Interspeech 2024
1-5 September 2024, Kos, Greece

RW-VoiceShield: Raw Waveform-based Adversarial Attack on One-shot Voice Conversion

Ching-Yu Yang, Shreya G. Upadhyay, Ya-Tse Wu, Bo-Hao Su, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

gino0950150@gmail.com, shreya@gapp.nthu.edu.tw, crowpeter@gapp.nthu.edu.tw, borrissu@gapp.nthu.edu.tw, cclee@ee.nthu.edu.tw

Abstract

In recent years, there have been significant advancements in one-shot voice conversion (VC), enabling the alteration of speaker traits with just a single sentence. However, as this technology matures and generates increasingly realistic utterances, it becomes vulnerable to privacy concerns. In this paper, we propose RW-VoiceShield to shield voice from replication. This is achieved by effectively attacking one-shot VC models through the application of imperceptible noise generated from a raw waveform-based generative model. Our method undergoes testing using the latest one-shot VC model, conducting subjective and objective evaluations under both black-box and white-box scenarios. Our results indicate significant disparities in speaker characteristics between the utterances generated by the VC model and those of the protected speaker. Furthermore, even with adversarial noise introduced to protected utterances, the speaker's distinct characteristics remain recognizable.

Index Terms: voice conversion, adversarial attack, speaker verification, speaker representation

1. Introduction

Deepfakes, combining deep learning and fake elements, revolutionize the creation of realistic media across various domains such as avatars, chatbots, and artistic creations [1, 2, 3]. Driven by recent advancements in deep learning technology and generative models, this technology seamlessly alters faces [4, 5, 6], voices [7, 8, 9, 10, 11, 12], and expressions [13, 14] in images, audio, and videos, blurring the boundary between reality and fiction. However, it also raises significant concerns regarding privacy and security, including forging legal evidence, facilitating identity theft and financial scams.

Audio deepfakes, a major subset of deepfake technology, specialize in cloning human voices. While initially driven by positive motives, research on audio deepfakes has led to fascinating and practical applications like AI singers [15, 16, 17] or cross-lingual voice conversion [18], which also aided in data augmentation for speech-related studies [19]. However, these advancements have also been exploited by malicious actors. For instance, in 2019, fraudsters utilized AI-based software to mimic a CEO's voice, resulting in a telephone scam that defrauded over USD 243,000 [20]. Among the techniques employed in audio deepfakes, Voice Conversion (VC) is particularly popular but carries significant risks. VC is a technique that alters a source speaker's voice to mimic a particular target style, including elements like speaker identity, prosody, and emotion while preserving the linguistic content. Some studies [21, 22] have shown that VC can effectively fool automatic speaker verification (ASV) systems and speech classification tasks, posing a serious threat to security and privacy. Furthermore, recent advancements in VC, particularly in one-shot approaches [7, 8, 11], enable the synthesis of realistic and high-quality utterances for any speaker using only one example utterance without fine-tuning. This makes cloning someone's voice easier, requiring only a single sample of the target speaker's voice. Therefore, addressing the privacy/safety concerns posed by VC is urgent and warrants discussion.

Conventional method for addressing these issues caused by audio deepfakes is Audio Deepfake Detection, which employs learning-based algorithms to distinguish between genuine and fabricated audio, yielding promising results [23, 24]. However, while effective in detecting disingenuous samples, these methods fail to actively prevent voice replication, leaving copied voices vulnerable to malicious exploitation. Huang et al. [25] recently proposed a method to shield voice characteristics of speech samples from replication by applying perturbations to audio files, introducing an adversarial attack as a defense mechanism. Subsequent studies [26, 27] have also shown promising results. Typically, these studies employ a two-stage attack framework: extracting acoustic features from audio files, applying privacy-preserving perturbations to these features, and synthesizing utterances using a vocoder. However, this approach often encounters feature mismatch, where acoustic features differ from those used in vocoder training. Moreover, vocoders can neutralize privacy-preserving perturbations and even limit the feature size used while applying adversarial attacks. These methods require training a separate vocoder for models with varying feature sizes to avoid attack failures, which is not practical. In this paper, we focus on addressing these issues by directly generating adversarial perturbations in the original raw waveform to shield voices from being cloned.

