

Diffusion-Based Adversarial Purification for Speaker Verification

Yibo Bai , Xiao-Lei Zhang , and Xuelong Li , *Fellow, IEEE*

Abstract—Recently, automatic speaker verification (ASV) based on deep learning is easily contaminated by adversarial attacks, which is a new type of attack that injects imperceptible perturbations to audio signals so as to make ASV produce wrong decisions. This poses a significant threat to the security and reliability of ASV systems. To address this issue, we propose a Diffusion-Based Adversarial Purification (DAP) method that enhances the robustness of ASV systems against such adversarial attacks. Our method leverages a conditional denoising diffusion probabilistic model to effectively purify the adversarial examples and mitigate the impact of perturbations. DAP first introduces controlled noise into adversarial examples, and then performs a reverse denoising process to reconstruct clean audio. Experimental results demonstrate the efficacy of the proposed DAP in enhancing the security of ASV and meanwhile minimizing the distortion of the purified audio signals.

Index Terms—Adversarial defense, diffusion model, speaker verification.

I. INTRODUCTION

AUTOMATIC speaker verification (ASV) aims to verify individuals based on their unique voiceprint characteristics. It has been widely used in biometric authentication. However, ASV systems are vulnerable to attackers [1], which raises a concern in enhancing their security in real-world applications. Particularly, recently a new type of attacker named *adversarial attack*, which adds imperceptible perturbations to original utterances, can easily contaminate an ASV system by making it e.g. accept speakers that should have been rejected or just the opposite, without changing the perception quality of the utterances to human. The polluted utterances are called *adversarial examples*, while the original utterances are also called *genuine examples*. In recent years, there has been growing interest in studying the susceptibility of ASV systems to adversarial attacks. For example, [1], [2] found that the state-of-the-art ASV models are highly vulnerable to adversarial attacks. [3] conducted transferable gray-box attacks on ASV systems across different features and different models.

Manuscript received 24 October 2023; revised 22 May 2024; accepted 16 June 2024. Date of publication 24 June 2024; date of current version 11 September 2024. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yu Tsao. (*Corresponding author: Xiao-Lei Zhang.*)

Yibo Bai is with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong (e-mail: baiyibo@connect.hku.hk).

Xiao-Lei Zhang is with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, also with the Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd., Beijing 100033, China, and also with the Research and Development Institute, Northwestern Polytechnical University, Shenzhen 518063, China (e-mail: xiaolei.zhang@nwpu.edu.cn).

Xuelong Li is with the Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd., Beijing 100033, China (e-mail: xuelong_li@ieee.org).

Digital Object Identifier 10.1109/LSP.2024.3418715

In response to this emerging threat landscape, researchers have begun investigating techniques to enhance the robustness and security of ASV systems against adversarial attacks. Current defense methods can be classified into three categories: adversarial training, adversarial detection and adversarial purification [4]. Specifically, adversarial training mainly utilizes the adversarial examples to retrain the ASV model. One weakness of this kind of methods is that it needs to modify the parameters of the original model [5]. Adversarial detection adds a detection head in front of the ASV model to reject adversarial examples as an input into the system [6]. However, it may hinder human access to ASV when his/her voice was polluted by adversarial perturbations. Adversarial purification aims to purify all incoming inputs to eliminate adversarial perturbations [4]. It overcomes the weaknesses of the first two kinds of methods, which is our focus in this paper.

Adversarial purification for ASV can be divided into two categories: preprocessing and reconstruction. Preprocessing methods apply empirical knowledge to the input signals. They are typically data-free and have low computational complexity. For example, [7] applied median, mean and Gaussian filters to the input utterance. [8] proposed to add white noise with different variance to the entire utterance. Reconstruction methods focus on recover the original audio or its acoustic features from the adversarial examples. [9] proposed a separation network to estimate adversarial noise for restoring the clean speech. [10] proposed to reproduce the acoustic features with a self-supervised model. Although existing purification methods have demonstrated effectiveness in defending ASV systems, the quality of the reconstructed audio signals was not guaranteed to high level. Some methods introduce additional noise into the purified samples, while others produce unexpected distortion to the audio signals which make the signals deviate significantly from their origins.

