VoiceShield: Real-Time Speech Protection with Preserved Machine Recognizability[*]

Jeeeun Park, Soona Lee, Yeonhee Kim, Gyurim Kim, and Jongkil Kim[†]

Ewha Womans University, Seoul, Korea
{lucykorea414, soon22a, heedong, 2371075, jongkil}@ewha.ac.kr

Abstract

The rapid advancement of neural speech synthesis and voice cloning technologies has raised severe concerns about the misuse of speech data, leading to threats such as voice phishing, identity theft, and authentication bypass. Existing defenses—including post-hoc detection, watermarking, and noise injection—struggle to balance real-time protection with the preservation of machine recognition accuracy. To address this gap, we propose VoiceShield, a real-time speech protection framework that safeguards user voices while maintaining practical usability in speech-to-text (STT)-based applications. VoiceShield integrates three modules: BandMaskNet, a spectrum masking network that applies perceptually natural yet adversarial perturbations; Selective Distortion, which targets formant-related frequency bands critical for speaker recognition; and Prosody Modification, which perturbs speech rhythm while preserving timbre. Experimental results demonstrate that VoiceShield significantly reduces the effectiveness of voice cloning attacks, achieving full protection in Korean and over 92% robustness in English, while keeping STT accuracy within usable bounds. These findings highlight VoiceShield as a practical and efficient solution for securing speech data in real-time communication and service environments without sacrificing usability.

1 Introduction

The rapid advancement of Artificial Intelligence(AI)-based speech synthesis technologies has opened up innovative possibilities across a wide range of applications [35]. Representative positive examples include voice assistants, automatic caption generation, and accessibility support. However, the unauthorized collection and use of speech data can lead to severe security and privacy concerns [16]. In particular, in environments where speech is routinely recorded, such as call centers and customer support services, the accumulated data, if leaked, pose significant risks of being exploited for training speech synthesis or cloning models.

Such misuse can directly lead to social crimes such as voice phishing, and the emergence of sophisticated AI-driven voice forgery technologies (e.g., deep voice and voice cloning) has further amplified the scale of potential damage [9]. In practice, voice cloning technologies are capable of replicating not only the spoken content but also the unique timbre and speaking style of the speaker, making it easier for victims to mistake the cloned voice for that of acquaintances or trusted institutions. Consequently, this poses a serious threat across multiple domains, including financial transactions, identity theft, and authentication systems in corporate and governmental organizations.

A variety of countermeasures have been investigated in response to these threats. Conventional approaches to detecting speech forgeries have primarily relied on exploiting subtle spectral

[*] Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 11, December 16-18, 2025, Sapporo , Japan. © The copyright of this paper remains with the author(s).

[†] Corresponding Author

discrepancies or discontinuities in speech characteristics between synthetic and natural audio, or on employing deep learning-based classifiers to determine authenticity [13]. However, such methods are inherently post-hoc in nature and therefore struggle to provide immediate protection in real-time communication or streaming scenarios. Watermarking- and encryption-based protection techniques can mitigate the risk of unauthorized data reuse, but they often introduce perceptual degradation and, more importantly, impair machine recognition of speech, reducing their applicability in practical services [15] such as speech-to-text (STT) systems. Moreover, other recently proposed privacy preservation methods [22] have also focused primarily on protecting data during model training, and they are less effective in controlling real-time threats that arise at the user end. Therefore, existing methods face inherent challenges in protecting speech data while maintaining machine recognition. In particular, STT systems—serving as a foundation for diverse applications such as customer support, automatic caption generation, and accessibility services—become one of the most negatively affected services when protective mechanisms are applied, as they are highly sensitive to recognition accuracy.

To address these limitations, we introduce VoiceShield, a real-time speech protection framework that preserves automatic speech recognition performance while significantly degrading the effectiveness of voice cloning and speech synthesis attacks. Unlike prior approaches that primarily rely on post-hoc detection or disruptive perturbation, VoiceShield ensures both proactive defense and practical usability in real-world communication settings. We quantitatively evaluate both speech recognition accuracy and speaker verification robustness. Through these contributions, this study demonstrates that VoiceShield offers a practical means of safeguarding speech data while preserving its utility for legitimate applications.