Specifically, we propose a system named RW-VoiceShield to launch adversarial attacks using raw waveform-based generative models on a state-of-the-art one-shot VC model by drawing inspiration from the embedding attack methodology introduced by Huang et al. [25] and the framework developed by Xie et al. [28] for conducting adversarial attacks on ASV models. Our approach is evaluated through objective and subjective assessments on the CSTR VCTK Corpus [29]. The objective ASV analysis indicates substantial differences in speaker characteristics between the VC model's generated utterances and those of the protected speaker, with adversarial perturbations minimally affecting speaker attributes and speech quality. Moreover, RW-VoiceShield maintains a higher signal-to-noise ratio and achieves better attack effectiveness compared to the baseline. The subjective tests reveal that the attacked VC model has difficulties in convincingly replicating the protected speaker's voice according to human perception. Our analysis confirms that our idea effectively distances the speaker characteristics replicated by the VC from those of the protected speaker.

10.21437/Interspeech.2024-458

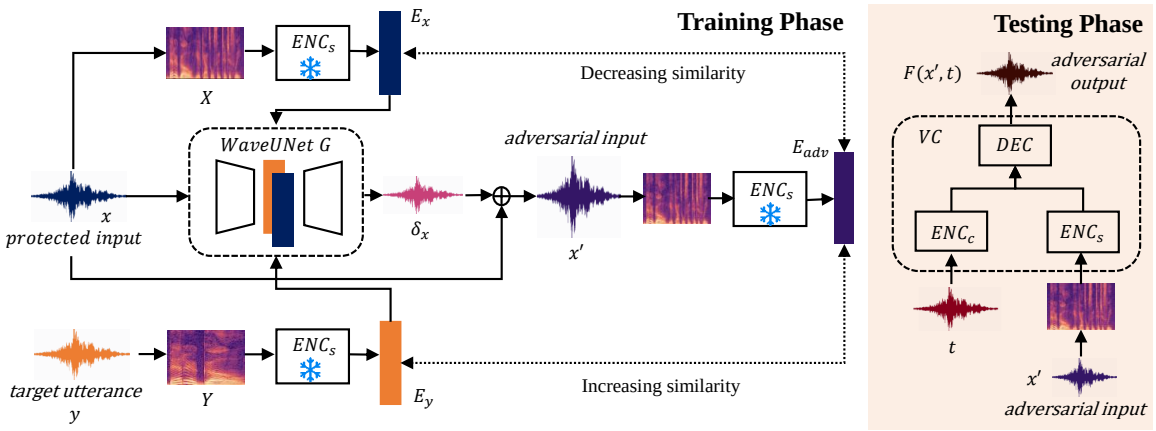

Figure 1: *Training and Testing phase of RW-VoiceShield. The dashed double arrow section represents the target of the loss update.*

2. Methodology

2.1. Dataset

The dataset is divided into two parts: one part is utilized for adversarial training of the generative model, while the other part is used for retraining the speaker encoder under the *black-box* scenario. For the first part, we initially divide 108 speakers from CSTR VCTK Corpus dataset [29] into two groups based on their gender. Each group is then further merged into the training, validation, and testing sets in a ratio of 3:1:1. The second part involves merging all the speech data from CSTR VCTK [29], VoxCeleb1 [30], and VoxCeleb2 [31].

2.2. Proposed Model

Our proposed method is shown in Figure 1. We select Free-VC [7], an advanced one-shot VC model, to assess the effectiveness of RW-VoiceShield in preventing voice replication by VC models. Like other common one-shot VC models, Free-VC adopts an encoder-decoder architecture at the testing phase in Figure 1. In this setup, the encoder includes a speaker encoder ENC_s to capture the speaker's traits from the utterance x, along with a content encoder ENC_c to extract content details from the utterance t. These output embeddings, $ENC_s(x)$ and $ENC_c(t)$, are then processed by the decoder to produce sentences that mimic the speaker's characteristics of x while preserving the content from t. Our goal is to alter the output of ENC_s to make it dissimilar from the embedding of the protected speaker.

