MDD: a Mask Diffusion Detector to Protect Speaker Verification Systems from Adversarial Perturbations

Yibo Bai*, Sizhou Chen†, Michele Panariello*, Xiao-Lei Zhang‡ § ¶, Massimiliano Todisco* and Nicholas Evans*
*Digital Security Department, EURECOM, France
†School of Computer Science, The University of Sydney, Australia
‡School of Marine Science and Technology, Northwestern Polytechnical University, China
§Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd., China
¶Research and Development Institute of Northwestern Polytechnical University in Shenzhen, China

***Abstract*—Speaker verification systems are increasingly deployed in security-sensitive applications but remain highly vulnerable to adversarial perturbations. In this work, we propose the Mask Diffusion Detector (MDD), a novel adversarial detection and purification framework based on a *text-conditioned masked diffusion model*. During training, MDD applies partial masking to Mel-spectrograms and progressively adds noise through a forward diffusion process, simulating the degradation of clean speech features. A reverse process then reconstructs the clean representation conditioned on the input transcription. Unlike prior approaches, MDD does not require adversarial examples or large-scale pretraining. Experimental results show that MDD achieves strong adversarial detection performance and outperforms prior state-of-the-art methods, including both diffusion-based and neural codec-based approaches. Furthermore, MDD effectively purifies adversarially-manipulated speech, restoring speaker verification performance to levels close to those observed under clean conditions. These findings demonstrate the potential of diffusion-based masking strategies for secure and reliable speaker verification systems.**

I. Introduction

Automatic Speaker Verification (ASV) plays a key role in providing secure access control to services, smart phones and devices. However, ASV systems are vulnerable to adversarial attacks [1]–[3] whereby subtle and even imperceptible perturbations are added to an acoustic input to manipulate normal system behaviour, i.e. to obtain unauthorised access to protected services or devices. Such vulnerabilities pose a challenge to verification reliability [4]–[6]. There is hence an interest to develop robust detection methods to protect against the threat of adversarial attacks.

Current research in defences against adversarial attacks falls into two categories [5], [6]. Proactive defences, such as adversarial training, requires advance knowledge of specific attacks and continual retraining/adaptation of the ASV model to new attacks [5]. Passive defences, in contrast, can be used to detect or eliminate adversarial perturbations without retraining. Among these, plug-in detection-based methods have received considerable attention due to their convenience and flexibility [5].

Various such detection schemes have been proposed recently. These include methods which rely on statistical analysis [7], [8], auxiliary classifiers like learnable mask networks [9], approaches based on self-supervised learning [10], and neural codec-based audio reconstruction [11]. However, these approaches often face limitations. For instance, some are too computationally intensive for real-time applications or require large-scale pretraining, while others are dependent on prior knowledge of specific attack types. This reliance can leave such systems vulnerable to new, unknown attacks [12], [13].

We propose a novel diffusion-based adversarial attack detector: the Mask Diffusion Detector (MDD). Our approach builds upon prior work in diffusion-based audio processing [14]–[16]. However, whereas the forward process of traditional diffusion models typically acts to progressively reduce the different between the input and Gaussian noise [17], [18], that of MDD is used to progressively produce a noised masked Mel-spectrogram. An embedding of the transcription text is applied during the conditional diffusion process to help preserve key information in the reconstruction. In the reverse denoising and reconstruction phase, the text-conditioned diffusion model learns to recover an estimate of the original, clean spectrogram. As a result, it can effectively identify and mitigate malicious, adversarial perturbations. Notably, MDD can be trained on bona fide data only, without any adversarial data. This attacker-independent approach promotes generalisation. Compared to existing works, the main contributions are as follows.

- We introduce MDD, a novel adversarial attack detection framework built upon a text-conditioned masked diffusion model. By applying spectral masking in both the forward and reverse diffusion processes, MDD effectively adapts diffusion-based generative modeling to the task of adversarial detection in speaker verification.
- MDD requires neither adversarial training data nor large-scale pretraining, yet it achieves strong detection performance, demonstrating robustness and generalisation across attack types.
- Beyond adversarial detection, MDD preserves the perfor-

Corresponding author: Yibo Bai (e-mail: bai@eurecom.fr)

This work is supported with funding received from the French Agence Nationale de la Recherche (ANR) via the P-SPIKE (ANR-23-CE39-0005), and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China under Grant GJHZ20240218114401004.

mance of speaker verification (ASV) systems on clean data, ensuring practical applicability in real-world scenarios without compromising verification accuracy.