2 Related Work

Speech synthesis technologies have evolved rapidly from early concatenative text-to-speech (TTS) and hidden Markov model (HMM)-based statistical TTS approaches to modern deep learning-based neural TTS systems. Representative works include the Tacotron series (Wang et al., 2017 [34]; Shen et al., 2018 [28]), Deep Voice (Arik et al., 2017 [2]), and FastSpeech (Ren et al., 2019 [26]). Furthermore, the advent of neural vocoders such as WaveNet (Oord et al., 2016 [30]) and HiFi-GAN (Kong et al., 2020 [14]) has greatly improved the naturalness of synthesized speech. More recently, research on zero-shot voice cloning has gained momentum, with systems such as SV2TTS (2018 [12]), AutoVC (Qian et al., 2019 [23]), YourTTS (Casanova et al., 2022 [3]), VALL-E (Wang et al., 2023 [32]), and OpenVoice (2024 [24]) enabling the reproduction of a new speaker's vocal characteristics from only a short speech sample. These technological advancements have significantly lowered the barriers to voice cloning, thereby accelerating the realization of tangible attack threats.

Research on countermeasures against speech synthesis and voice cloning attacks has generally progressed along three main directions. The first focuses on noise-injection-based defenses. For example, VSMask (2023 [4]) attempted to disrupt the training of synthesis models by inserting real-time predicted perturbations, but its high computational complexity limited applicability in real-time scenarios. CloneShield (2025 [37]) explored the use of universal perturbations to provide wide-coverage protection; however, when applied with high intensity, it was reported to cause noticeable speech quality degradation. VocalCrypt (2025 [7]) proposed the insertion of pseudo-timbre to prevent evasion by detection systems, though this approach entails the risk of distortion in practical communication environments.

The second line of research centers on detection-based approaches. For instance, AntiFake (2023 [10]) and VoiceBlock (2024 [29]) proposed classifier-based methods designed to detect synthetic speech. However, such approaches are inherently post-hoc, making it fundamentally difficult to prevent harm before it occurs.

The third direction involves watermarking and prosody modification techniques. For example, RoVo (2025 [19]) and SafeSpeech (2025 [36]) proposed approaches that incorporate protective signals at the speaker-embedding level or alter prosodic rhythms of speech. However, when deployed in real-time communication or streaming environments, these methods often lead to performance degradation, and their generalization capability across diverse languages remains limited.

Existing studies have primarily focused on post-hoc detection or limited forms of preventive perturbation, which impose fundamental constraints on achieving effective defense against synthetic speech attacks while preserving STT quality in real-time communication environments. These limitations underscore the necessity for new defense strategies. In this work, we propose a speech protection framework designed to overcome such constraints by ensuring both real-time operability and the preservation of STT accuracy.

3 Threat Models

Speech data are routinely collected and stored in various contexts, including call center interactions, customer service recordings, and online meetings. However, once such recordings are leaked, they can be exploited by AI-based voice cloning technologies, posing severe security threats. Recent studies have reported that speech synthesis is no longer limited to simple deepfake generation but can serve as a critical tool for social crimes such as financial fraud, voice phishing, and identity theft (Not My Voice!, 2024 [11]; Assessing the Risks of Voice Cloning, 2024 [33]). In particular, the theft of call center or counseling session recordings may expose individuals to sophisticated synthetic voice attacks without their awareness (Pitch, 2023 [20]). In this context, the need for techniques that move beyond post-hoc detection and instead provide real-time protection while preserving the quality of STT-based services is becoming increasingly evident.