During the training process, shown as Figure 1, we make modifications to the WaveUNet proposed by Stoller *et al.* [32], a raw waveform-based generative model, denoted as G, to serve as our base model. In the forward propagation phase, for each waveform of an utterance to be protected, denoted as x or *protected input*, we randomly select another utterance from a different speaker, denoted as y or *target utterance*. We compute the mel-spectrogram of x and y, input them into ENC_s, and obtain speaker embeddings, E_x and E_y, note that the ENC_s here is pretrained and frozen, meaning it does not participate in updates. These embeddings are then concatenated with a hidden layer of G. G takes three inputs, namely, x, E_x, and E_y, to generate the adversarial perturbation, denoted as δ_x. We center δ_x around zero by subtracting its mean along the time axis, as observed in experiments that without this step, G tended to produce a positive constant value, directly added to x. While such attacks might succeed, they are vulnerable to countermeasures and easily invalidated. We scale δ_x so that when added to x, the SNR equals a noise constraint constant z. Finally, we obtain the waveform resulting from the mixture of x and δ_x, denoted as x' or *adversarial input*. Here, "input" refers to the input of the VC model during the testing phase as shown in Figure 1.

Our training objective is to minimize the similarity between the speaker embedding of the adversarial input, E_{adv}, and embedding of the protected input, E_x, while maximizing the similarity between E_{adv} and E_y. By doing so, when E_{adv} is input into the decoder of the VC model and synthesized, the resulting voice will be distinguishable from the voice of x's speaker. This ensures the protection of the speaker characteristics of x from being replicated by the VC model. We utilize mean square error (MSE) and cosine similarity (CS) as loss functions, as shown in Equation 1 and 2, and note that there is a positive-valued hyperparameter λ balancing the importance between the two loss terms. Pseudocode is provided in Algorithm 1.

$$L_{CS} = 1 - cos(E_{adv}, E_x) + \lambda \cdot \max(0, cos(E_{adv}, E_y)) \quad (1)$$

$$L_{MSE} = -MSE(E_{adv}, E_x) + \lambda \cdot MSE(E_{adv}, E_y)) \quad (2)$$

2.3. Attack Scenarios

Here, we explore two distinct scenarios. In one scenario, the attacker has unrestricted access to the target model, including full knowledge of its architecture and trained parameters. This allows for direct application of adversarial attacks, commonly referred to as the *white-box* scenario. In this scenario, we utilize the pretrained speaker encoder provided by Free-VC as the ENC_s component of RW-VoiceShield, ensuring consistency with the ENC_s used during the testing phase. The second scenario, known as the *black-box* scenario, involves the attacker lacking direct access to the parameters of the target model. We make slight modifications to the model by adjusting the hidden layer size and subsequently train a new speaker encoder, using it as the ENC_s during the training phase. However, during the testing phase, we maintain consistency by using the ENC_s provided by Free-VC. Note that different ENC_s are employed for the two phases. We compare the effectiveness of attacks in the two scenarios in subsequent analyses.

3. Experiment Results and Analysis

3.1. Experimental Settings and Evaluation Metrics

In all experiments, the Adam optimizer is utilized with an initial learning rate of 0.001. Additionally, a cyclic learning rate is applied with a maximum learning rate of 0.001 and a min-

Algorithm 1 Training procedure of RW-VoiceShield

Require: Training pairs $D = \{(x_1, y_1), ..., (x_n, y_n)\}$,Speaker Encoder ENC_s, noise constraint parameter z, constant weight λ