II. Related work

A. Automatic speaker verification

A typical ASV system consists of two stages: speaker embedding extraction and speaker similarity computation. Embeddings are extracted from an enrolment utterance x_e and a test utterance x_t. The similarity between the pair of embeddings is then computed to determine whether the speaker in x_e matches that in x_t. Common speaker embedding extractors include x-vector [19], ECAPA-TDNN [20] and self-supervised learning-based models [21].

B. Adversarial attacks

Given a bona fide (clean) utterance x_t, an attacker generates an adversarially perturbed version x_{adv} which causes the ASV system to verify incorrectly the similarity between x_e and x_{adv}. The adversarial example x_{adv} is constrained to be perceptually similar to x_t so that $\|x_{adv} - x_t\|_p \leq \varepsilon$, where $|\cdot|_p$ denotes the ℓ_p-norm and ε is a small positive constant. Common attack methods include the basic iterative method (BIM) [22], projected gradient descent (PGD) [23], and the fast gradient sign method (FGSM) [24].

C. Diffusion models

Diffusion models have emerged recently as powerful generative frameworks in various domains. They operate by adding noise gradually to input data and then by learning a de-noising process so that they can generate high-quality, clean data from pure noise inputs. By approximating source distributions through such an iterative process, diffusion models have been shown to perform well for generative tasks [25], [26] and enhancement tasks [27], [28].

III. Mask diffusion detector

In this section we describe the proposed Mask Diffusion Detector (MDD), a novel framework designed for the detection of adversarial perturbations aimed at deceiving ASV systems. MDD consists of two main components: a diffusion model adapted with a specific mask strategy and noise schedule, and a back-end detector to distinguish bona fide from adversarially-manipulated speech data.

A. Diffusion model with mask strategy

An illustration of the MDD is shown in Fig. 1. At the core is a diffusion model which iteratively denoises the masked Mel-spectrogram. The model encompasses a forward diffusion process and a reverse reconstruction process.

1) Forward diffusion process: The forward diffusion process, denoted by $q(x_1, \ldots, x_T|x_0, c)$, takes an initial clean Mel-spectrogram x_0 and a text condition c, and progressively adds noise and applies masking over a sequence of T steps to produce x_T.

First, a masked version of the input, denoted as x_m, is generated from the original, clean spectrogram x_0. This is achieved through random masking of 16×16 patch regions.

Second, a composite noise target, N_{target}, is formulated according to:

$$N_{target} = x_m + \sigma \cdot \epsilon, \tag{1}$$

where ϵ is random noise sampled from a standard Gaussian distribution $\mathcal{N}(0, I)$ and with the same dimensions as x_m, and σ is a scalar coefficient used to control the signal-to-noise ratio in N_{target}.

The noisy spectrogram x_t for a given timestep t is subsequently generated according to a pre-defined noise schedule. It is generated from the initial clean spectrogram x_0, the new noise target N_{target}, and a noise schedule parameter $\bar{\alpha}_t$, which represents the cumulative product of noise variances up to step t. In our modified framework, the standard Gaussian noise term is replaced by our specifically formulated N_{target}. Thus, x_t can be expressed by the relation:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} N_{target}. \tag{2}$$

The forward diffusion hence steers x_0 towards a state which is a controlled combination of x_0 itself and the composite noise target N_{target}. Consequently, the masked input x_m becomes an integral part of the degradation process which the reverse reconstruction process must learn to invert.

2) Reverse reconstruction process: The reverse process, $p_\theta(x_{t-1}|x_t, c)$, aims to reconstruct the initial clean Mel-spectrogram x_0 from corrupt version x_t, with parameter θ and a text condition c. Given that the forward process uses N_{target} to noise and mask x_0, the model learns to denoise x_t as well as to fill the masked area in x_m. At each step, the training objective is to minimise the reconstruction error between the predicted noise and N_{target}. By learning the distribution of clean data, MDD is able to remove adversarial perturbations from masked inputs.

