



Mimic Blocker: Self-Supervised Adversarial Training for Voice Conversion Defense with Pretrained Feature Extractors

Gwangyeol Yu*, Junhyeok Lee*, Seoryeong Kim*, Jimin Lee, Jehyuk Lee†

Department of AI, Big Data & Management, Kookmin University, Republic of Korea

{rhkdduf627, jh2020, tjfud0216, dlwlals922, jehyuk.lee}@kookmin.ac.kr

Abstract

Voice conversion (VC) enables natural speech synthesis with minimal data; however, it poses security risks, e.g., identity theft and privacy breaches. To address this, we propose Mimic Blocker, an active defense mechanism that prevents VC models from extracting speaker characteristics while preserving audio quality. Our method employs adversarial training, an audio quality preservation strategy, and an attack strategy. It relies on only publicly available pretrained feature extractors, which ensures model-agnostic protection. Furthermore, it enables self-supervised learning using only the original speaker's speech. Experimental results demonstrate that our method achieves robust defense performance in both white-box and black-box scenarios. Notably, the proposed approach maintains audio quality by generating noise imperceptible to human listeners, thereby enabling protection while retaining natural voice characteristics in practical applications.

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

1. Introduction

Recent advancements in deep learning have furthered the development of voice conversion (VC) technology significantly [1, 2, 3, 4]. Currently, VC technology is widely employed in various applications, including voice conversion services [5, 6, 7] and audio data augmentation [8]. However, the increasing sophistication of VC models introduces critical risks, particularly the generation of audio deepfakes, and such misuse can lead to serious concerns, including identity theft and privacy breaches [9].

Among VC models, one-shot VC models have garnered attention because these models can synthesize high-quality speech using only a single sample of a target speaker's voice. One-shot VC effectively retains the linguistic content of the content speech while adapting the timbre to match the target speaker. In addition, recent advancements have further enhanced the capability of one-shot VC models, enabling natural and highly realistic VC with minimal input data [10, 11].

To mitigate the risks associated with VC technology, previous studies have investigated audio deepfake detection and post-hoc defense mechanisms [12, 13, 14, 15, 16]. However, these approaches primarily focus on identifying the converted speech rather than proactively preventing the conversion process. Consequently, these methods are passive and offer limited fundamental protection. In addition, research on dedicated defense mechanisms against VC models remains insufficient,

and many existing methods struggle to balance defense effectiveness with sufficient audio quality. While some approaches incorporate specific VC models during training to enhance defense performance [17, 18, 19, 20], their generalizability to unseen VC models remains uncertain, and many methods require both style and target speech during training, thereby imposing structural constraints that increase data requirements.

To address these challenges, we propose Mimic Blocker, an active defense mechanism against VC models. The Mimic Blocker leverages adversarial training and a quality-preserving strategy in the waveform domain to maintain audio quality while maximizing defense effectiveness. By utilizing publicly available feature extractors [21, 22] and applying attack strategies at the embedding level, the proposed method ensures model-agnostic defense. This design enables the Mimic Blocker to maintain consistent defense effectiveness across various VC models without depending on any specific architectures. Unlike previous methods [17, 18, 19, 20], the proposed Mimic Blocker achieves defense effectiveness through self-supervised learning using only a single style speech, while maintaining attack strength by directly adding noise to the original style waveform.

In addition, the adversarial noise introduced by the Mimic Blocker is imperceptible to human listeners while preventing speaker identity features from being captured by VC models, thus providing effective protection. Through this research, we aim to prevent the misuse of VC technology and contribute to the secure and ethical advancement of voice conversion techniques

2. Proposed Methodology

The primary objective of the Mimic Blocker is to prevent one-shot VC models from replicating a speaker's unique vocal characteristics. To achieve this, adversarial noise is added to the style waveform, which serves as input to the VC model. Note that the generated noise must be sufficiently subtle to be imperceptible to human listeners while effectively preventing the VC model from extracting the speaker's characteristics. Figure 1 shows the overall framework of the proposed Mimic Blocker, which comprises three key components, i.e., the generator, the discriminator, and the feature extractor. The proposed method is built on three core strategies, including adversarial training, a quality preservation strategy, and an attack strategy.