Ensure: Trained generative model $G(\cdot)$

for number of iterations **do**
  for number of steps **do**
    $x, y \leftarrow$ minibatch of m pairs from D
    $E_x \leftarrow ENC_s(Mel(x))$
    $E_y \leftarrow ENC_s(Mel(y))$
    $\delta_x \leftarrow G(x, E_x, E_y)$
    $\delta_x \leftarrow \delta_x - Mean(\delta_x)$
    $\delta_x \leftarrow \frac{RMS(x)}{RMS(\delta_x) \cdot 10^{\frac{z}{20}}} \cdot \delta_x$
    $E_{adv} \leftarrow ENC_s(x + \delta_x)$
    Loss $\leftarrow 1 - \cos(E_{adv}, E_x)$
      $+\lambda \cdot \max(0, \cos(E_{adv}, E_y))$
    minimize Loss to update $G(\cdot)$
  end for
end for

imum learning rate of 0.00001. The batch size was set to 4, and the noise constraint parameter z is initially set to 20, while λ is set to 0.1. We conduct evaluations every 10,000 steps and implement early stopping with a patience of 5. All results are calculated using the toolkit [33] at a 95% confidence level and recorded in Table 1. To explore the relationship between the noise constraint parameter z and the model's performance, when the early stopping criteria are met, the model is saved, and the value of z is incremented by 1 before continuing training. This process continues for z values ranging from 20 to 25, and the results are recorded in Table 2. To evaluate the performance of the proposed idea we use the attack success rate (ASR) and preservation success rate (PSR) over 1050 test sets. In addition, we estimate the PESQ, SNR, and STOI of the adversarial input as metrics for comparing perceptual quality.

We reimplement the embedding attack proposed by Huang *et al.*[25] as our baseline and test it in the *white-box* scenario. The difference between this approach and RW-VoiceShield is that it applies adversarial perturbation to the spectrogram of the x instead of the waveform. This perturbation is a trainable tensor updated based on the mean square error between the E_x and E_y. The spectrogram with the added perturbation is then synthesized into adversarial utterance using the Griffin-Lim algorithm [34]. We test this approach on our testing dataset and present the results in Table 1.

3.2. Objective Tests

We adopt an automatic speaker verification (ASV) model, utilizing the pretrained ECAPA-TDNN [35], to assess whether the attack is successful and whether the adversarial noise affects the speaker characteristics of the protected input. It encodes two input utterances into embeddings and computes their similarity. Utterances surpassing a predefined threshold are considered to be uttered by the same speaker. The threshold is determined based on the equal error rate (EER) when verifying randomly sampled utterance pairs from the VCTK corpus. Specifically, we select 256 utterances for each speaker in the dataset and divide them equally into positive and negative samples. Positive samples were compared against random utterances from the authentic speaker, while negative samples were compared against random utterances from other randomly selected speakers. This yielded a threshold of 0.328, with an EER of 0.027.

Table 1: *Performance comparison with baseline. The abbreviations represent different configurations of the proposed model, "CS" and "MSE" denote the types of loss utilized, and "w" and "b" indicate* white-box *and* black-box *scenarios. The values within parentheses represent the lower and upper bounds of the confidence interval.*

Model	ASR	PSR	SNR	PESQ	STOI
B-w [25]	0.53 (0.46, 0.61)	0.99 (0.98, 1.00)	4.01 (3.92, 4.06)	1.61 (1.56, 1.67)	0.77 (0.76, 0.78)
P-CS-w	**0.94** (0.90, 0.98)	0.82 (0.76, 0.88)	19.98 (19.97, 19.99)	1.79 (1.74, 1.84)	0.81 (0.79, 0.83)
P-MSE-w	0.70 (0.61, 0.79)	**0.99** (0.97, 1.00)	19.99 (19.98, 20.00)	1.97 (1.91, 2.01)	**0.84** (0.82, 0.86)
P-CS-b	0.85 (0.83, 0.87)	0.95 (0.92, 0.98)	20.03 (20.01, 20.05)	1.90 (1.86, 1.94)	0.81 (0.79, 0.83)
P-MSE-b	0.74 (0.64, 0.83)	0.97 (0.94, 0.99)	19.98 (19.97, 19.99)	**1.99** (1.93, 2.03)	0..82 (0.80, 0.84)

Table 2: *Performance comparison of the proposed P-CS-b model under different SNR variations.*