B. Waveform reconstruction

Following the reverse diffusion process, the denoised Mel-spectrogram $\hat{x}_0$ is transformed back into a time-domain waveform using a pretrained HiFi-GAN vocoder [29]. HiFi-GAN is a non-autoregressive, GAN-based neural vocoder that maps Mel-spectrogram features to high-fidelity audio waveforms. In our framework, the vocoder operates as a fixed post-processing module, independent of the diffusion model training. Its purpose is to synthesise speech from the reconstructed spectrogram $\hat{x}_0$ such that it can be evaluated by downstream automatic speaker verification (ASV) systems.

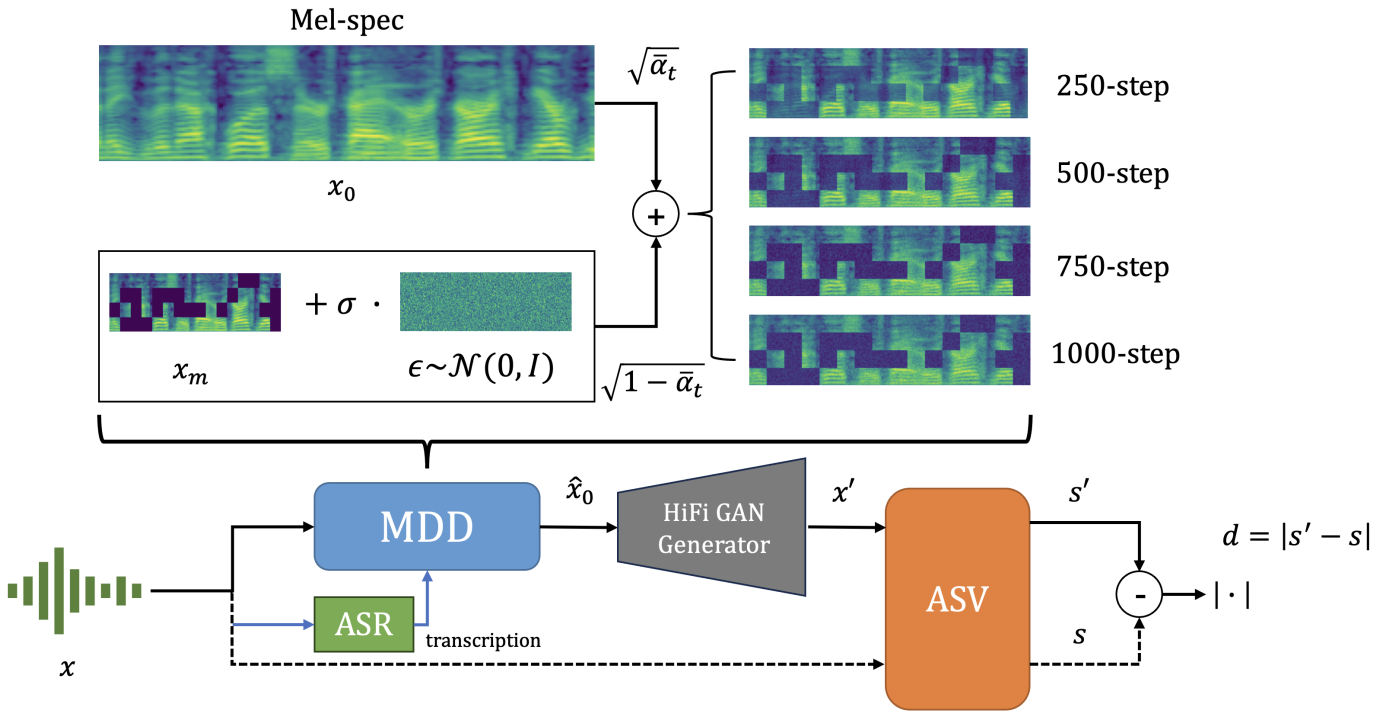

Fig. 1. An illustration of the workflow of the proposed MDD method. Given an input Mel-spectrogram, MDD first generates a masked version of the feature and mixes it with Gaussian noise. During the forward process, the mask diffusion model applies masking and noise to the input step by step. In the reverse process, it performs unmasking and denoising operations conditioned with the transcription. A HiFi-GAN vocoder then reconstructs the waveform from the processed feature. Finally, both the original and the regenerated waveforms are fed into the ASV system to compute the score difference.

