encoder effectively separates speakers by gender, validating the use of gender-specific distributions as stable adversarial targets. Specifically, NaVo establishes male and female target distributions and trains individual LoRA modules for various acoustic categories (e.g., rain, babble) within each gender. This modular design ensures universal effectiveness, allowing the defense to protect unseen speakers without any additional optimization.

Architecturally, NaVo applies LoRA exclusively to the cross-attention layers within the AudioLDM2 UNet. Specifically, low-rank adapters are added to the key (W_K) and value (W_V) projection matrices in each transformer block:

$$W_K' = W_K + B_K A_K, \quad W_V' = W_V + B_V A_V \tag{2}$$

where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times d}$, and rank $r \ll d$. During training, only these LoRA parameters are optimized, while all other components—including the VAE, vocoder, and text encoders—remain frozen. This selective adaptation leverages the fact that the K and V projections in cross-attention are pivotal for controlling acoustic characteristics based on text conditions. Consequently, it achieves maximum adversarial effectiveness while preserving the model's underlying generative prior.

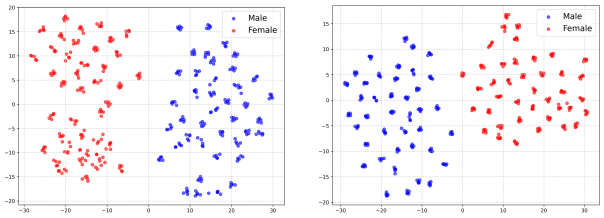

Figure 2: ***t-SNE visualization of speaker embeddings.*** *The embeddings are extracted from speaker encoders, showing clear clustering by gender (male: blue, female: red).*

2.3. Optimization Objective

NaVo is trained by jointly optimizing two primary loss functions: a **Diffusion Loss** [11], which preserves the perceptual naturalness and acoustic quality of the generated audio, and a **Distributional Target Loss**, which drives the generated audio to be highly adversarial against speaker encoders.

Diffusion Loss. To ensure the LoRA-adapted UNet generates category-consistent natural ambient sounds, we employ the denoising diffusion objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|^2\right] \tag{3}$$

where z_0 is the clean latent representation encoded from the reference ambient audio, $\epsilon \sim \mathcal{N}(0, I)$ is the added noise, t is the diffusion timestep, c is the text conditioning, and $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$. This loss objective prevents the adversarial optimization process from degrading the audio quality, ensuring that NaVo maintains high-fidelity generative performance.

Distributional Target Loss via α-Divergence. At each training step, the ambient sound $\hat{a}$ estimated from the UNet's noise prediction is mixed with a source speech s at a prescribed SNR:

$$x_{\text{mix}} = s + \gamma \cdot \hat{a}, \quad \gamma = \frac{\text{RMS}(s)}{\text{RMS}(\hat{a})} \cdot 10^{-\text{SNR}/17} \tag{4}$$

where γ is a scaling factor adjusting the ambient sound's amplitude. The SNR is fixed at 17 dB across all training scenarios to guarantee the clear intelligibility of the original speech while the ambient sound acts as a natural background [12]. This mixing process is fully differentiable, enabling gradients from the target loss to propagate back through the speaker encoder to the LoRA parameters in an end-to-end manner.

The mixed audio x_{mix} is fed into a pre-trained, frozen speaker encoder $f(\cdot)$ to extract a speaker embedding $e = f(x_{\text{mix}})$. Since the encoder parameters are fixed, any shift in the embedding space is exclusively driven by the generated ambient audio. Our objective is to guide this embedding toward a pre-defined target distribution P, effectively concealing the original speaker's identity. As shown in Figure 2, the speaker embedding space exhibits clear gender-based clusters. Leveraging this acoustic distinctness, NaVo defines the target distribution $P = \mathcal{N}(\mu_P, \text{diag}(\sigma_P^2))$ by shifting the speaker's identity toward the opposite gender's distribution. Specifically, to protect male speakers, we set the target as the female embedding distribution, and vice versa. Before training, we construct P by extracting embeddings from a diverse set of speakers of the target gender $\mathcal{X}_{\text{target}}$ and computing their statistical properties as follows:

$$\begin{aligned} \mu_P &= \frac{1}{|\mathcal{X}_{\text{target}}|} \sum_{x \in \mathcal{X}_{\text{target}}} f(x), \\ \sigma_P^2 &= \frac{1}{|\mathcal{X}_{\text{target}}|} \sum_{x \in \mathcal{X}_{\text{target}}} (f(x) - \mu_P)^2 \end{aligned} \tag{5}$$