SNR	ASR	PSR	PESQ	STOI
20	0.84	0.95	1.90	0.81
21	0.80	0.97	1.98	0.82
22	0.78	0.98	2.05	0.84
23	0.73	0.99	2.17	0.85
24	0.70	1.00	2.28	0.87
25	0.66	1.00	2.40	0.88

Among the 21 speakers in our test dataset, we randomly select 50 utterances from each speaker as the protected input x. To evaluate the effectiveness of our attacks, we pair each x with a randomly chosen utterance t from a different speaker and a target utterance y from a speaker of the opposite gender of x. This test set, denoted as (x, y, t), must meet a specific condition: the pair $(F(x, t), x)$ should pass the ASV system. Here, $F(x, t)$, denoted as *original output*, represents the synthesized speech obtained by encoding x with a speaker encoder and t with the content encoder of FreeVC. This condition is imposed to ensure that before the attack, the VC successfully synthesizes speech that closely resembles the speaker characteristics of x. Subsequently, We feed (x, y) into RW-VoiceShield and baseline model to generate adversarial input x'.

Next, we feed x' and t into the ENC_s and ENC_c of the VC, shown as the testing phase in Figure 1, to obtain $F(x', t)$, denoted as the *adversarial output*. In the ideal scenario, for the pair $(F(x', t), x)$, we expect it to fail to pass the ASV system, indicating the success of our attack and effectively protecting the voice of x from voice cloning. Furthermore, for the pair (x', x), we expect it to pass the ASV system, demonstrating that even after adding adversarial perturbations, we are able to successfully preserve the speaker characteristics of x. Table 1 shows all the performance results.

Table 1 summarizes the results: B-w for the baseline model in *white-box* scenario, P-CS-w and P-MSE-w for the proposed model with CS and MSE loss in *white-box* scenario, P-CS-b and P-MSE-b in *black-box* scenario. In the *white-box* scenario, both P-CS-w and P-MSE-w achieve higher ASR at a significantly higher SNR compared to the baseline model. For instance, the ASR of RW-VoiceShield is 0.94 at an SNR of 19.98, while that of the baseline model is 0.53 at an SNR of 4.01. Additionally, P-CS-w exhibits a higher PESQ at 1.79, compared to 1.61 for the baseline model, and a better STOI at 0.81, compared to 0.77 for the baseline model. The significant performance gap arises due to the two-stage issue of the baseline model. As the input

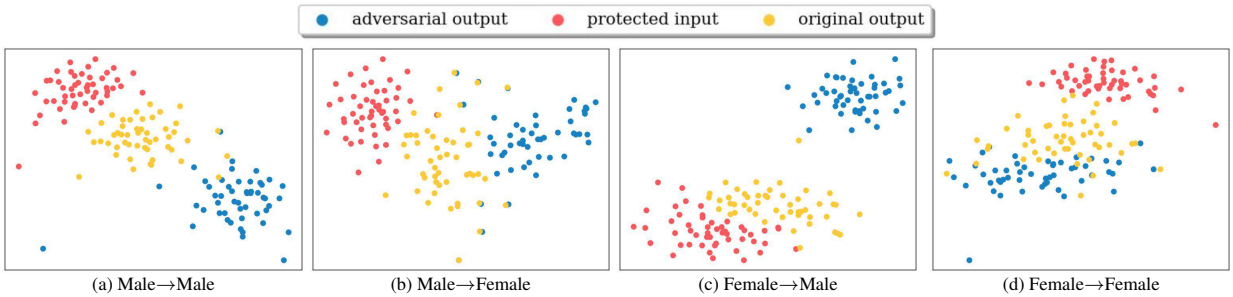

Figure 2: *t-SNE analysis of objective tests, the title indicates the gender of speaker of x and gender of speaker of y respectively.*