To address the above issue, we propose the Diffusion-Based Adversarial Purification (DAP) method. This novel method utilizes a diffusion model to purify the impact of adversarial attacks by reconstructing the original speech waveform, which defends ASV systems with high reliability. Our contributions can be summarized as follows: We propose the first adversarial defense diffusion model for ASV systems. Our method achieves the state-of-the-art performance on the ASV purification task, and retains the essential information of the original speech signal.

II. RELATED WORK

A. Automatic Speaker Verification

Speaker verification aims to determine whether a test utterance belongs to a speaker that it declares to. Most of the current ASV systems comprise three components: an acoustic

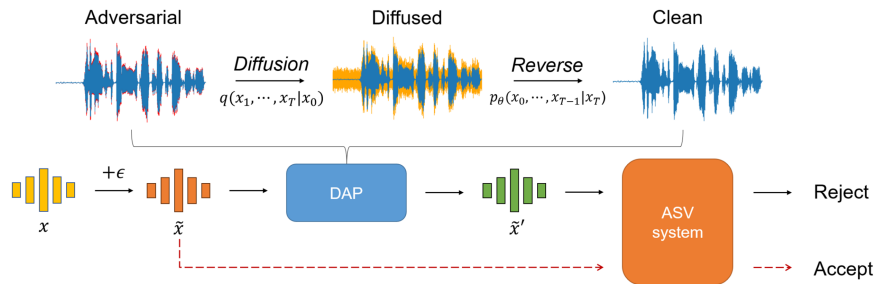


Fig. 1. A speaker verification pipeline with DAP method. Initially, the adversarial example is introduced into a diffusion model positioned before the ASV system for processing. Subsequently, the diffusion model employs a “diffusion” process on the adversarial input, followed by the reversal of this process to reconstruct the original clean audio. Finally, the ASV system produces the correct verification outcome.

TABLE I
ECAPA-TDNN PERFORMANCE RESULTS FOR GENUINE EXAMPLES ON THE
SPEAKER VERIFICATION TASK

Trials	EER(%)	minDCF
VoxCeleb-O	1.069	0.107
VoxCeleb-E	1.201	0.131
VoxCeleb-H	2.288	0.226
1,000 trials	0.828	0.021

feature extractor, an encoder front-end which yields a speaker embedding from the acoustic features, and a scoring back-end which evaluates the similarity of the representations of two utterances. Commonly used acoustic features include Mel-frequency cepstral coefficients (MFCCs) or logarithmic filter-banks (LogF-Bank). Given a test utterance x^t and an enrollment utterance x^e , the scoring process of ASV can be defined as:

$$s = S(F(x^t), F(x^e)) \quad (1)$$

where $S(\cdot)$ denotes the scoring back-end, $F(\cdot)$ is the encoder front-end, and s is the similarity score between x^t and x^e . By comparing the similarity score with a predefined threshold, the system determines whether to accept the test utterance.

B. Adversarial Attack to ASV

Given a genuine audio utterance x from a speaker i , an adversarial attack creates a perturbation signal ϵ to x . The adversarial example is formulated as $\tilde{x} = x + \epsilon$ subject to the condition $\|\tilde{x} - x\|_p \leq \epsilon$ which guarantees that \tilde{x} is similar to x , where ϵ is a very small number that controls the energy of ϵ , and $\|\cdot\|_p$ is the ℓ_p -norm. As shown in Fig. 1, \tilde{x} aims to cause an error of the ASV system.

C. Denoising Diffusion Probabilistic Models

The denoising diffusion probabilistic models [11] are a type of generative models used to produce data similar to the input. Specifically, the diffusion model works by progressively adding Gaussian noise to blur the training data and then learning how to denoise it and recover the original input. Once trained, the diffusion model can reverse the diffusion process to generate new data from random noise.

Recently, diffusion models have garnered interest among researchers. They utilize diffusion and denoising processes for

high-quality content generation. By incorporating specific generation conditions, the outcomes of diffusion models can be controlled [12] to satisfy different applications such as speech enhancement, speech command recognition, image reconstruction, and remote sensing [13], [14], [15], [16].