In this context, μ_P and σ_P^2 denote the centroid and variance of the target embeddings, respectively. The most intuitive approach for achieving speaker defense would be to minimize the ℓ_2 distance between the embedding and the target centroid, i.e., $\|e - \mu_P\|^2$. However, this only pulls the embedding toward a single point and ignores the variance of the target distribution. Since target speakers naturally span a region in the embedding space, using a point estimate as the target is an oversimplification.

To address this, NaVo adopts a distributional approach. Specifically, for each training batch, we model the distribution of extracted embeddings from the source gender group as a diagonal Gaussian $Q = \mathcal{N}(\mu_Q, \text{diag}(\sigma_Q^2))$, where μ_Q and σ_Q are the empirical mean and variance of the defended embeddings within the batch. To minimize the α-divergence between Q and the target distribution P, we employ the Bhattacharyya distance, a closed-form instance of the α-divergence for $\alpha = 0.5$:

$$D_B(Q\|P) = \frac{1}{d} \sum_{j=1}^{d} \left[\frac{1}{2} \ln \frac{\sigma_{P,j}^2 + \sigma_{Q,j}^2}{2\,\sigma_{P,j}\,\sigma_{Q,j}} + \frac{(\mu_{P,j} - \mu_{Q,j})^2}{4(\sigma_{P,j}^2 + \sigma_{Q,j}^2)} \right] \tag{6}$$

where d is the embedding dimension and j indexes each dimension. The first term encourages the variances of Q and P to match, while the second term aligns their means. Compared to the ℓ_2 approach, this formulation jointly optimizes both central tendency and dispersion, promoting a structured and robust shift of the embedding distribution toward the target.

Total Loss. The total loss combines $\mathcal{L}_{\text{diff}}$ for maintaining acoustic naturalness with D_B for speaker protection. To ensure optimization stability, the adversarial embedding loss is applied exclusively during low-noise stages ($t < t_{\text{max}}$), where the denoised estimates are most reliable:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{diff}} \mathcal{L}_{\text{diff}} + \lambda_{\text{emb}} \cdot \mathbb{I}[t < t_{\text{max}}] \cdot D_B(Q\|P) \tag{7}$$

2.4. Inference-Time Defense

During inference, NaVo enables instant defense by selecting a LoRA module corresponding to the user's gender and target acoustic category. This modular design allows for environment-specific protection without further optimization. The resulting adversarial audio ensures robust generalization to unseen speakers and utterances while preserving the high-fidelity acoustic quality of AudioLDM2.

Table 1: *Overall DSR performance: NaVo vs. Enkidu.*

	DSR (%)↑		
	Resemblyzer	ECAPA-TDNN	ResNet
Enkidu	25.2	60.6	45.5
NaVo	**78.0**	**83.6**	**79.8**