of ENC_s consists of low-resolution spectrograms, the baseline model can only apply perturbations to these low-resolution spectrograms in the *white-box* scenario. However, the perturbed spectrograms cannot be properly reconstructed by the vocoder, leading to ineffective attacks and poor perceptual quality. This highlights the advantage of RW-VoiceShield, which conducts adversarial attacks directly on the waveform, thereby avoiding the issues caused by the two-stage process in the baseline. Additionally, the experimental results indicate that in the *black-box* scenario, although the ASR is slightly lower than in the *white-box* scenario, P-CS-b still reaches 0.85. Comparing the two types of loss, CS loss yields a slightly higher ASR, but the PSR is slightly lower compared to MSE loss. We also analyze the relationships between ASR, PSR, PESQ, and SNR in Table2, revealing that as SNR increases, ASR decreases while PSR and PESQ increase. This can be explained by the higher probability of successful attacks with the addition of more perturbations, while also experiencing a certain loss in perceptual quality and the speaker's characteristics of the protected input.

We conduct four sets of experiments using P-CS-b, exploring combinations of speakers' genders for x and y. For each set, we generate 50 test sets (x, y, t), with x and y selected randomly from distinct speakers. The protected input x, along with the generated adversarial output and original output, are input into the ASV model to calculate embeddings. We then utilize t-SNE to analyze their distributions, as depicted in Figure 2. Our analysis reveals distinct clusters, with the distributions of protected inputs and original outputs being relatively close, indicating effective VC replication before the attack. However, adversarial outputs display a notable separation from original outputs, with their distance from original inputs even greater, suggesting RW-VoiceShield effectively distinguishes the distribution of adversarial outputs from the original ones.

3.3. Subjective Tests

We conduct subjective evaluations using P-CS-w and P-CS-b. We randomly select two protected inputs from 21 test speakers, resulting in 42 sets (x, y, t). Each set comprise three pairs of utterances, where one was the protected input, and the others were either the adversarial input, adversarial output, or original output. Participants were instructed to determine whether two given utterances originated from the same speaker by choosing one of four options: (I) Different, absolutely sure; (II) Different, but not very sure; (III) Same, but not very sure; and (IV) Same, absolutely sure. Each pair was evaluated by 6 subjects, and to mitigate outliers, we excluded two extreme ballots on both ends for each pair. The percentages of ballots are depicted in Figure 3.

The results align with the objective test. Initially, 72.62% of original outputs are successfully replicated by the VC, perceived as from the same speakers as the protected inputs. Following our attack only left 11.9% and 20.84% of adversarial outputs in both scenarios are identified as from the same speaker as the protected inputs. Notably, 75% and 61.9% of respondents in both scenarios believed that adversarial outputs definitely came from different speakers than the protected inputs. This demonstrates our success in neutralizing the VC's ability to mimic the speaker characteristics of the protected inputs. Bars (b) and (d) reveal that 97.02% and 98.81% of the adversarial inputs were recognized as originating from the same speaker as the protected utterances. This suggests that adversarial inputs preserve the speaker characteristics of the protected inputs. For a demonstration and access to the source code, visit `https://github.com/gino0950150/RW_VoiceShield`.

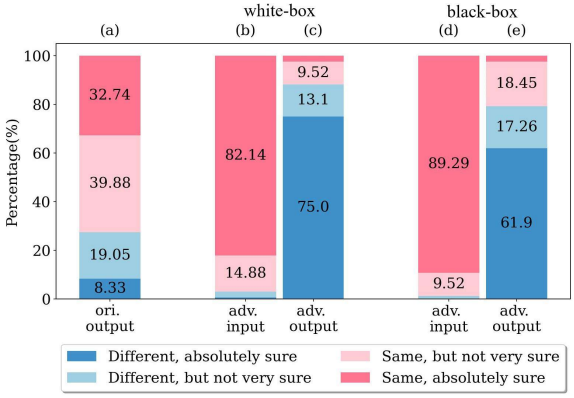

Figure 3: *Subjective tests result with RW-VoiceShield.*

4. Conclusion

We introduce a novel approach to attacking VC by adding subtle noise to protected utterances, hindering one-shot VC's ability to replicate speaker characteristics. Using a waveform-based generative model, we apply perturbations directly in the time domain, avoiding the two-stage challenge of previous methods. We train a general model to infer adversarial noise for different inputs, eliminating the need to retrain for each protected utterance. Evaluated on state-of-the-art one-shot VC models in both *black-box* and *white-box* scenarios, our method significantly reduces the VC's ability to mimic voices while preserving the original speaker characteristics. This is supported by objective speaker verification tests and subjective evaluations. Future research will focus on enhancing the model's generalizability to effectively attack various VC models.