III. METHODOLOGY

A. Framework

The objective of our research is to develop a robust adversarial purification model $D(\cdot)$ for ASV, which eliminates the perturbation ϵ in \tilde{x} and produces a purified audio \tilde{x}' . It can be formulated as a problem of $\tilde{x}' = D(\tilde{x})$ subject to:

$$S(\tilde{x}', x^e) = S(x, x^e) \quad (2)$$

This paper proposes to take the denoising diffusion probabilistic model [11] as $D(\cdot)$ to purify adversarial perturbation and transform these adversarial examples into clean data for ASV. The proposed DAP method is illustrated in Fig. 1. Given an adversarial audio \tilde{x} , DAP purifies it to \tilde{x}' for satisfying (2). Specifically, DAP first introduces noise to \tilde{x} via a *forward process* with a diffusion timestep T . Subsequently, it reconstructs the clean audio signal \tilde{x}' via a reverse denoising process before feeding it into the ASV system for analysis. In the next subsection, we will present the denoising diffusion probabilistic model in detail.

B. Diffusion-Based Audio Purification

A T -step denoising diffusion probabilistic model consists of two processes: the diffusion process, a.k.a. *forward process*, and the reverse process, which can both be represented as a T -step parameterized Markov chain. In its diffusion process, the model adds T rounds of noise to the real example x_0 to obtain the noised sample x_T . The reverse process aims to recover the original x_0 based on x_T . Following the Markov assumption, the state at t step in the diffusion process, only depends on the state at $t - 1$ step, so the process can be defined as:

$$q(x_1, \dots, x_T | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}), \quad (3)$$

where $q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t \mathbf{I})$. In other words, x_t is sampled from a Gaussian distribution with mean $\sqrt{1 - \beta_t}x_{t-1}$ and variance β_t , where β_t is a hyperparameter

determined by a predefined strategy, usually satisfying $\beta_1 < \beta_2 < \dots < \beta_T$.

Then, employing the recursive reparameterization trick, x_t can be represented in terms of x_0 :

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + (1 - \bar{\alpha}_t)\epsilon, \quad (4)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. We have:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (5)$$

Similarly, for the reverse process that transforms x_T back to x_0 , we have:

$$p_\theta(x_0, \dots, x_{T-1}|x_T) = \prod_{t=1}^T p_\theta(x_{t-1}|x_t), \quad (6)$$

where $p_\theta(\cdot)$ is used to estimate $q(\cdot)$ in (3) and $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 \mathbf{I})$ with the parameterized μ_θ and σ_θ^2 described as [17]:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t) \right), \quad (7)$$

and

$$\sigma_\theta(x_t, t)^2 = \tilde{\beta}_t, \quad (8)$$

where $\tilde{\beta}_t = \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t$ for $t > 1$ and $\tilde{\beta}_1 = \beta_1$, and $\epsilon_\theta(x_t, t)$ represents a deep neural network model used to predict Gaussian noise ϵ from x_t and t .

According to [11], we train diffusion model with the following unweighted objective function:

$$\mathbb{E}_{x_0, t, \epsilon} \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{(1 - \bar{\alpha}_t)}\epsilon, t) \right\|_2^2, \quad (9)$$

where t is uniformly sampled from the range 1 to T . After training, ϵ_θ is able to predict ϵ well. In the inference stage, the diffusion model first begins with a diffused sample x_T , then the reverse process iteratively utilizes ϵ_θ to get the mean μ_θ and finally recover x_0 from x_T .

Existing theorems have proven that in the forward process (3) of the diffusion model, the KL divergence between the distribution of clean data and the distribution of adversarial examples monotonically decreases [18]. This indicates that the two distributions gradually become more similar as t increases, enabling the use of the reverse process to reconstruct clean inputs from adversarial examples. This comprehensive process constitutes the foundation of our innovative defense approach against adversarial attacks. While the previous method in [18] focuses on purifying images with fixed width and height, our method can flexibly handle audio signals with variable length. Given an adversarial audio \tilde{x} , we initiate the forward process with $x_0 = \tilde{x}$. According to the above theorem in [18], there exists a timestep t^* that minimizes the KL divergence between the two distributions. Therefore, by starting the reverse process with $T = t^*$, the diffusion model can recover corresponding clean audio of \tilde{x} . The resulting audio utterance is then passed to the ASV system.

