Recoverable Audio Adversarial Examples for Voice Protection in One-shot Voice Conversion

Chenshuai Shu† , Tianpeng Zheng†, Yanxiang Chen*
School of Computer Science and Information Engineering
Hefei University of Technology, Hefei, China
E-mail: {scs123,zhengtp}@mail.hfut.edu.cn, chenyx@hfut.edu.cn

***Abstract*—The advancements in voice conversion (VC) technology, specifically one-shot VC, have raised significant concerns regarding privacy and identity verification. Researchers use adversarial examples to protect voice data, as they can mislead algorithms. However, existing methods for protection via adversarial examples lack recoverability and reversibility, rendering them ineffective as robust protection mechanisms. To address this issue, we propose a recoverable audio adversarial example generation approach. It not only effectively protects speech data but also allows authorized users to restore the original speaker characteristics when needed. Our approach is tested using one-shot VC, with both subjective and objective evaluations conducted under white-box and black-box scenarios. The results demonstrate that the output of the VC model differs significantly from the protected speaker's voice, and our approach allows authorized users to restore the utterance for normal use.**

***Index Terms*—voice conversion, recoverable audio adversarial examples, speaker verification, privacy protection**

I. Introduction

Voices shared by users on social media platforms can be manipulated via one-shot voice conversion (VC) to transform any utterance into the voice of a reference speaker. Voice conversion technology replicates a speaker's vocal characteristics, offering significant benefits, such as enabling the creation of AI singers[1]–[3] and facilitating cross-lingual voice conversion [4]–[6]. However, advancements in VC technology also risk misuse by malicious users. For example, cloning political leaders' or executives' voices could manipulate public opinion or spread misinformation. Moreover, VC technology poses a serious threat to security and privacy by-passing speaker verification and recognition systems[7], [8]. The development of one-shot voice conversion [9]–[13], in particular, enables cloning a speaker's voice from a single sample. Consequently, the privacy and security challenges posed by VC technology necessitate urgent and multifaceted consideration to address their implications effectively.

One effective approach to mitigating this issue is adversarial examples [14]–[18], which are designed to mislead and disrupt intelligent algorithms, causing them to produce incorrect outputs. Using adversarial examples to extract speaker features prevents the VC model from accurately replicating the target speaker's voice. Audio adversarial example generation can be divided into two scenarios: white-box [19] and black-box [20], [21]. In the white-box setting, adversarial examples are crafted with complete knowledge of the VC model parameters and architecture. For instance, Huang et al. [22] proposed an optimization-based method to iteratively generate adversarial examples. While this approach avoids additional training, it is often time-consuming. On the other hand, Dong et al. [23] leveraged Generative Adversarial Networks (GANs) to produce adversarial examples, enabling faster generation at the cost of reduced stealthiness. Furthermore, the transferability of adversarial examples [24] allows their use in black-box scenarios, where the defender lacks access to the model details. Black-box scenarios make adversarial examples highly practical in real-world applications, posing a significant threat to the reliability of deep neural networks (DNN). Cheng et al. [25] introduced a black-box adversarial linguistic feature-based attack pipeline to generate adversarial example.

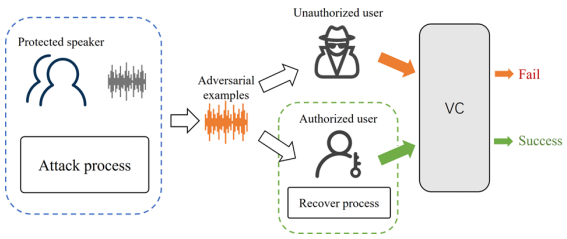

Fig. 1. Protection scenarios for recoverable adversarial samples.