2.1. Adversarial Training

The adversarial training process is designed to generate perturbations that effectively interfere with VC models while maintaining the naturalness of the speech, which is achieved through the interplay between the generator and the discrimina-

*These authors contributed equally.

† corresponding author

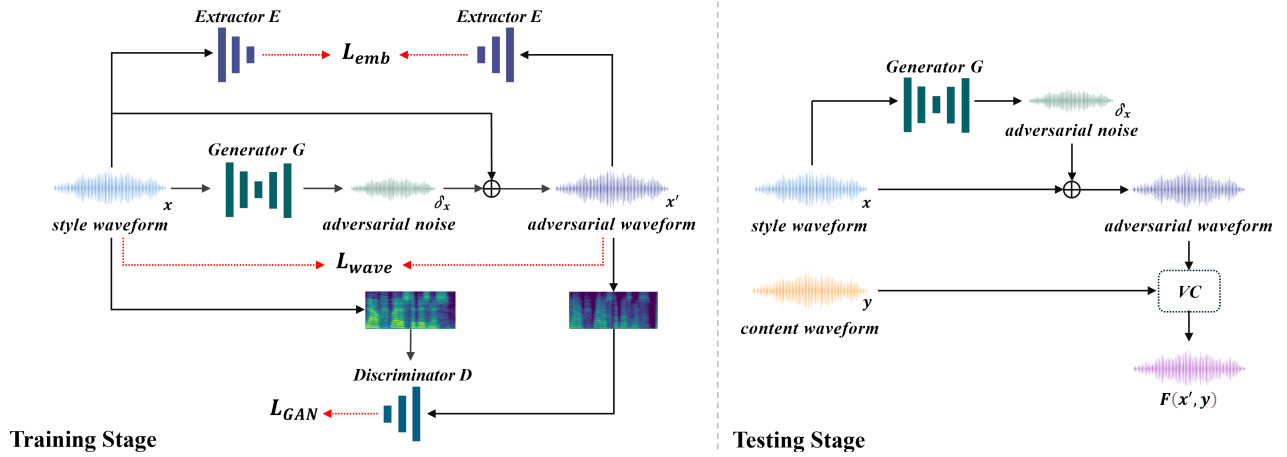


Figure 1: Training and testing stage of Mimic Blocker. The red arrows represent the computational pathways used to calculate the loss functions.

tor. Specifically, the generator creates adversarial noise, and the discriminator learns to distinguish between clean and perturbed waveforms.

Generator : The generator receives the style waveform x as input and generates noise to hinder the VC model. The generator is composed of 1D-convolution and transposed convolution layers, with the final output restricted to the range $[-1, 1]$ using the hyperbolic tangent activation function. Then, combine the noise with the original style waveform to produce a new waveform.

Discriminator : The discriminator operates in the Mel-spectrogram domain to differentiate between the original style waveform x and the adversarial waveform x' . It consists of convolutional layers and fully connected layers, ultimately producing a classification score using a sigmoid activation function. To enhance the stability of the adversarial training, the final linear layer of the discriminator is initialized dynamically for each batch, and an adaptive balance between the discriminator and the generator is maintained.

The generator and discriminator are trained jointly using the following loss functions.

$$x' = x + G(x) \quad (1)$$

$$L_{GAN} = \mathbb{E}[\log(D(\text{mel}(x)))] + \mathbb{E}[\log(D(\text{mel}(x')))] \quad (2)$$

In Equation (1), x denotes the style waveform used as the input to the VC model, $G(\cdot)$ denotes the generator, and x' denotes the waveform with added noise. In Equation (2), $D(\cdot)$ denotes the output of the discriminator, and $\text{mel}(\cdot)$ denotes the Mel-spectrogram transformation.

2.2. Quality Preservation Strategy

Two methods are implemented to ensure the imperceptibility of the noise and preserve the audio quality. First, we introduced wave loss, which minimizes the mean squared error (MSE) between the original style waveform and the adversarial waveform, which helps ensure that the generated noise remains imperceptible to human listeners. The wave loss is defined as follows.

$$L_{wave} = \text{MSE}(x, x') \quad (3)$$

Second, previous studies have applied adversarial attacks in the Mel-spectrogram domain, subsequently reconstructing the waveform using a vocoder. However, this technique has notable weaknesses, including degraded audio quality and reduced defense effectiveness due to the distortions introduced during the spectrogram-to-waveform conversion process [19]. In contrast, the proposed method introduces noise directly into the waveform domain. This approach preserves the audio quality while simultaneously maximizing the defense effectiveness against VC models.