5. References

[1] J. Zhang, Z. Jiang, D. Yang, H. Xu, Y. Shi, G. Song, Z. Xu, X. Wang, and J. Feng, "Avatargen: a 3d generative model for animatable human avatars," in *European Conference on Computer Vision*. Springer, 2022, pp. 668–685.

[2] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen *et al.*, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4563–4573.

[3] J. Oppenlaender, "The creativity of text-to-image generation," in *Proceedings of the 25th International Academic Mindtrek Conference*, 2022, pp. 192–202.

[4] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4834–4844.

[5] Y. Li, C. Ma, Y. Yan, W. Zhu, and X. Yang, "3d-aware face swapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 705–12 714.

[6] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577.

[7] 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.

[8] Y. Yang, Y. Kartynnik, Y. Li, J. Tang, X. Li, G. Sung, and M. Grundmann, "Streamvc: Real-time low-latency voice conversion," *arXiv preprint arXiv:2401.03078*, 2024.

[9] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[10] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," *Advances in neural information processing systems*, vol. 36, 2024.

[11] Y. A. Li, C. Han, and N. Mesgarani, "Styletts-vc: One-shot voice conversion by knowledge transfer from style-based tts models," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 920–927.

[12] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," *arXiv preprint arXiv:2304.09116*, 2023.

[13] N. Shah, M. Singh, N. Takahashi, and N. Onoe, "Nonparallel emotional voice conversion for unseen speaker-emotion pairs using dual domain adversarial network & virtual domain pairing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] G. Zhang, Y. Qin, W. Zhang, J. Wu, M. Li, Y. Gai, F. Jiang, and T. Lee, "iemotts: Toward robust cross-speaker emotion transfer and control for speech synthesis based on disentanglement between prosody and timbre," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[15] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, "Deepsinger: Singing voice synthesis with data mined from the web," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1979–1989.

[16] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 10, 2022, pp. 11 020–11 028.

[17] Y.-J. Luo, C.-C. Hsu, K. Agres, and D. Herremans, "Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3277–3281.

[18] S. Zhao, H. Wang, T. H. Nguyen, and B. Ma, "Towards natural and controllable cross-lingual voice conversion based on neural tts model and phonetic posteriorgram," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5969–5973.

[19] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario." in *Interspeech*, 2020, pp. 4382–4386.

[20] C. Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," *The Wall Street Journal*, 2022.

[21] X. Tian, R. K. Das, and H. Li, "Black-box Attacks on Automatic Speaker Verification using Feedback-controlled Voice Conversion," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 159–164.

[22] Z. Ye, T. Mao, L. Dong, and D. Yan, "Fake the Real: Backdoor Attack on Deep Speech Classification via Voice Conversion," in *Proc. INTERSPEECH 2023*, 2023, pp. 4923–4927.

[23] H. Ilyas, A. Javed, and K. M. Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," *Applied Soft Computing*, vol. 136, p. 110124, 2023.

[24] A. Hamza, A. R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, and R. Borghol, "Deepfake audio detection via mfcc features using machine learning," *IEEE Access*, vol. 10, pp. 134 018–134 028, 2022.

[25] 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.

[26] Z. Liu, Y. Zhang, and C. Miao, "Protecting your voice from speech synthesis attacks," in *Proceedings of the 39th Annual Computer Security Applications Conference*, 2023, pp. 394–408.

[27] 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.

[28] 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, no. 16, 2021, pp. 14 129–14 137.

[29] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," sound, 2019.

[30] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[31] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[32] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[33] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals

[34] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[35] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 3830–3834.