3.1 Threat Models in Voice Cloning Attacks

The rapid advancement of voice cloning technologies poses a direct threat to voice security systems. First, synthetic speech can be exploited to bypass or disable speaker verification mechanisms, enabling attacks against authentication frameworks [8]. Second, zero-shot cloning techniques, which allow the generation of high-quality synthetic voices from only a small amount of speech data, significantly amplify the risks of data leakage [17]. Third, embedding-level manipulation, in which speaker embeddings are directly altered to forge identities, introduces novel attack vectors that challenge existing authentication infrastructures [18]. These attacks not only impose severe challenges to conventional security mechanisms but also disrupt the reliable operation of STT systems, thereby amplifying their potential impact. Consequently, the necessity of developing new defense mechanisms against such threats is becoming increasingly urgent.

4 VoiceShield

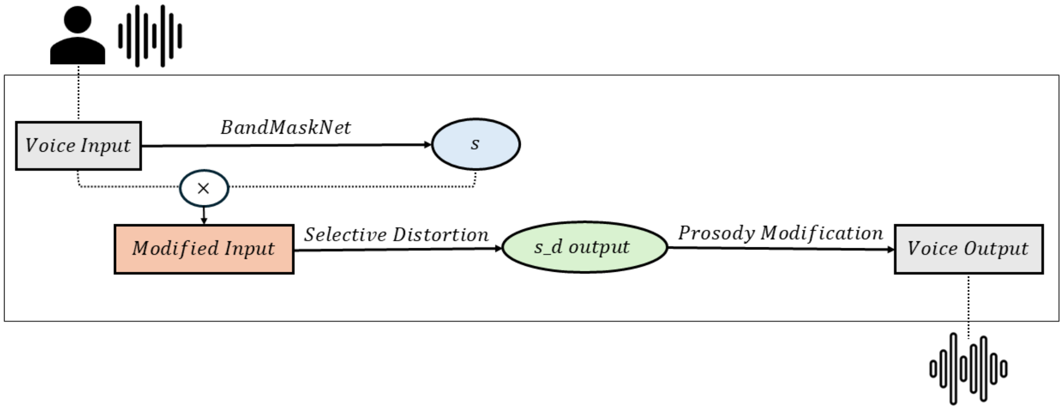

Figure 1: Overall system architecture of the proposed VoiceShield framework.

In this study, we propose VoiceShield, a real-time voice protection framework designed to prevent speech data from being exploited in the training of synthesis models. VoiceShield aims to preserve the accuracy of STT, while maintaining perceptual quality for human listeners, and at the same time disrupting the stable learning of speaker embeddings and voice cloning models. To meet these requirements, the proposed framework is composed of three core modules. The overall system pipeline is illustrated in Figure 1, where each module is sequentially integrated to ensure real-time operation. The input speech is first processed by BandMaskNet, which applies spectrum-based masking. Subsequently, the selective distortion module perturbs frequency bands that are critical for speaker recognition. Finally, in the prosody modification stage, the rhythm of speech is altered to destabilize prosodic features. The resulting modified speech is then delivered to STT engines or communication systems, making the framework suitable for real-time applications such as video conferencing and customer service systems.

4.1 BandMaskNet: Proposed Spectrum Masking Module

In this work, we propose BandMaskNet, a spectrum-based masking module specifically designed for real-time streaming environments. BandMaskNet leverages the time–frequency representation of speech to selectively amplify or attenuate certain regions, thereby disrupting the training process of synthesis models. Unlike conventional noise injection methods, BandMaskNet applies learned perturbation patterns that are natural to human perception yet adversarial to machine learning models.

As illustrated in Figure 2, the network receives a log-magnitude spectrogram as input and passes it through a lightweight architecture composed of depthwise-pointwise convolutional(DW/PW Conv) blocks, each followed by a Gaussian Error Linear Unit(GELU) activation, prior to processing by a bidirectional Gated Recurrent Unit(GRU) for temporal context modeling. The convolutional layers capture local spectral patterns such as formant structures, while the GRU ensures frame-level consistency across time. The final 1×1 convolution produces a multiplicative mask constrained to the range $[0.8, 1.2]$, which is applied directly to the input spectrogram.