2.3. Attack Strategy

In the proposed method, we also introduce an embedding loss to enhance the generator's ability to obstruct the extraction of the speaker's characteristic by VC models. This approach maximizes the L2 distance between the embedding vectors of the original style waveform and the adversarial waveform, the proposed method effectively disrupts the VC models' ability to extract the speaker's characteristics.

$$L_{emb} = - \| E(x) - E(x') \|_2 \quad (4)$$

Here, $E(\cdot)$ denotes the pretrained feature extractor, where $E(x)$ denotes the embedding vector of the original style waveform and $E(x')$ denotes the embedding vector of the adversarial waveform. By maximizing the embedding distance between the original style waveform and the adversarial waveform, the proposed method effectively disrupts the VC models' ability to extract the speaker's characteristics.

Previous studies have applied attacks by minimizing the distance between the original style waveform and a designated target waveform [17, 18, 20]; however, the proposed method employs a self-supervised learning attack technique that separates the embeddings of the original style waveform and the adversarial waveform without requiring additional target waveforms. The training objective of the generator is defined through the following loss function.

$$L_{total} = \lambda_{wave}L_{wave} + \lambda_{emb}L_{emb} + L_{GAN} \quad (5)$$

This approach allows the noise to degrade the performance of the VC model while preserving the natural audio qual-

Table 1: Performance comparison of different methods in FreeVC.

Model	ASR \uparrow	PSR \uparrow	PESQ \uparrow	STOI \uparrow
DYV_w [20] (Defending Your Voice White-box)	0.530 (0.460, 0.610)	0.990 (0.980, 1.000)	1.610 (1.560, 1.670)	0.770 (0.760, 0.780)
RW_cs_w [20] (RW-VoiceShield Cosine Similarity White-box)	0.940 (0.900, 0.980)	0.820 (0.760, 0.880)	1.790 (1.740, 1.840)	0.810 (0.790, 0.830)
RW_mse_w [20] (RW-VoiceShield MSE White-box)	0.700 (0.610, 0.790)	0.990 (0.970, 1.000)	1.970 (1.910, 2.010)	0.840 (0.820, 0.860)
RW_cs_b [20] (RW-VoiceShield Cosine Similarity Black-box)	0.850 (0.830, 0.870)	0.950 (0.920, 0.980)	1.900 (1.900, 1.940)	0.810 (0.790, 0.830)
RW_mse_b [20] (RW-VoiceShield MSE Black-box)	0.740 (0.830, 0.870)	0.970 (0.940, 0.990)	1.990 (1.930, 2.030)	0.820 (0.800, 0.840)
P_wavlm_w (Proposed method wavlm White-box)	0.883 (0.862, 0.901)	1.000 (0.996, 1.000)	3.633 (3.633, 3.661)	0.942 (0.939, 0.945)
P_hubert_b (Proposed method hubert Black-box)	0.880 (0.859, 0.898)	1.000 (0.996, 1.000)	3.633 (3.633, 3.661)	0.942 (0.939, 0.945)

ity. Compared with existing methods, the proposed method demonstrates superior defense capabilities through a simpler design while remaining independent of specific VC models [17, 18, 19, 20]. Here, λ_{wave} and λ_{emb} are the hyperparameters that balance the contributions of the wave loss and the embedding loss, respectively. These values were set to $\lambda_{wave} = 0.8$ and $\lambda_{emb} = 0.2$ following a manual search, where we selected the configuration that minimized the loss function.

2.4. Inference

During the inference process, the trained generator adds optimized noise to the given style waveform. The generated adversarial noise is designed to effectively obstruct the VC model’s extraction of the speaker’s characteristics while preserving audio quality that is similar to that of the original style waveform. As shown in the right portion of Figure 1, the trained generator generates the optimal adversarial noise δ_x for a given the original style waveform input x . The generated noise δ_x is injected into the original style waveform to form the adversarial waveform x' , and the resulting x' serves as input to the VC model, ultimately hindering accurate preservation of the speaker’s style in the converted waveform $F(x', y)$.