As shown in Figure 1, adversarial samples can prevent unauthorized users from outputting correct results using the VC model. However, existing audio adversarial examples lack recoverability and reversibility, limiting their effectiveness as a robust protection mechanism. To address this, we propose a method for generating recoverable audio adversarial examples, inspired by [26]. Unlike existing methods, as shown in Figure 1, we provide a recovery process for authorized users. The main contributions of this paper are as follows:

- We propose a recoverable audio adversarial example generate method to protect the speaker's speech from cloning by the VC system, while authorized users can obtain clean samples through the recovery process.
- We proposed three approaches to generate audio adversarial samples and recovery samples using common VC mod-els.
- Through objective and subjective tests, we found ad-

†These authors contributed equally to this work and should be considered co-first authors. *Corresponding author

versarial samples widen the gap between VC outputs and protected speaker features, while recovered samples effectively restore adversarial ones to clean samples.

II. Methodology

A. Problem Formulation

A common voice conversion model uses an encoder-decoder structure, with the encoder divided into a content encoder and a speaker encoder. Figure 2 shows the VC model structure. The content encoder E_c processes an input utterance u to extract content information, producing $E_c(u)$, while the speaker encoder E_s encodes the speaker characteristics of an input utterance x into a latent vector $E_s(x)$. The decoder D then takes $E_c(u)$ and $E_s(x)$ as inputs to generate a spectrogram $F(u, x)$, which combines the content information from $E_c(u)$ and the speaker characteristics from $E_s(x)$.

Given a one-shot VC system $F(\cdot, \cdot)$, a protected speaker utterance x, and a content utterance u, the attack process generates an adversarial perturbation δ that is added to x. The recovery process, on the other hand, generates a recovery perturbation $\delta^{'}$. When $\delta^{'} = \delta$, the adversarial perturbation can be effectively removed, thereby recovering the adversarial example to its original clean state.

For the clean example, one-shot VC system $F(\cdot, \cdot)$ generates an utterance $F(x, u)$ such that:

$$F(x, u) = x \tag{1}$$

The attack process aims to generate a perturbation such that:

$$F(x + \delta, u) \neq x \tag{2}$$

The recovery process aims to generate a perturbation $\delta^{'}$ such that:

$$F(x + \delta - \delta^{'}, u) = x \tag{3}$$

Here, the protected speaker utterance input to the VC system is represented as a mel-spectrogram $x \in R^{C \times T}$,C represents the number of mel-frequency channels. The generated perturbation is denoted as $\delta \in R^{C \times T}$. Unless otherwise specified,the term "utterance" in the following text refers to its mel-spectrogram feature representation. Building on insights from prior research [22], we propose three approaches for executing the attack and recovery processes. These three approaches are designed based on the common encoder-decoder VC architecture: the end-to-end approach uses the output of the decoder as the target for generating adversarial or recovered samples, the embedding approach directly alters the embedding of the speaker's utterance using the speaker encoder, and the feedback approach combines the first two approaches.

B. End-to-end approach

A straightforward approach to generating adversarial examples is to target the decoder output $F(\cdot, \cdot)$, in an end-to-end method. As shown in Figure 2, given protected input x and target speaker utterance y, both targeted attacks and recovery

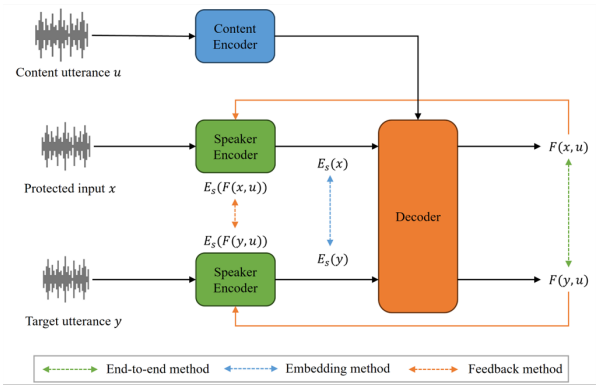

Fig. 2. The three proposed approaches. The dashed line in-dicates the update of the perturbation.

can be performed.

Attack process

$$\arg\min_{\delta} \begin{array}{l} \mathcal{L}(F(x+\delta, u), F(y, u)) \\ \quad - \lambda \cdot \mathcal{L}(F(x+\delta, u), F(x, u)) \end{array} \tag{4}$$

The first term in (4) aims to make the VC model's output sound more like y's speaker, while the second term ensures that the output diverges from x's speaker. Here, λ is a hyperparameter used to balance the importance between the source and target speakers.