C. Back-end adversarial detector

A straightforward back-end detector is used to determine whether an input audio sample is clean or whether it contains adversarial perturbations. The detector takes the form of a conventional ASV system which is applied separately to input signals x and diffusion-purified version x', thereby producing a pair of ASV scores s and s' respectively. The absolute score difference $d = |s - s'|$ is then calculated and thresholded for classification. For clean inputs, purification should produce $x' \approx x$ and $s' \approx s$, hence lower values of d. In the case of adversarial inputs, then purification should remove the perturbations which would otherwise act to compromise the ASV system. As a result, s' (purified of perturbations) should be lower than s (with perturbations) corresponding to comparatively higher values of d.

An input sample is classified as adversarial if d exceeds an empirically optimised detection threshold τ. We set τ_{det} from experiments involving a set $\mathbb{T}$ of clean data only (no adversarial examples) to achieve an arbitrarily-set target false positive rate FPR_{target} as follows:

$$\tau_{det} = \min_{\tau} \left\{ \frac{|\{x^i \in \mathbb{T} \mid d^i > \tau\}|}{\text{number of samples in } \mathbb{T}} \leq FPR_{\text{target}} \right\}, \quad (3)$$

where $\tau \in \mathbb{R}$ and d^i denotes the score difference for the i-th clean sample.

IV. Experiments

A. Experimental setup

We used the PGD method [23] and a subset of 1,000 clean utterances extracted from the VoxCeleb1 test dataset [30] to generate 1,000 adversarial examples with which to test the reliability of ECAPA-TDNN ASV and MDD systems in the white-box attack scenario. The input to the ECAPA-TDNN system is 80-dimensional log filterbank (LogFBank) components extracted with a 25 ms Hamming window and 10 ms frame shift. The model is trained using the VoxCeleb1 development set with the standard 512-channel architecture provided with the *Wespeaker* toolkit [31]. The PGD algorithm is applied with 50 attack iterations and ℓ_2-norm.

We evaluate detection performance using the Detection Rate (DR) metric, which quantifies the percentage of adversarial examples correctly identified by the system. Specifically, DR is measured at fixed false positive rate (FPR) thresholds, ensuring a controlled trade-off between detecting attacks and minimising false alarms on bona fide inputs. A higher DR indicates a stronger capability to detect adversarial perturbations without significantly impacting clean audio samples.

We trained six MDD models with 0%, 10%, 25%, 50%, 75% and 100% masked Mel-spectrograms. Each model is trained for 10,000 iterations with a batch size of 4 using the LibriSpeech train-clean-100 subset. In MDD, the noise control factor σ is set to 0.1, which we found to provide a good balance between mitigating adversarial effects and preserving the perceptual

fidelity of the output audio. Our implementation is based on the *audio-diffusion* toolkit[1], and follows a default 1000-step DDPM noise schedule [17]. We use the *whisper-small* Automatic Speech Recognition (ASR) model[2] for transcription during conditional generation, and *Stella*[3] to encode the text as conditional embeddings. A pretrained HiFi-GAN model [29] from AudioLDM[4] [26] serves as the vocoder in MDD.

B. Experimental results

The DR results reported in Table I show that the 10% masking configuration achieves the best detection performance across both FPR thresholds. Interestingly, the 0% masking case (i.e., unmasked input) performs slightly worse, highlighting the effectiveness of introducing partial spectral masking during diffusion. As the masking ratio increases beyond 10%, detection performance progressively declines due to greater information loss. These results suggest that moderate masking encourages the model to focus on key spectral regions relevant for detecting adversarial perturbations, while excessive masking degrades the model's ability to reconstruct meaningful features.

TABLE I
DR (%) RESULTS FOR THE MDD DEFENCE METHOD AT DIFFERENT MASK RATIOS.

Mask ratio	FPR=0.1	FPR=0.05
0% (unmasked)	96.2	95.0
10%	**98.0**	**96.9**
25%	96.0	94.0
50%	80.4	76.3
75%	60.6	57.3
100% (fully masked)	55.4	47.6

C. Comparison with other defence methods

We compare the adversarial detection performance of the proposed MDD with 10% masking against several existing defence methods, including the DAP diffusion model [15] and three neural codec-based approaches: AcademiCodec [32], SpeechTokenizer [33], and DAC [34]. The corresponding Detection Rate (DR) results under fixed false positive rate thresholds (FPR = 0.1 and 0.05) are reported in Table II.