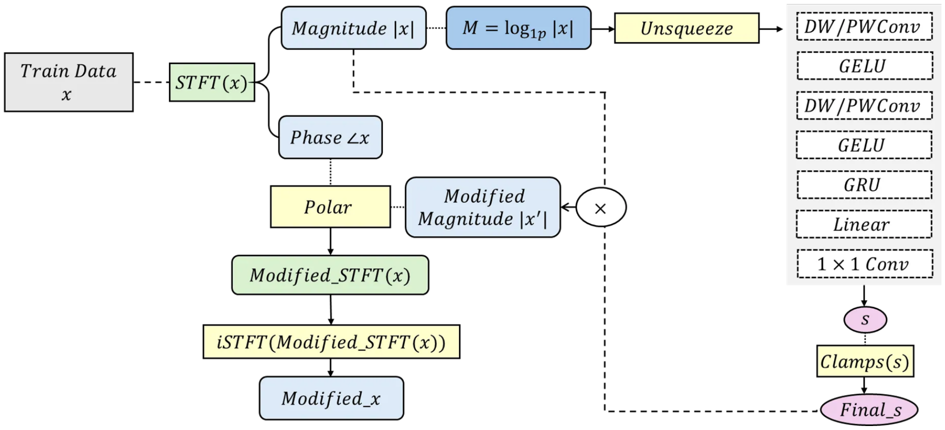

Figure 2: Proposed BandMaskNet module. A multiplicative time–frequency mask perturbs the input spectrum.

The key design principle of this module is to preserve machine recognition accuracy while destabilizing speaker-specific characteristics. To achieve this, BandMaskNet selectively perturbs frequency regions that are less perceptible to human listeners, while constraining the mask magnitude to prevent excessive degradation of audio quality. As a result, users can still perceive the utterance naturally, whereas speaker embedding models fail to obtain a stable representation for reliable training.

4.2 Selective Distortion

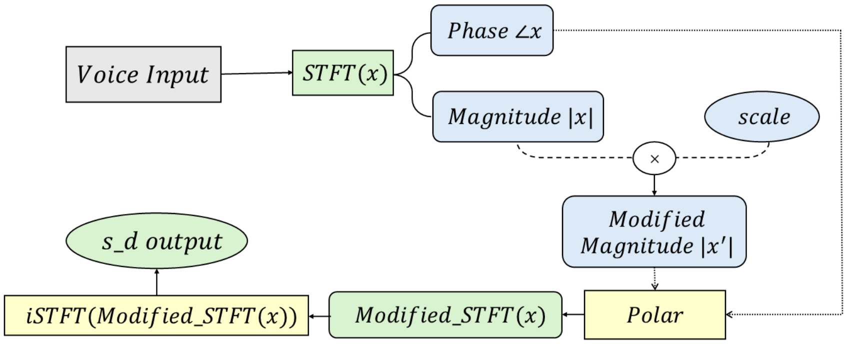

Figure 3: Stochastical perturbation of formant-related frequency bands (Selective distortion module)

The Selective Distortion module selectively perturbs specific frequency bands that are critical for speaker recognition, particularly those associated with formants [27, 6]. Instead of injecting random noise, it introduces subtle perturbations by stochastically scaling the magnitude spectrum at each frame. Figure 3 illustrates the overall structure of this process.

An important aspect of this design is that the phase information is preserved. Consequently, the reconstructed signal does not introduce unnatural artifacts such as metallic noise or phase distortion. While the perturbed speech remains perceptually natural to human listeners, speaker embedding models struggle to extract consistent features, thereby significantly reducing the success rate of voice cloning attacks.