IV. EXPERIMENTS

A. Experimental Settings

1) *Dataset*: We utilized VoxCeleb1 [19] for evaluating our DAP approach and VoxCeleb2 [20] for training the ASV

and DAP models. VoxCeleb1 contains over 1,000 hours of speech data with 148,642 utterances from 1,211 speakers. VoxCeleb2 extends VoxCeleb1 with more speakers. This dataset offers an expanded set of development data to train our ASV models, which enables them to learn from a more diverse set of speakers. To balance computational requirements and ensure a representative evaluation, we conducted our adversarial research by randomly selecting 1,000 trials from the VoxCeleb1-O subset.

2) *ASV System*: We employed ECAPA-TDNN [21] as the main victim ASV model for adversarial attacks, which consists of convolutional layers with 512 channels. We used the AAM-Softmax objective function [22] with hyperparameters $\{s=32, m=0.2\}$ for training, along with attentive statistical pooling. The input acoustic feature is an 80-dimensional LogFBank representation with a 25 ms hamming window and a 10 ms step size. Additionally, cepstral mean and variance normalization (CMVN) is applied to the features. Data augmentation techniques including speed perturbing, superimposed disturbance, and reverberation enhancement are employed. Cosine distance is used to produce similarity scores between embeddings.

3) *Adversarial Attack*: We employed PGD attack [23] and BIM attack [24], which adopt the same parameters $\{\epsilon=30, \alpha=1\}$ to generate the adversarial examples. The iteration step was normally set as 50. To ensure a consistent signal-to-noise ratio (SNR) between genuine and adversarial examples, we added Gaussian white noise to the genuine examples, with the noise level determined by the corresponding adversarial perturbations. As a result, the mean signal-to-noise ratio for the genuine examples was set to approximately 40 dB.

4) *Evaluation Metrics*: We measured the performance of the ASV systems and the effectiveness of the defense mechanism using two commonly used metrics: Equal Error Rate (EER) and minimum Detection Cost Function (minDCF) with $p = 0.01$ and $C_{miss} = C_{fa} = 1$ [25]. EER measures the point at which the false acceptance rate (FAR) equals the false rejection rate (FRR). minDCF is a cost-based metric that considers both false acceptance and false rejection errors, allowing for a more comprehensive evaluation of ASV system performance.

We evaluated the reconstruction performance of the purified signals using three objective metrics: Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Short-Time Objective Intelligibility (STOI), and Perceptual Evaluation of Speech Quality (PESQ). These metrics assess the quality and intelligibility of the audio, where higher values indicate better speech quality.

5) *Baseline and Proposed Methods*: For our Diffusion-Based Adversarial Purification (DAP) model, we adopted the same architecture as DiffWave [17]. To introduce controlled noise during the training stage, we set step T as 100 and utilized a linear noise schedule $\{0.0001, 0.035\}$ to apply β_t at each step. It makes β_t begin with 0.0001 and increase to 0.035 over 100 steps. In the inference stage, the six-step variance schedule $[0.0001, 0.001, 0.01, 0.05, 0.2, 0.35]$ is applied to set the value of γ in the fast sampling algorithm. In addition, We used adversarial examples generated by the PGD method for training the DAP model. We trained two DAP systems conditioned on 512-dimensional spectrogram, which were trained by 1 k and 80 k iterations, respectively.

We compared the proposed adversarial defense method with three adversarial defense methods [7], [8], [10]. (i) The TERA method [10] employed a self-supervised model to reconstruct

TABLE II

EER(%) RESULTS OF THE VICTIM ASV MODEL FOR GENUINE AND ADVERSARIAL EXAMPLES, GIVEN THE DEFENSE MODELS OF TERA, SPATIAL SMOOTHING, ADDING NOISE AND THE PROPOSED DAP METHOD

	N/A	TERA [10]		Spatial smoothing [7]			Adding noise [8]					DAP (proposed)	
		1*TERA	9*TERA	Median	Mean	Gaussian	$\sigma=0.002$	$\sigma=0.005$	$\sigma=0.01$	$\sigma=0.02$	$\sigma=0.05$	iter=1k	iter=80k
genuine	0.828	20.890	35.818	27.853	1.161	10.973	1.161	1.547	2.277	3.675	9.731	2.505	1.253
adv-PGD	91.683	48.356	42.360	27.557	88.935	26.722	23.188	5.609	3.727	4.762	9.524	3.340	7.087
adv-BIM	92.133	26.499	39.545	26.087	76.789	13.872	6.576	2.901	2.484	3.868	9.731	2.321	2.070

The term "N/A" means that no defense model is applied.