Recovery process

$$\arg\min_{\delta'} -\lambda \cdot \mathcal{L}(F(x+\delta-\delta^{'}, u), F(y, u)) \tag{5}$$

The recovery process aims to distance the adversarial sample $x + \delta$ from y's speaker, thereby generating a recovery perturbation $\delta^{'}$ to remove the original perturbation δ as much as possible, ultimately recover to the clean example.

C. Embedding approach

In the VC model, the speaker encoder E converts utterances into latent vectors, where same speaker samples cluster closely and different speaker samples separate. As shown in Figure 2,the embedding approach uses the speaker encoder to directly modifying speaker embeddings.

Attack process

$$\arg\min_{\delta} \begin{array}{l} \mathcal{L}(E_s(x+\delta), E_s(y)) \\ \quad - \lambda \cdot \mathcal{L}(E_s(x+\delta), E_s(x)) \end{array} \tag{6}$$

The attack process induces a geometric transformation in the embedding space, causing the latent representation of the protected speaker to converge toward the target y's cluster while diverging from the x's speaker characteristic distribution.

Recovery process

$$\arg\min_{\delta'} -\lambda \cdot \mathcal{L}(E_s(x+\delta-\delta^{'}), E_s(y)) \tag{7}$$

In this approach, both the attack and recovery processes rely solely on the speaker encoder. Since only the speaker

encoder is involved, the embedding method proves to be a more efficient approach.

D. Feedback approach

The third approach attempts to integrate the previously mentioned two approaches by feeding the spectrogram $F(x,u)$ output by the decoder D back into the speaker encoder E_s,while also considering the speaker embeddings obtained through this process. Consequently, the attack and recovery processes can be expressed as:

Attack process

$$\arg\min_{\delta} \begin{array}{l} \mathcal{L}(E_s(F(x+\delta,u)),E_s(y)) \\ \qquad -\lambda\cdot\mathcal{L}(E_s(F(x+\delta,u)),E_s(x)) \end{array} \tag{8}$$

Recovery process

$$\arg\min_{\delta'} -\lambda\cdot\mathcal{L}(E_s(F(X+\delta-\delta^{'},u)),E_s(F(y,u))) \tag{9}$$

III. Experimental Settings

We conducted experiments on the models proposed by Chou et al [9] (referred to as the Chou's model hereafter), VQVC+ [10], and FreeVC [13]. These models are capable of performing one-shot VC for unseen speakers without requiring finetuning, making them well-suited for our application scenario.We consider two parties in the experiments. The defender's objective is to safeguard the protected speaker utterance against voice cloning by using the adversarial examples. Conversely, the attacker aims to exploit various VC models to clone the target speaker's voice as effectively as possible.

A. Protect scenarios

Two scenarios were tested in this study. In the first scenario, the defender knows full architecture and trained parameters of the VC model used by the attacker. This scenario is referred to as the white-box setting and is exclusively conducted on the Chou's model. Specifically, the defender generates adversarial examples or recovery examples using the Chou's model.

In the second scenario is known as the black-box setting, the defender still uses Chou's model to generate adversarial or recovery examples. The attacker then applies two defender unknown models, VQVC+ and FreeVC, for voice conversion.

B. Defence procedure

We chose the MSE loss as $\mathcal{L}(\cdot,\cdot)$ and set $\lambda = 0.5$. The perturbation δ or $\delta^{'}$ was initialized with a standard normal distribution, and the Adam optimizer was employed to iteratively update δ or $\delta^{'}$ based on the loss functions defined in equations (4), (5), (6), (7), (8), and (9). The learning rate was set to 0.001,and the number of iterations was fixed at 1500. This configuration was applied to both the attack and recovery processes.