3. Experiment

3.1. Experimental Setup

Voice Cloning Models. To evaluate defense performance, we adopt four models with strong zero-shot capabilities. We conduct white-box experiments using SV2TTS [13] and CosyVoice [14], and black-box experiments using Tortoise [15] and a commercial API, ElevenLabs [16]. This setup ensures NaVo's robustness across both open-source and real-world commercial scenarios.
Dataset. To cover diverse speakers and domains, we utilized multiple datasets widely adopted in voice cloning studies [17]: VCTK [18], FST [19], MCV [20], CSNED [21], CSUKIED [22], and LibriSpeech [23]. The data was partitioned into train, validation, and test sets with 196, 42, and 42 speakers, respectively. With 10 utterances per speaker, we employed 1,960, 420, and 420 samples per split. The target distribution P was constructed from non-overlapping data to ensure rigorous evaluation. For high-quality generation, we extracted 100 samples per category from AudioSet [24]. NaVo's robustness against unseen datasets was evaluated using the test set.
Evaluation Metrics. The proposed method was evaluated using Defense Success Rate (DSR) for defense performance and CLAP score [25] for acoustic quality. DSR is defined as the ratio of synthesized utterances that successfully achieve defense—by failing speaker verification—to the total number of defense-processed samples. A defense is successful if the similarity score falls below a pre-defined threshold τ, measured across three SV models: ECAPA-TDNN [26], Resemblyzer [27], and ResNet [28]. The CLAP score ensures high-fidelity naturalness by calculating the cosine similarity between text prompts and generated ambient sounds. Since NaVo protects speech by mixing it at a low SNR of 17dB, the resulting audio remains perceptually natural; however, we prioritized the CLAP score over DNN-based quality metrics (e.g., MOS), as such models inherently penalize any mixed audio regardless of actual naturalness. Higher values in both metrics signify superior defense effectiveness and generative fidelity, respectively, as further evidenced by our demo samples.
Baseline methods. To objectively evaluate NaVo, we adopt Enkidu [29] as our baseline, given its comparable real-time proactive defense scenario.

3.2. Experimental Result

Overall Performance. Evaluated on unseen datasets, NaVo demonstrated superior defense effectiveness and generalization. As shown in Table 1, NaVo achieved at least a 20% DSR improvement over the baseline Enkidu across all SV models, notably providing a threefold increase against Resemblyzer. These results confirm NaVo's ability to maintain high defense success rates even against unfamiliar voice characteristics.

White-box & Black-box Evaluation. As summarized in Table 2, NaVo demonstrated consistently high defense performance across all acoustic styles in white-box scenarios. Specifically, we utilized the GE2E-based encoder [30] from the SV2TTS frame-

Table 2: *White-box evaluation results across genders and acoustic styles. ↑ indicates higher values are better. "No UAA" and "Default" denote unprotected speech and audio from the backbone model without LoRA defense, respectively.*

Gender	Style	CLAPScore ↑			DSR (%) ↑					
					Resemblyzer		ECAPA-TDNN		ResNet	
		Default	GE2E	CAM++	SV2TTS	CosyVoice	SV2TTS	CosyVoice	SV2TTS	CosyVoice
Female	No UAA	–	–	–	3.7	1.0	22.1	2.9	23.0	0.5
	Raindrop	0.3003	**0.3604**	0.2248	88.4	**57.8**	**89.7**	**57.4**	**95.1**	**51.1**
	Babble	0.3604	0.3453	0.2099	68.6	55.1	66.4	52.1	79.2	49.3
	Office	0.3750	0.2057	**0.2362**	**89.5**	50.4	89.2	40.1	93.0	41.2
Male	No UAA	–	–	–	1.1	0.2	24.9	1.3	19.3	0.0
	Raindrop	0.3003	0.2327	0.2484	69.8	**58.3**	89.4	**62.1**	93.8	**51.1**
	Babble	0.3604	0.1300	0.3120	56.4	41.3	74.5	55.7	81.1	50.4
	Office	0.3750	**0.2874**	**0.3512**	**93.7**	52.7	**97.6**	43.6	**97.9**	39.7

Table 3: *Black-box evaluation results across genders and acoustic styles. ↑ indicates higher values are better. "No UAA" and "Default" denote unprotected speech and audio from the backbone model without LoRA defense, respectively.*