3. Experiment

This study used the CSTR VCTK Corpus 0.92, which comprises 110 speakers [23]. We randomly partitioned the dataset into training (60%), validation (20%), and test (20%) sets while ensuring gender balance. Additionally, we resampled all audio recordings to 16 kHz and normalized them to maintain consistent volume levels. We trained the proposed Mimic Blocker model using the Adam optimizer with a batch size of 32 and a learning rate of 1e-4 for 10 epochs. The training process on an NVIDIA RTX 4090 GPU required approximately 25 hours for completion. The inference time is less than 1 second per sample. We evaluated the proposed Mimic Blocker in terms of the perceptual evaluation of audio quality (PESQ) [24], short-time objective intelligibility (STOI) [25], attack success rate (ASR), and preservation success rate (PSR) [20]. For speaker verification, we utilized the ECAPA-TDNN model [26], which was pre-trained on the VoxCeleb dataset [27, 28], using the same cosine similarity threshold (0.328) as RW-VoiceShield for comparative analysis [20].

3.1. Attack Scenarios

To evaluate the robustness of the proposed method, we investigated its defense performance in both white-box and black-box attack scenarios. In previous studies, the white-box scenario was defined as a scenario in which the target VC model’s speaker encoder was trained on existing data. Conversely, in the black-box scenario, the speaker encoder was retrained with new data before testing [17, 20]. The proposed method employs a publicly available pretrained feature extractor rather than the speaker encoder from a specific VC model; thus, the standard definitions from these previous studies could not be applied directly in this experimental evaluation.

Therefore, the following modifications were introduced to align the attack scenarios with the proposed method. In the white-box scenario, we used the publicly available version of WavLM, which has the same architecture as the speaker encoder used in the FreeVC. In the black-box scenario, we conducted attacks with the HuBERT feature extractor, which differs from the speaker encoder of FreeVC. This setting allowed us to analyze the impact of the encoder variations on the attack performance. In addition, the experimental design enabled a comprehensive evaluation of the defense effectiveness of the proposed Mimic Blocker under diverse conditions while ensuring generalizability across different speaker encoders.

3.2. Objective Tests

The experimental results shown in Table 1 demonstrate that the proposed Mimic Blocker outperformed the existing defense approaches while maintaining superior audio quality. For the comparative analysis, we considered two baseline methods, i.e., the Defending Your Voice (DYV) [17] and RW-VoiceShield (RW) [20] methods.

The abbreviations in the “Model” column of Table 1 represent the model architecture and experimental conditions. DYV, RW, and P denote the utilized models, where P refers to the “proposed model.” CS and MSE indicate the loss functions used for training RW. Lastly, w and b correspond to the white-box and black-box scenarios, respectively. Values in parentheses represent the 95% confidence interval lower and upper bounds. For methods other than the proposed model, the results were referenced from the literature [20] to facilitate an effective comparison with existing methods.

As Table 1 demonstrates, P_wavlm_w and P_hubert_b

achieved ASR of 0.883 and 0.880 respectively, outperforming all existing methods except RW_cs_w [20]. This result indicates that the proposed Mimic Blocker prevents VC models from extracting the speaker’s characteristics more effectively than existing methods. Additionally, while maintaining consistent ASR regardless of the feature extractors, the proposed method showed stable performance in both white-box and black-box experiments, whereas the existing methods exhibited large variations in specific scenarios. In contrast, the proposed Mimic Blocker demonstrated stable performance in both the white-box and black-box experiments.

Furthermore, the proposed Mimic Blocker significantly outperformed existing methods in terms of the audio quality metrics, including PSR, PESQ, and STOI. Notably, the proposed method achieved PESQ 3.633 and STOI of 0.942, which are substantially higher than those obtained by the previous models. These results indicate that the proposed method maintains excellent performance in both defense effectiveness and audio quality. Unlike existing methods that reduce audio quality to enhance adversarial robustness, the proposed method maintains both defense effectiveness and perceptual quality.

Table 2: Performance of TriAAN-VC.