The DAP method is based on a waveform-level diffusion model trained on unmasked audio using LibriSpeech, and fine-tuned on adversarial examples from the VoxCeleb1 development set. Although DAP achieves moderate DR values (78.0% and 71.7%), it falls short of the performance achieved by MDD, which attains 98.0% and 96.9% at the same FPR levels. Their score-difference distributions on bona fide and PGD adversarial data are shown in Fig. 2. This highlights the benefit of applying masking in the spectral domain and conditioning the reverse process with transcription in MDD.

[1]https://github.com/teticio/audio-diffusion
[2]https://huggingface.co/openai/whisper-small
[3]https://huggingface.co/NovaSearch/stella_en_400M_v5
[4]https://huggingface.co/cvssp/audioldm/tree/main/vocoder

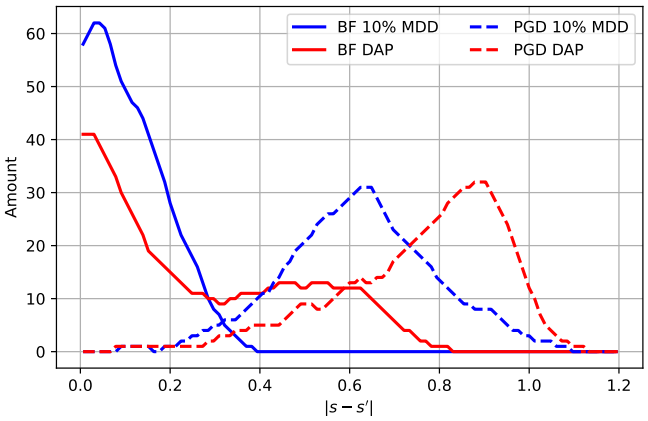

Fig. 2. The distributions of score differences when applying 10% MDD and DAP on bona fide (BF) and PGD adversarial data. A larger gap between the two lines of the same colour indicates that the PGD adversarial examples are more easily to be detected.

In contrast, the neural codec-based methods show substantially lower detection performance. While prior work has reported strong results using these codecs [11], they typically rely on large-scale pretraining, and their performance drops significantly when retrained under the same conditions as MDD. Specifically, AcademiCodec, SpeechTokenizer, and DAC achieve DRs of 58.0%, 65.7%, and 74.7% respectively at FPR=0.1, and even lower scores at FPR=0.05. These results suggest that neural codec-based detectors may lack generalizability and robustness when applied under constrained or mismatched training conditions.

Overall, the 10% masked MDD demonstrates state-of-the-art performance among all compared methods, benefiting from its text-conditioned reconstruction and partial masking strategy, which collectively enhance its ability to suppress adversarial perturbations while maintaining the fidelity of clean speech.

TABLE II
DR (%) COMPARISON BETWEEN THE 10% MASK MDD, DAP AND NEURAL CODEC-BASED METHODS UNDER FPR=0.1 AND FPR=0.05.

	Configuration	FPR=0.1	FPR=0.05
MDD	10% Mask	**98.0**	**96.9**
DAP [14]	-	78.0	71.7
AcademiCodec [32]	16k-320d-l-uni	58.0	45.4
SpeechTokenizer [33]	hubert_avg	65.7	59.9
DAC [34]	16k	74.7	65.8

V. PURIFICATION IMPACT ON ASV PERFORMANCE

While detection accuracy is a critical metric for evaluating adversarial defences, the ultimate objective of any defence method is to protect the downstream application, which is ASV in our case. A truly effective purification system must not only detect and counteract adversarial perturbations, but also preserve ASV performance under real-world conditions.

To evaluate this, we assess the impact of different purification methods on ASV performance using the Equal Error Rate

(EER) metric. Table III reports EERs for both bona fide and PGD adversarial trials. We distinguish between two types of trials: (i) *target vs. non-target*, which reflects performance for clean data, and (ii) *target vs. adversarial non-target*, which evaluates robustness under attack. Without purification, the ASV system achieves an EER of 1.4% for bona fide trials but fails dramatically under adversarial conditions, reaching 73.2%. This result highlights the vulnerability of unprotected ASV pipelines.

In contrast, MDD significantly reduces the EER for adversarial inputs, especially for masking ratios of 10% and 25%, which achieve 18.0% and 17.6% respectively, more than a fourfold reduction compared to the unprotected baseline. Importantly, the 10% MDD model also maintains a low EER (4.0%) for clean data, striking the best balance between robustness and reliability. Higher masking ratios (50% and above) lead to performance degradation for bona fide data due to excessive information loss during reconstruction. This underscores the importance of using a moderate masking strategy to retain speaker-discriminative features while suppressing adversarial noise.