4.3 Prosody Modification

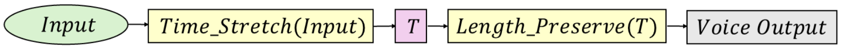

Figure 4: Perturbation of speech rhythm with pitch preservation (Prosody modification module)

In the final stage, prosody modification, one of two approaches is employed. In general, prosodic manipulation can be categorized into: (i) pitch-shifting, which adjusts timbre by altering the pitch, and (ii) time-stretching, which subtly modifies temporal characteristics such as speech rhythm and rate [21]. In this study, we preserve the speaker's pitch to maintain their inherent timbre and identity, while instead applying temporal adjustments. This design encourages speaker embedding models to lose temporal consistency across utterances, thereby reducing the accuracy of both voice cloning and speaker identification.

This process has minimal impact on the performance of the STT engine. While users can naturally perceive the utterance with its original meaning intact, synthesis models face difficulty generalizing speaker characteristics due to unstable rhythmic patterns. In particular, when combined with BandMaskNet and Selective Distortion, the prosody modification module enhances the multidimensional defense effect, thereby providing stronger security guarantees compared to standalone techniques.

5 Experiments

5.1 Experimental Setup

The performance of VoiceShield was evaluated in two aspects. First, STT accuracy was measured to examine whether the speech data remain practically usable in service applications after the application of VoiceShield. Second, the defensive effectiveness against voice cloning attacks was validated through speaker verification experiments.

The Whisper-base model [25] was employed for STT, where word error rate (WER) was measured for English, and character error rate (CER) as well as phoneme error rate (PER) were measured for Korean. For speaker verification, the ECAPA-TDNN model from SpeechBrain [5] was utilized. This model was trained on the VoxCeleb1/2 datasets and determines speaker identity by computing the cosine similarity between embedding vectors of two speech samples. The evaluation datasets consisted of Korean speech (50 pairs from the AI-Hub counseling corpus [1]) and English speech (50 pairs from the VCTK corpus [31]). For each language, four conditions were compared: (i) original speech, (ii) VoiceShield-protected speech, (iii) cloned speech generated from original speech, and (iv) cloned speech generated from VoiceShield-protected speech.

Data	Avg CER(%)	Avg PER(%)
Benchmark [25]	15.2	–
Original Speech	3.62	2.96
VS* Applied Speech	14.59	10.49

* VS: VoiceShield

(a) Korean STT Results

Data	Avg WER(%)
Benchmark [25]	4.5
Original Speech	1.18
VS* Applied Speech	1.16

* VS: VoiceShield

(b) English STT Results

Table 1: Comparison of STT performance before and after applying VoiceShield: (a) CER/PER for Korean, (b) WER for English.

5.2 Speech-to-Text Accuracy

The Korean results presented in Table 1(a) demonstrate that applying VoiceShield substantially increased the CER (from 3.62% to 14.59%) and the PER (from 2.96% to 10.49%). While CER is sensitive to orthographic mismatches—even when the underlying pronunciations are acoustically similar—PER captures structural variations at the level of phonetic units (i.e., initial, medial, and final consonants). This observation aligns with the intended effect of the Prosody Modification module, which introduces temporal and rhythmic distortions that reshape phonetic realizations. Hence, the results indicate that VoiceShield does not merely induce superficial spelling mismatches but fundamentally alters the phonological composition of the signal, thereby preventing synthesis models from establishing stable acoustic correspondences.

Moreover, both CER and PER exhibited comparable relative increases, with CER rising by nearly four times and PER by approximately 3.5 times. Such proportional growth across the two metrics suggests that the perturbations introduced by VoiceShield were not confined to isolated phonemes but systematically affected the phonological structure as a whole. This consistency across the two metrics empirically supports VoiceShield's multi-domain perturbation strategy, where the combined effects of BandMaskNet, Selective Distortion, and Prosody Modification jointly destabilize the speech manifold at different acoustic levels. Importantly, despite this distortion, the resulting CER remained below the Korean benchmark of Whisper-large (15.2%). These findings imply that VoiceShield successfully strikes a balance between robustness and usability: it preserves the utility of STT-based applications (e.g., meeting transcription and voice command services) while simultaneously undermining the phonetic regularities that are critical for effective synthesis-based attacks.