Model	ASR \uparrow	PSR \uparrow	PESQ \uparrow	STOI \uparrow
P_wavlm	0.992 (0.985, 0.996)	1.000 (0.996, 1.000)	3.632 (3.602, 3.661)	0.941 (0.938, 0.943)
P_hubert	0.990 (0.981, 0.994)	1.000 (0.996, 1.000)	3.632 (3.602, 3.661)	0.941 (0.938, 0.943)

Table 2 shows the experimental results in the TriAAN-VC environment. P_wavlm and P_hubert achieved ASR of 0.992 and 0.990 respectively. These results indicate that the proposed Mimic Blocker maintains consistent defense performance across different VC models, demonstrating its robustness and applicability to various voice conversion architectures.

In summary, the proposed Mimic Blocker effectively maximizes the defense performance while preserving the audio quality. Notably, its stable performance across both white-box and black-box scenarios, coupled with its robust defense against diverse VC models, is a key factor in enhancing the practicality of the proposed Mimic Blocker method.

3.3. Subjective Tests

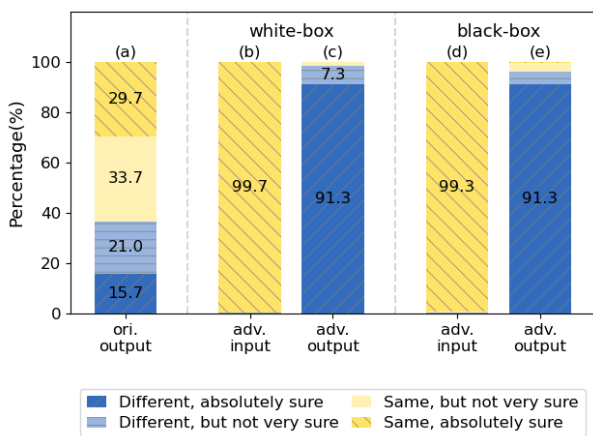


Figure 2: Subjective tests result with Mimic Blocker.

To further assess the effectiveness of the proposed Mimic Blocker, a subjective evaluation was performed involving 21 test speakers (9 male, and 12 female participants). For the male speakers, three utterances were selected randomly from seven participants, and two utterances were selected from the remaining two participants. For the female speakers, three utterances were selected from a single participant, and two utterances were selected randomly from each of the remaining eleven participants, resulting in a total of 50 evaluation sets.

Here, each set comprised five pairs based on a single style waveform: (a) the style waveform and the original output, (b) the style waveform and the adversarial waveform (white-box), (c) the style waveform and the adversarial output (white-box), (d) the style waveform and the adversarial waveform (black-box), and (e) the style waveform and the adversarial output (black-box). In addition, a panel of six evaluators was tasked with determining whether the speakers in each speech pair were the same. The participants selected from the following four confidence-based options: (I) different, absolutely sure; (II) different, but not very sure; (III) same, but not very sure; and (IV) same, absolutely sure.

The corresponding evaluation results are shown in Figure 2 as percentage values. This subjective evaluation also yielded high performance. For example, in the white-box and black-box scenarios ((b) and (d)), the adversarial waveform was perceived as belonging to the same speaker as the style waveform with confidence levels of 99.7% and 99.3%, respectively. In contrast, in scenarios where the style waveform was compared with the adversarial output ((c) and (e)), 91.3% of the participants confidently identified them as different speakers. These findings confirm that the proposed protection method successfully disrupts VC while preserving the speaker’s vocal characteristics. The demonstration and source code are available at <https://github.com/yugwangyeol/Mimic-Blocker>.

4. Conclusion

This paper proposes the Mimic Blocker, a robust speech protection approach designed to prevent VC models from replicating a speaker’s vocal characteristics. Unlike conventional defense techniques, the proposed method does not rely on specific VC models by utilizing a publicly available feature extractor and does not require target speech during training.

Mimic Blocker employs an adversarial training model that operates directly in the waveform domain, providing strong defense performance. The experimental results demonstrate that the proposed method consistently preserved its protective effectiveness across both white-box and black-box scenarios, exhibiting minimal performance variability and demonstrating high generalization capability to various VC models.