Gender	Style	CLAPScore ↑		DSR (%) ↑					
				Resemblyzer		ECAPA-TDNN		ResNet	
		Default	Ensemble based	Tortoise	Elevenlabs	Tortoise	Elevenlabs	Tortoise	Elevenlabs
Female	No UAA	–	–	0.5	0.0	11.7	0.0	1.3	0.5
	Raindrop	0.3003	0.1426	**91.1**	**97.1**	**93.5**	**95.2**	**91.0**	**96.2**
	Babble	0.3604	0.3362	88.9	94.8	91.1	90.0	90.3	80.5
	Music	0.3750	0.4466	56.5	83.8	77.5	49.5	67.5	43.8
Male	No UAA	–	–	3.3	0.0	13.6	0.0	5.4	0.0
	Raindrop	0.3003	0.2484	77.1	**92.9**	**94.4**	**95.2**	**89.5**	**97.1**
	Babble	0.3604	0.1568	**80.8**	84.3	84.8	81.9	78.9	67.6
	Music	0.4642	0.4403	75.4	42.9	60.3	33.3	61.9	32.9

work and the CAM++ encoder [31] from CosyVoice as target speaker encoders. While the unprotected baseline (No UAA) was completely vulnerable to the state-of-the-art CosyVoice model with near-zero DSR, NaVo effectively elevated the DSR to approximately 40-50%, proving its efficacy against advanced zero-shot synthesis. For the black-box evaluation in Table 3, NaVo was trained using an ensemble of white-box encoders to enhance transferability to unseen architectures. This approach achieved remarkable success against entirely different models, including Tortoise and the commercial ElevenLabs engine. Notably, NaVo reached over 90% DSR in the Raindrop style against ElevenLabs, verifying its robust applicability in real-world environments. Crucially, the CLAP scores remained comparable to the original backbone model across all configurations, confirming that NaVo preserves high-fidelity acoustic naturalness while maintaining its defensive strength.

Robustness to Adaptive Attacks. To evaluate practical robustness, we simulated adaptive attacks using WaveGuard [32], which applies signal processing and DNN-based speech enhancement [33] to remove the defense. Despite these attempts, NaVo's defense remained effective, with DSR staying above 80% in most cases. Furthermore, we observed that applying these purification techniques rather led to a noticeable degradation in speech quality. Since the UAA is not easily separated from the speech, the purification process further hindered the extraction of speaker features by distorting essential acoustic components. This confirms that NaVo provides sturdy protection even under active countermeasures, as the attack ultimately makes it more difficult to capture distinct speaker characteristics.

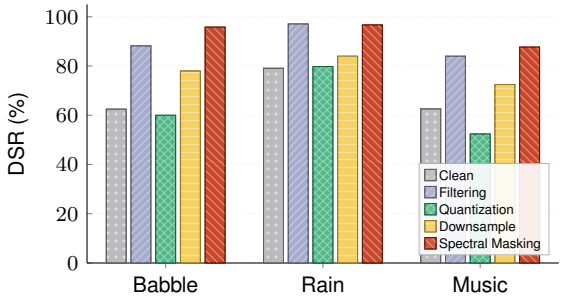

Figure 3: *Robustness to Adaptive Attacks: DSR (%) performance of NaVo under various audio purification.*

4. Conclusion

This paper proposed NaVo, a universal proactive defense against voice cloning. By integrating a distribution-based objective with modular LoRA adaptation, NaVo generates high-fidelity Universal Adversarial Audio (UAA) that maintains semantic alignment with text prompts. Unlike optimization-based methods, NaVo enables real-time protection via a single forward pass and simple mixing, while ensuring superior effectiveness. Extensive evaluations confirm consistently high Defense Success Rates (DSR) across diverse SV models, acoustic styles, and both white-box and black-box scenarios. Notably, NaVo exhibits strong robustness against adaptive purification attacks, verifying its practical utility. By leveraging natural ambient sounds as structured semantic perturbations, NaVo establishes a new generative paradigm for robust and perceptually natural voice protection.

5. Generative AI Use Disclosure

The core ideas, experimental designs, analysis, and scientific interpretations in this manuscript were conducted entirely by the authors. During the drafting process, generative AI tools, including ChatGPT and Gemini, were utilized solely for the purposes of language refinement, grammatical correction, and ensuring terminological consistency. Additionally, these tools were employed to assist with minor code implementation and debugging during the experimental phase. All final content and research integrity remain the full responsibility of the authors.