In contrast, the English results shown in Table 1(b) reveal that the WER exhibited virtually no variation, shifting only marginally from 1.18% to 1.16%, while remaining substantially lower than the Whisper-large benchmark for English (4.5%). This finding suggests that, within English speech, VoiceShield preserves recognition accuracy almost entirely, thereby maintaining the usability of STT systems, while concurrently disrupting speaker-specific traits in a manner that strengthens defenses against synthesis-based attacks. The robustness of this dual effect is further corroborated by the speaker verification results presented in the subsequent section.

In summary, VoiceShield perturbs pronunciation structures in Korean, as evidenced by the increases in CER and PER, while preserving the functional usability of speech recognition. In English, it maintains WER at a stable level yet continues to provide protective effects against synthesis-based attacks. These results demonstrate that the proposed approach achieves a balanced trade-off between security and usability, while also exhibiting adaptability across different language contexts.

5.3 Speaker Verification Robustness

In the speaker verification experiments, the defensive effectiveness was evaluated using similarity scores computed by the ECAPA-TDNN model. For Korean speech, all 50 pairs yielded similarity scores below 0.5 and were thus not recognized as originating from the same speaker. This result indicates a complete suppression of speaker identity features, demonstrating that VoiceShield successfully removes the speaker's acoustic fingerprint (voiceprint) in Korean utterances. The outcome also reinforces the finding from the STT results—namely, that the system perturbs pronunciation and spectral patterns deeply enough to disrupt speaker embeddings used in synthesis and verification models.

In the English dataset, 46 out of 50 pairs (92%) were classified with similarity scores below 0.5, indicating a high level of defensive performance. However, four pairs recorded scores above 0.5 and were incorrectly identified as originating from the same speaker. This outcome can be attributed to two factors. First, since ECAPA-TDNN was trained primarily on large-scale English corpora such as VoxCeleb, it is possible that certain speaker characteristics were partially preserved in English utterances despite the perturbations introduced by VoiceShield. Second, because English possesses a relatively simpler phonemic structure compared to Korean, residual speaker-specific patterns may persist to some extent even under prosodic and spectral modifications. Together, these observations suggest that while VoiceShield robustly defends against cloning in both languages, fine-grained adaptation of perturbation intensity may further enhance its resilience against models trained on language-specific embeddings.

Additionally, it was observed that increasing the decision threshold to the range of 0.55–0.6 could convert some of the same-speaker classifications in the English dataset into different-speaker judgments, thereby enhancing defensive performance. However, such threshold adjustment inevitably entails the risk of false negatives, where genuine same-speaker pairs are misclassified. This highlights an important practical consideration: operational settings can flexibly tune VoiceShield's trade-off curve between security sensitivity and user accessibility depending on the application domain—for instance, employing stricter thresholds in financial or authentication systems, and more relaxed thresholds in general user services.

6 Conclusion

In this work, we propose VoiceShield, a novel real-time voice protection framework designed to counter AI-driven speech synthesis and cloning attacks. VoiceShield incorporates BandMaskNet, Selective Distortion, and Prosody Modification modules to introduce multilayer perturbations in the spectral, frequency-band, and prosodic domains, thereby effectively disrupting the stable learning of synthesis models.

Experimental results demonstrated that VoiceShield achieves a strong balance between security and practicality. In particular, Whisper-based STT evaluations confirmed that the proposed method preserves recognition accuracy within usable bounds — maintaining negligible WER variation in English and acceptable CER/PER increases in Korean — while ECAPA-TDNN-based speaker verification experiments verified over 92% defense effectiveness against voice cloning attacks. These results suggest that users can seamlessly utilize STT-based services (e.g., meeting transcription and voice commands) while effectively preventing the illicit misuse of voice data.