C. Evaluation metrics

We employed Speaker Verification (SV) to assess whether the attacker successfully cloned the characteristic of the protected speaker, using the Attack Success Rate (ASR) to quantify the effectiveness of our proposed approach. ASR indicates the success rate of an attacker using VC model to clone protected speaker characteristics. Additionally, we utilized Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) to evaluate the audio quality of the adversarial examples and recovery examples. Higher PESQ and STOI scores indicate better audio quality. Furthermore, to ensure comprehensive testing, we also incorporated subjective evaluations.

IV. Experiment Results

A. Objective tests

In the objective evaluation, we employed speaker verification accuracy as the attack success rate. The speaker verification system employed in this study first encodes two input utterances into embeddings and then computes the similarity between them. If the similarity exceeds a predefined threshold,the system determines that the two utterances are from the same speaker. During testing, we consistently compared generated utterances with utterances from the protected speaker.The speaker verification accuracy, used in the following analysis, is the percentage of cases where the system identifies two utterances from the same speaker.

The verification system employed in this study is based on a pretrained d-vector model1 [27], which is different from the speaker encoders of the three attack models. Following the procedure described below, the threshold was determined based on the Equal Error Rate (EER) obtained from verifying randomly sampled utterance pairs in the VCTK corpus. We extracted 10 utterance samples for each of the 94 speaker pairs in the dataset. For positive samples, the similarity was computed between different utterances from the same speaker,while for negative samples, it was computed between random utterances from other randomly selected speakers. Therefore,the threshold was set to 0.657, with an EER of 0.041.

From the 109 speakers in the VCTK corpus, we randomly selected 108 speakers to provide speaker features, while the remaining one provided content utterances. The 108 speakers were divided into two groups: 54 speakers were assigned as protected speakers x, and the other 54 served as target speakers y for attack and recovery reference. For each speaker, 10 utterance samples were randomly chosen. These samples were then combined via random pairings to construct the (u,x,y) test dataset.

In the testing phase, $F(x,u)$ is considered as the original input (OI), $F(x+\delta,u)$ as the adversarial input (AI), and $F(x+\delta-\delta^{'},u)$ as the recovery input (RI). Here, $x+\delta$ denotes the adversarial sample generated based on Chou's method; $x+\delta-\delta^{'}$ represents the recovery sample generated based on Chou's method. Table 1 shows the attack success rate for three approaches,end-to-end, embedding, and feedback, under

TABLE I

THE ATTACK SUCCESS RATE FOR THREE APPROACHES. THE ABBREVIATIONS REPRESENT DIFFERENT INPUT OF VC MODEL, "OI", "AI" AND "RI" DENOTE THE INPUT TYPE, ORIGINAL INPUT, ADVERSARIAL INPUT AND RECOVERY INPUT.

Model	End-to-end			Embedding			Feedback		
	OI	AI	RI	OI	AI	RI	OI	AI	RI
Chou's	80.56	40.00	60.00	80.37	37.41	41.30	77.96	36.11	42.04
VQVC+	47.59	32.41	45.00	47.78	30.00	36.30	46.11	29.07	38.33
FreeVC	90.93	74.26	78.52	91.48	16.11	72.22	89.26	15.56	69.63

TABLE II

THE ADVERSARIAL EXAMPLE AND RECOVERY EXAMPLE QUALITY. THE "A" AND "R" INDICATE ADVERSARIAL AND RECOVERY. THE "E2E", "EMB" AND "FB" INDICATE END-TO-END, EMBEDDING AND FEEDBACK.

Approach	PESQ		STOI	
	a	r	a	r
e2e	2.52	2.91	0.80	0.84
emb	2.52	2.79	0.80	0.83
fb	2.71	2.97	0.81	0.83

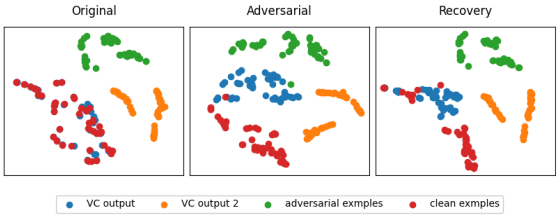

Fig. 3. t-SNE analysis of objective tests, the title indicates the VC model input type.