6. Acknowledgment

This work was supported by a Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No. RS-2026-25530576, Development of Robust Real-time Deepfake Disruption Technology via Adversarial Noise Injection for Privacy Protection) and an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02215142, Development of Pseudonymization Technology for Suspected Criminal Information).

7. References

[1] B. AI, "Deepfake ceo fraud: $50m voice cloning threat cfos," *Brightside AI Blog*, Oct 2025. [Online]. Available: https://www.brside.com/blog/deepfake-ceo-fraud-50m-voice-cloning-threat-cfos

[2] T. Ramsey, "Scammers use ai voice cloning to make fraudulent payments," *Which?*, Feb 2026. [Online]. Available: https://www.which.co.uk/news/article/beware-of-survey-phone-scams-a3SEH9I5fwuD

[3] Z. Yan, Y. Zhao, and H. Wang, "{VoiceWukong}: Benchmarking deepfake voice detection," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 4561–4580.

[4] H.-H. Nguyen-Le, V.-T. Tran, T. Nguyen, and N.-A. Le-Khac, "A survey on proactive deepfake defense: Disruption and watermarking," *ACM Computing Surveys*, vol. 58, no. 5, pp. 1–37, 2025.

[5] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 552–559.

[6] 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*, 2023, pp. 460–474.

[7] 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," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 4581–4600.

[8] "NaVo," Sound samples: https://anonymous.4open.science/w/NaVo/, 2026.

[9] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.

[11] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[12] International Organization for Standardization, "ISO 9921:2003: Ergonomics — assessment of speech communication," International Organization for Standardization, Geneva, Switzerland, Standard, 2003.

[13] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[14] Z. Du, Q. Chen, S. Zhang, K. Hu, H. Lu, Y. Yang, H. Hu, S. Zheng, Y. Gu, Z. Ma *et al.*, "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.

[15] J. Betker, "Better speech synthesis through scaling," *arXiv preprint arXiv:2305.07243*, 2023.

[16] ElevenLabs, "Elevenlabs: Prime voice AI," https://elevenlabs.io/, 2024, [Accessed: 26-February-2026].

[17] K. Wang, M. Chen, L. Lu, J. Feng, Q. Chen, Z. Ba, K. Ren, and C. Chen, "From one stolen utterance: Assessing the risks of voice cloning in the aigc era," in *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2025, pp. 4663–4681.

[18] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019, university of Edinburgh. The Centre for Speech Technology Research (CSTR).

[19] Surfing.AI, "ST-AEDS-20180100 1, Free ST American English Corpus," https://www.openslr.org/45/.

[20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marseille, France, 2020, pp. 4218–4222.

[21] Google, "Crowdsourced high-quality Nigerian English speech data set," https://openslr.elda.org/70/.

[22] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the English accents in the British Isles," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marseille, France, 2020, pp. 6532–6541.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, 2015, pp. 5206–5210.

[24] Google Research, "AudioSet: An ontology and human-labeled dataset for audio events," https://research.google.com/audioset/, 2017, [Accessed: 26-February-2026].

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[27] R. AI, "Resemblyzer: A toolkit for analyzing and comparing voices," 2019, accessed: 2025-01-10. [Online]. Available: https://github.com/resemble-ai/Resemblyzer

[28] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak *et al.*, "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.

[29] Z. Feng, J. Chen, C. Zhou, Y. Pu, Q. Li, T. Du, and S. Ji, "Enkidu: Universal frequential perturbation for real-time audio privacy protection against voice deepfakes," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 11 638–11 647.

[30] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[31] H. Wang, S. Zheng, Y. Chen, L. Cheng, and Q. Chen, "Cam++: A fast and efficient network for speaker verification using context-aware masking," *arXiv preprint arXiv:2303.00332*, 2023.

[32] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "WaveGuard: Adversarial Example Detection and Purification in Speech Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 425–16 434.

[33] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 7–19, 2014.