Despite these results, VoiceShield still presents several limitations. First, its defensive performance was evaluated against a limited set of synthesis models; hence, experiments with more diverse and contemporary voice cloning architectures are necessary to validate robustness against evolving attack methods. Second, as the experiments were primarily conducted on English and Korean datasets, further validation in multilingual and cross-dialectal environments is necessary to assess language generalizability. Finally, the current evaluation mainly relies on objective metrics such as WER, CER, and embedding distance; incorporating subjective human evaluations (e.g., Mean Opinion Score for perceptual quality and anonymization perception) would provide a more comprehensive understanding of user experience and perceived naturalness.

For future research, we plan to address these limitations by expanding the evaluation scope to include multiple languages, dialects, and contemporary VC models, as well as by integrating human-centered perceptual assessments. Furthermore, future work will include analyzing computational efficiency and user experience in interactive and streaming scenarios to further advance VoiceShield into a fully deployable defense framework for security-critical domains such as finance, authentication, and customer service.

7 Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155966, Artificial Intelligence Convergence Innovation Human Resources Development (Ewha Womans University)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00357189).

References

[1] AI-Hub counseling corpus. https://www.aihub.or.kr/aihubdata/data/view.do?&dataSetSn=100. Accessed: 2024-09-09.

[2] Sercan Ö. Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi. Deep Voice: Real-time Neural Text-to-Speech. https://arxiv.org/abs/1702.07825, 2017.

[3] Edresson Casanova, Julian Weber, Christopher Dane Shulby, Eren Gölge, José Soares, Anderson da Silva Soares, Sandra Aluisio, and Moacir Antonelli Ponti. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. https://arxiv.org/abs/2112.02418, 2022.

[4] Yufei Chen, Qiushi Huang, Jiahao Pan, Yifan Yang, Yiming Li, Zhan Qin, and Kui Ren. VSMask: Defending Against Voice Synthesis Attacks via Imperceptible Perturbations. https://arxiv.org/abs/2305.05736, 2023.

[5] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. https://arxiv.org/abs/2005.07143, 2020.

[6] Qing Fei, Wei Hou, Xin Hai, and Xu Liu. VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect. https://doi.org/10.48550/arXiv.2502.10329, 2025.

[7] Qingyuan Fei, Wenjie Hou, Xuan Hai, and Xin Liu. VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect. https://arxiv.org/abs/2502.10329, 2025.

[8] Priyanka Gupta, Hemant A. Patil, and Rodrigo Capobianco Guido. Vulnerability issues in Automatic Speaker Verification (ASV) systems. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024.

[9] Qiushi Huang, Yufei Chen, Jiahao Pan, Yifan Yang, Yiming Li, Zhan Qin, and Kui Ren. Partial Fake Speech Attacks in the Real World Using Deepfake Voice Samples. *Journal of Cybersecurity and Privacy*, 5(1):6, 2025.

[10] Yihao Huang, Jianhua Yin, Songyang Zhang, Hao Chen, Yaliang Li, Bolin Ding, Ying Shen, Kai Lei, and Xiaohui Liang. AntiFake: Improving Audio Deepfake Detection with Contrastive Learning. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2023.

[11] Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. https://arxiv.org/abs/2402.01708, 2024.

[12] Corentin Jemine. SV2TTS: Real-Time Voice Cloning. https://arxiv.org/abs/1806.04558, 2018.

[13] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. How Deep Are the Fakes? Focusing on Audio Deepfake: A Survey. https://arxiv.org/abs/2111.14203, 2021.

[14] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. https://arxiv.org/abs/2010.05646, 2020.

[15] Yann Kowalczuk and Jan Holub. Evaluation of digital watermarking on subjective speech quality. *Scientific Reports*, 11(1):20185, 2021.

[16] Jingjin Li, Chao Chen, Lei Pan, Mostafa Rahimi Azghadi, Hossein Ghodosi, and Jun Zhang. Security and Privacy Problems in Voice Assistant Applications: A Survey. *Computers & Security*, 2023.