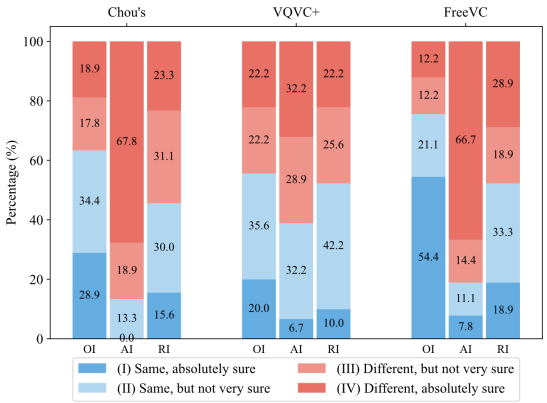

Fig. 4. Subjective evaluation results with end-to-end approach for write-box and black-box."OI", "AI" and "RI" denote the input type,original input, adversarial input and recovery input.

both white-box and black-box scenarios. In the white-box scenario, all three approaches demonstrate effective protection capabilities, with the end-to-end approaches showing the best performance. After protection, the speech fails to pass the speaker verification system effectively, when subjected to one-shot voice conversion. Moreover, for the speech with added adversarial perturbations, the recovery process can restore it as much as possible,resulting in a certain improvement in speaker verification accuracy.

In the black-box scenario, we analyzed two different one-shot voice conversion models. VQVC+ did not perform well with the original input at a threshold of 0.657, but still demonstrated some effectiveness. In contrast, FreeVC achieved relatively favorable results in both the embedding and feedback approaches, with significant effects observed in both attack and recovery scenarios.

In addition to attack success rate, we also utilized the PESQ and STOI as metrics for comparing perceptual quality. Higher PESQ,along with STOI scores, indicate better audio quality. As shown in Table 2, the fb approach performed better in PESQ. There was no significant difference in STOI among the three approaches.

To effectively test our attack and recovery performance,each utterance x was paired with a randomly selected utterance u from a different speaker and utterances y from a speaker of the opposite gender of x. We sampled 50 pairs of protected samples (u, x, y), and used the end-to-end approach to generate adversarial and recovery samples. Subsequently,we used the clean samples, adversarial samples, and recovery samples to generate VC output samples. Then, we utilized t-SNE to analyze their distributions, as shown in Figure 3. The VC output using clean samples input is closer to the clean samples, indicating that VC was effective before the attack. The VC output using adversarial samples input is away from clean examples and closer to adversarial samples, demonstrating the success of attack. The VC output using recovery samples input moves away from adversarial samples and closer to clean samples. Under three input conditions, the VC output exhibited no convergence toward the other speaker. It indicates that our approach can effectively protect the speaker's characteristics from being cloned by the VC model and can successfully restore the adversarial samples to clean samples.

B. Subjective tests

Objective tests might be insufficient to determine if the attack and recovery successfully protected speaker characteristic in human perception. Therefore, we also perform a subjective assessment under both white-box and black-box scenarios using the end-to-end approach. We randomly selected 10 speakers from the objective evaluation samples as protected speakers.For each, we randomly chose a target speaker to generate adversarial and recovered samples.Subjects then selected one of four options to judge if the two voices

came from the same speaker: (I) Same, absolutely sure; (II) Same, but not very sure; (III) Different, but not very sure, and (IV) Different, absolutely sure.

The results align with the findings from the objective tests in Figure 4. In both the white-box (Chou's) and black-box (VQVC+ and FreeVC) scenarios, adversarial samples are used as inputs for the VC system, only 0%, 6.7%, and 7.8% deem VC output originate from the same speaker as the protected input. Conversely, when the recovered samples were used as VC inputs, the proportion of VC outputs perceived as coming from the same speaker rises to a level similar to the original inputs. This demonstrates that the adversarial samples can effectively shield the speaker's features, and the proposed recovery process can successfully remove the adversarial.