TABLE III

QUALITY OF THE AUDIO SIGNALS THAT ARE FIRST GENERATED FROM PGD ADVERSARIAL EXAMPLES AND THEN PURIFIED BY DEFENSE MODELS

Defender	SI-SDR	STOI	WB-PESQ	NB-PESQ
N/A	35.099	0.991	4.412	4.397
Median filter [7]	-16.359	0.632	1.152	1.397
Adding Noise [8]	12.572	0.867	1.556	2.353
DAP(proposed)	11.467	0.932	3.120	3.717

the acoustic feature. We pretrained it with the same setting in [10] except on 80-dim LogFbank features. (ii) The spatial smoothing method [7] used median, mean, and Gaussian filters to process the input audio. (iii) The noise-based method [8] added Gaussian noise to the entire audio signal. In our experiments, the standard deviation of the noise was set to $\{0.002, 0.005, 0.01, 0.02, 0.05\}$.

B. Experimental Results

1) *Defense Performance*: Table II compares the DAP method with three purification methods on the ASV defense task. DAP outperforms others in defending against adversarial examples while maintaining performance on genuine examples. Furthermore, DAP is capable of defending against both ℓ_∞ and ℓ_2 attacks. In addition, although self-supervised models like TERA are not trained for adversarial purification tasks, they could still demonstrate some defensive capabilities. We also evaluated another pretrained self-supervised model—WavLM [26]. It reaches an EER of 9.427% after purifying the PGD examples.

2) *Reconstruction Performance*: In terms of audio reconstruction quality, DAP surpasses other methods as shown in Table III. Fig. 2 gives a visualized comparison between the original audio and its corresponding adversarial examples after processed by different defense methods. From the figure, we see that the adversarial example purified by DAP are observed to be more similar to the original signal compared to those purified by the other approaches, e.g. the highlighted part in the red box. Moreover, DAP can reduce the noise and reverberation component of the speech signal. We also conducted a human listening test on 50 DAP-purified samples with five listeners. In 52% of the cases, listeners could not discern any differences, while 36% of the samples were perceived as similar to the original audio but with slight noise. For the remaining 12%, listeners noticed significant noise.

3) *Effect of Attack Settings*: In this subsection, we study the robustness of the proposed method against different attackers and with different victim models. Table IV lists the performance of the ECAPA-TDNN and Fast-ResNet34 victim models under the ℓ_2 PGD and ℓ_∞ BIM attack methods, where the standard

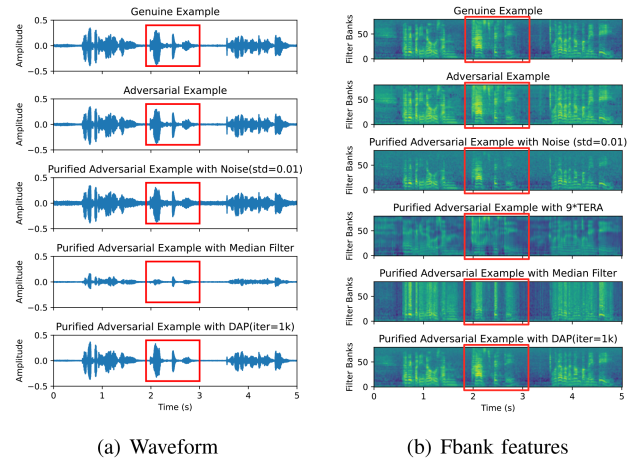


Fig. 2. A comparison example between the original audio and its adversarial example with different defenders. The genuine example is from id10270/5r0dWxy17C8/00024.wav of VoxCeleb1. As TERA method focuses on feature-level purification, it is not included in Fig. 2(a).