[17] Renyuan Li, Zhibo Liang, Haichuan Zhang, Tianyu Shi, Zhiyuan Cheng, Jia Shi, Carl Yang, and Mingjie Tang. CloneShield: A Framework for Universal Perturbation Against Zero-Shot Voice Cloning. https://doi.org/10.48550/arXiv.2505.19119, 2025.

[18] Ze Li, Yao Shi, Yunfei Xu, and Ming Li. Adversarial Attacks and Robust Defenses in Speaker Embedding based Zero-Shot Text-to-Speech System. https://doi.org/10.48550/arXiv.2410.04017, 2024.

[19] Peiyu Liu, Yuepeng Zhang, Ziqian Zeng, Yuxuan Zhang, Yinpeng Dong, Hang Su, and Jun Zhu. RoVo: Robust Voice Obfuscation against Unauthorized Voice Cloning. https://arxiv.org/abs/2505.12686, 2025.

[20] Govind Mittal, Arthur Jakobsson, Kelly Marshall, Chinmay Hegde, and Nasir Memon. Pitch: Ai-assisted tagging of deepfake audio calls using challenge-response. In *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, ASIA CCS '25, page 559–575, New York, NY, USA, 2025. Association for Computing Machinery.

[21] Matthew Morrison, Lukas Rencker, Zhiyao Jin, Nicholas J. Bryan, Juan P. Caceres, and Bryan Pardo. Context-aware prosody correction for text-based speech editing. In *ICASSP*, 2021.

[22] Eoin O'Reilly. Limitations of Post-Hoc Watermarking Techniques for Speech. https://arxiv.org/abs/2504.10782, 2025.

[23] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. https://arxiv.org/abs/1905.05879, 2019.

[24] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. OpenVoice: Versatile Instant Voice Cloning. https://arxiv.org/abs/2312.01479, 2023.

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision (Whisper). https://arxiv.org/abs/2212.04356, 2022.

[26] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech:

Fast, Robust and Controllable Text to Speech. https://arxiv.org/abs/1905.09263, 2019.

[27] Brian Roberts, Robert J. Summers, and Philip J. Bailey. Formant-frequency variation and informational masking of speech by extraneous formants: Evidence against dynamic and speech-specific acoustical constraints. *Journal of Experimental Psychology: Human Perception and Performance*, 40(1):295–305, 2014.

[28] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (Tacotron 2). https://arxiv.org/abs/1712.05884, 2018.

[29] Mingjie Sun, Yinpeng Dong, Hang Su, and Jun Zhu. VoiceBlock: Robust Unauthorized Speech Synthesis Detection via Adversarial Perturbation. In *Proceedings of NeurIPS*, 2022.

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. https://arxiv.org/abs/1609.03499, 2016.

[31] Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit. The Centre for Speech Technology Research (CSTR), University of Edinburgh.

[32] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yan Wu, Yao Qian, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Huaming Wang, Zhihua Wei, Eric Sun, and Hsiao-Wuen Hon. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers (VALL-E). https://arxiv.org/abs/2301.02111, 2023.

[33] Kun Wang, Meng Chen, Li Lu, Jingwen Feng, Qianniu Chen, Zhongjie Ba, Kui Ren, and Chun Chen. From One Stolen Utterance: Assessing the Risks of Voice Cloning in the AIGC Era. https://openreview.net/pdf?id=c08w2AG1kh, 2025.

[34] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis. https://arxiv.org/abs/1703.10135, 2017.

[35] Xiang Yin, Hang Su, and Jun Zhu. A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, 9(19):4050, 2019.

[36] Qing Yuan, Xinyu Yang, Yutong He, Wenchao Xu, and Shu-Tao Xia. SafeSpeech: Real-Time Defense Against Zero-Shot Voice Cloning. https://arxiv.org/abs/2504.09839, 2025.

[37] Haotian Zhang, Yucheng Shi, Tianyi Liu, Xueqiang Yan, Yinpeng Dong, Hang Su, and Jun Zhu. CloneShield: Robust Real-Time Defense Against Unauthorized Voice Cloning. https://arxiv.org/abs/2505.19119, 2025.