V. Conclusions

We introduce a reversible approach to protect speaker characteristics by adding subtle perturbation to utterance, preventing one-shot VC systems from cloning them. We also add a recovery mechanism to create recoverable adversarial samples, allowing authorized users to use the system normally and enhancing the protection mechanism's effectiveness and practicality. We proposed three approaches and evaluated them on three advanced VC models. Results show our approaches significantly reduce VC's ability to mimic voices and can restore speaker characteristics in black-box scenarios. Future work will explore more complex models and larger datasets to improve the generation speed of adversarial and recovered-samples.

References

[1] S. Liu, Y. Cao, D. Su, and H. Meng, "Diffsvc: A diffusion probabilistic model for singing voice conversion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, pp. 741–748.

[2] Y. Zhou and X. Lu, "Hifi-svc: Fast high fidelity cross-domain singing voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6667–6671.

[3] S. Liu, Y. Cao, N. Hu, D. Su, and H. Meng, "Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation," in *2021 ieee international conference on multimedia and expo (icme)*, IEEE, 2021, pp. 1–6.

[4] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[5] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "Using joint training speaker encoder with consistency loss to achieve cross-lingual voice conversion and expressive voice conversion," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2023, pp. 1–8.

[6] K. Ezzine, J. Di Martino, and M. Frikha, "Any-to-one non-parallel voice conversion system using an autoregressive conversion model and lpcnet vocoder," *Applied Sciences*, vol. 13, no. 21, p. 11 988, 2023.

[7] X. Tian, R. K. Das, and H. Li, "Black-box attacks on automatic speaker verification using feedback-controlled voice conversion," *arXiv preprint arXiv:1909.07655*, 2019.

[8] D. Cai, Z. Cai, and M. Li, "Identifying source speakers for voice conversion based spoofing attacks on speaker verification systems," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[9] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[10] D.-Y. Wu, Y.-H. Chen, and H.-Y. Lee, "VQVC+: One-shot voice conversion by vector quantization and u-net architecture," *arXiv preprint arXiv:2006.04154*, 2020.

[11] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 5954–5958.

[12] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," *arXiv preprint arXiv:2109.13821*, 2021.

[13] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[14] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Enabling fast and universal audio adversarial attack using generative model," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 14 129–14 137.

[15] 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.

[16] Z. Yu, Y. Chang, N. Zhang, and C. Xiao, "SMACK: Semantically meaningful adversarial audio attack," in *32nd USENIX security symposium (USENIX security 23)*, 2023, pp. 3799–3816.

[17] P. Wang, H. Gao, X. Guo, Z. Yuan, and J. Nian, "Improving the security of audio captchas with adversarial examples," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 2, pp. 650–667, 2023.

[18] C.-Y. Yang, S. G. Upadhyay, Y.-T. Wu, B.-H. Su, and C.-C. Lee, "Rw-voiceshield: Raw waveform-based adversarial attack on one-shot voice conversion," in *Proc. Interspeech 2024*, 2024, pp. 2730–2734.

[19] S. Chen, L. Chen, J. Zhang, K. Lee, Z. Ling, and L. Dai, "Adversarial speech for voice privacy protection from personalized speech generation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 11 411–11 415.

[20] Y. Chen, X. Yuan, J. Zhang, *et al.*, "{Devil's} whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2667–2684.

[21] G. Chen, S. Chenb, L. Fan, *et al.*, "Who is real bob? adversarial attacks on speaker recognition systems," in *2021 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2021, pp. 694–711.

[22] 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.

[23] S. Dong, B. Chen, K. Ma, and G. Zhao, "Active defense against voice conversion through generative adversarial network," *IEEE Signal Processing Letters*, vol. 31, pp. 706–710, 2024.

[24] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[25] P. Cheng, Y. Wang, P. Huang, *et al.*, "Alif: Low-cost adversarial audio attacks on black-box speech platforms using linguistic features," in *2024 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2024, pp. 1628–1645.

[26] J. Zhang, J. Wang, H. Wang, and X. Luo, "Self-recoverable adversarial examples: A new effective protection mechanism in social networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 562–574, 2022.

[27] 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.