Notably, the proposed method maintains high defense performance while preserving the original speaker’s audio quality, setting it apart from existing techniques. Mimic Blocker retains audio quality more effectively than previous defense methods while significantly degrading the conversion performance of VC models. This demonstrates its ability to overcome the common trade-off in adversarial defenses, where audio quality is typically compromised, thereby offering practical and high-quality speech protection.

The findings of this study suggest that the proposed Mimic Blocker method can serve as a robust voice protection mechanism, potentially preventing social issues caused by the misuse of VC technology at an early stage.

5. References

- [1] S. Kovala, R. Valle, A. Dantrey, and B. Catanzaro, "Any-to-any voice conversion with F0 and timbre disentanglement and novel timbre conditioning," in *Proc. ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [2] M. Chu *et al.*, "E-DGAN: An encoder-decoder generative adversarial network based method for pathological to normal voice conversion," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, pp. 2489–2500, May 2023.
- [3] X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Voice conversion with denoising diffusion probabilistic GAN models," in *Proc. International Conference on Advanced Data Mining and Applications*, Nanjing, China, Sep. 2023, pp. 154–167.
- [4] S. Dhar, N. D. Jana, and S. Das, "GLGAN-VC: A guided loss-based generative adversarial network for many-to-many voice conversion," *IEEE Transactions on Neural Networks and Learning Systems*, Dec. 2023, early access.
- [5] 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 *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Diego, USA, Aug. 2020, pp. 1979–1989.
- [6] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "DiffSinger: Singing voice synthesis via shallow diffusion mechanism," in *Proc. AAAI Conference on Artificial Intelligence*, Vancouver, Canada, Feb. 2022, pp. 11 020–11 028.
- [7] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, Jun. 2021, pp. 5969–5973.
- [8] S. Shah Nawazuddin, 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 *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 4382–4386.
- [9] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, "Hello, it's me: Deep learning-based speech synthesis attacks in the real world," in *Proc. ACM SIGSAC Conference on Computer and Communications Security*, Seoul, Korea, Nov. 2021, pp. 235–251.
- [10] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [11] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "TriAAN-VC: Triple adaptive attention normalization for any-to-any voice conversion," in *Proc. ICASSP 2023 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [12] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, and N. Evans, "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," *arXiv preprint arXiv:2109.00537*, Sep. 2021.
- [13] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Proc. INTERSPEECH*, Brno, Czechia, Sep. 2024, pp. 4259–4263.
- [14] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *Proc. ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, Jun. 2021, pp. 6369–6373.
- [15] X. Li, N. Li, C. Weng, S. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with Res2Net architecture," in *Proc. ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, Jun. 2021, pp. 6354–6358.
- [16] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake speech detection through emotion recognition: A semantic approach," in *Proc. ICASSP 2022 – IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, May 2022, pp. 8962–8966.
- [17] C. Huang, Y. Y. Lin, H. Lee, and L. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proc. IEEE Spoken Language Technology Workshop*, Shenzhen, China, Jan. 2021, pp. 552–559.
- [18] J. Li, D. Ye, L. Tang, C. Chen, and S. Hu, "Voice Guard: Protecting voice privacy with strong and imperceptible adversarial perturbation in the time domain," in *Proc. International Joint Conference on Artificial Intelligence*, Macao, China, Aug. 2023, pp. 4812–4820.
- [19] S. Dong, B. Chen, K. Ma, and G. Zhao, "Active defense against voice conversion through generative adversarial network," *IEEE Signal Processing Letters*, 2024.
- [20] 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*, Kos Island, Greece, Sep. 2024, pp. 2730–2734.
- [21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, and Z. Chen, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022. [Online]. Available: <https://drive.google.com/file/d/12-cB34qCTvByWT-QtOcZaqwwO21FLSqU/view>
- [22] W. N. Hsu, B. Bolte, Y. H. H. Tsai, K. Lakhota, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: <https://huggingface.co/facebook/hubert-large-ls960-ft>
- [23] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," p. 15, 2017.
- [24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 2001, pp. 749–752.
- [25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.
- [26] B. Desplanques, J. Thienpondt, and K. Demuyne, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 3830–3834.
- [27] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 2616–2620.
- [28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1086–1090.