TABLE IV

EER(%) RESULTS OF ECAPA-TDNN AND FAST-RESNET34 UNDER PGD AND BIM ATTACK METHODS

Attacker	Steps	ECAPA-TDNN			Fast-ResNet34		
		N/A	Noise	DAP	N/A	Noise	DAP
PGD	10	83.232	3.340	3.758	85.339	2.714	2.128
	20	95.633	3.704	4.115	96.378	3.132	2.277
	50	98.668	4.983	4.293	98.992	3.833	2.899
	100	98.998	5.010	4.528	99.398	3.967	2.901
BIM	10	59.872	3.311	3.520	62.128	2.505	1.852
	20	84.258	3.448	3.868	87.628	2.714	2.244
	50	94.821	3.643	4.348	95.899	3.132	2.996
	100	95.834	3.883	4.348	96.604	3.292	3.125

deviation of the adding noise method was set to 0.01. The results show that our method is effective under different attack scenarios and different ASV architectures.

V. CONCLUSION

In this letter, we propose a DAP method for the ASV defense against adversarial attacks. DAP utilizes a diffusion model to purify the adversarial examples and mitigate the perturbations in audio inputs. We conducted experiments in scenarios where the attackers are unaware of the defense method. The experimental results indicate that our approach outperforms the representative purification methods. It also introduces the minimal distortion to the genuine examples over the comparison methods.

REFERENCES

- [1] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Proc. INTERSPEECH*, 2020, pp. 4233–4237.
- [2] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6579–6583.
- [3] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1962–1966.
- [4] H. Wu, J. Kang, L. Meng, H. Meng, and H.-Y. Lee, "The defender's perspective on automatic speaker verification: An overview," 2023, *arXiv:2305.12804*.
- [5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2014.
- [6] X. Li et al., "Investigating robustness of adversarial samples detection for automatic speaker verification," in *Proc. INTERSPEECH 2020*.
- [7] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense against adversarial attacks on spoofing countermeasures of asv," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6564–6568.
- [8] L.-C. Chang, Z. Chen, C. Chen, G. Wang, and Z. Bi, "Defending against adversarial attacks in speaker verification systems," in *Proc. IEEE Int. Perform., Comput., Commun. Conf.*, 2021, pp. 1–8.
- [9] H. Zhang, L. Wang, Y. Zhang, M. Liu, K. A. Lee, and J. Wei, "Adversarial separation network for speaker recognition," in *Proc. INTERSPEECH*, 2020, pp. 951–955.
- [10] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-Y. Lee, "Improving the adversarial robustness for speaker verification by self-supervised learning," *IEEE/ACM Trans Audio, Speech Lang. Process.*, vol. 30, pp. 202–217, 2022.
- [11] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.
- [12] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denosing diffusion probabilistic models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14347–14356.
- [13] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 7402–7406.
- [14] W. Yu, Y. Xu, and P. Ghamisi, "Universal adversarial defense in remote sensing based on pre-trained denoising diffusion models," 2023, *arXiv:2307.16865*.
- [15] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," in *Proc. 11th Int. Conf. Learn. Representations*, 2022.
- [16] I. Alkhoury, S. Liang, R. Wang, Q. Qu, and S. Ravishanker, "Diffusion-based adversarial purification for robust deep MRI reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 12841–12845.
- [17] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [18] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16805–16827.
- [19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Proc. INTERSPEECH*, 2017.
- [20] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Proc. INTERSPEECH*, 2018.
- [21] B. Desplanques, J. Thienpondt, and K. Demuyne, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc.*, 2020, pp. 3830–3834.
- [22] Y. Liu, L. He, and J. Liu, "Large margin softmax loss for speaker verification," in *Proc. INTERSPEECH*, 2019.
- [23] L. Ruff et al., "A unifying review of deep and shallow anomaly detection," *Proc. IEEE Proc. IRE*, vol. 109, no. 5, pp. 756–795, May 2021.
- [24] S. Joshi, J. Villalba, P. ZČelasko, L. Moro-Velázquez, and N. Dehak, "Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4811–4826, 2021.
- [25] A. Nagrani et al., "Voxsrc 2020: The second voxceleb speaker recognition challenge," 2020, *arXiv:2012.06867*.
- [26] S. Chen et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.