Compared to other approaches, including the diffusion-based DAP model and several neural codec-based methods (AcademiCodec, SpeechTokenizer, DAC), MDD consistently achieves superior EERs across both clean and adversarial conditions. These alternative methods suffer from poorer generalisation and higher EERs, even when retrained under the same conditions. In summary, MDD provides not only strong adversarial detection, but also practical and effective purification for robust ASV, demonstrating its potential as a viable defence mechanism for real-world speaker verification systems.

TABLE III
PURIFIED ASV EER (%) RESULTS ON BONA FIDE DATA AND PGD ADVERSARIAL DATA WITH DIFFERENT METHODS. "TAR" REFERS TO TARGET SPEAKER TRIALS, "NON-TAR" REFERS TO NON-TARGET SPEAKER TRIALS, AND "ADV" REFERS TO ADVERSARIAL NON-TARGET TRIALS.

Purification method	**Bona Fide EER (%) tar vs. non-tar**	**PGD EER (%) tar vs. adv**
No Purification	1.4	73.2
0% MDD (unmasked)	6.0	19.2
10% MDD	**4.0**	18.0
25% MDD	6.2	**17.6**
50% MDD	18.8	27.6
75% MDD	40.6	43.4
100% MDD (fully masked)	49.4	50.0
DAP [14]	31.2	31.4
AcademiCodec [32]	51.4	51.4
SpeechTokenizer [33]	40.6	40.8
DAC [34]	33.0	35.0

VI. CONCLUSIONS

In this work, we introduced MDD, a novel adversarial defense framework designed to protect automatic speaker verification (ASV) systems against imperceptible perturbations. MDD leverages a *text-conditioned masked diffusion model*, which progressively denoises masked Mel-spectrograms while preserving essential speaker information through transcription-based conditioning.

Unlike many existing approaches, MDD does not rely on adversarial training or large-scale pretraining, yet it achieves strong performance in both detection and purification tasks. Our experiments demonstrate that a moderate masking ratio (specifically 10%) yields the best trade-off, allowing MDD to effectively identify adversarial examples while maintaining high verification accuracy on clean speech.

We performed comprehensive comparisons with state-of-the-art diffusion-based and neural codec-based purification methods. MDD consistently outperforms these baselines in terms of detection rate and purified ASV equal error rate (EER), confirming its robustness and practical applicability.

Importantly, we emphasise that the ultimate goal of adversarial defence is not only to detect attacks, but to ensure the reliability of ASV systems in real-world deployment scenarios. Our results show that MDD meets this goal, providing a lightweight, effective, and generalisable solution for securing voice-based authentication systems.

REFERENCES

[1] M. Todisco, M. Panariello, X. Wang, H. Delgado, K.-A. Lee, and N. Evans, "Malacopula: Adversarial automatic speaker verification attacks using a neural-based generalised hammerstein model," in *Proc. ASVspoof Workshop 2024*, 2024.

[2] C. Szegedy, W. Zaremba, I. Sutskever, *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[3] J. Villalba, Y. Zhang, and N. Dehak, "X-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Interspeech*, 2020, pp. 4233–4237.

[4] M. Todisco, X. Wang, V. Vestman, *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *Interspeech 2019*, 2019.

[5] J. Lan, R. Zhang, Z. Yan, J. Wang, Y. Chen, and R. Hou, "Adversarial attacks and defenses in speaker recognition systems: A survey," *Journal of Systems Architecture*, vol. 127, p. 102 526, 2022.

[6] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu, "Adversarial attack and defense strategies of speaker recognition systems: A survey," *Electronics*, vol. 11, no. 14, p. 2183, 2022.

[7] I. U. Hassan, K. Panduru, and J. Walsh, "Review of data processing methods used in predictive maintenance for next generation heavy machinery," *Data*, vol. 9, no. 5, p. 69, 2024.

[8] H. Wu, H.-C. Kuo, Y. Tsao, and H.-y. Lee, "Scalable ensemble-based detection method against adversarial attacks for speaker verification," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 4670–4674.

[9] X. Chen, J. Wang, X.-L. Zhang, W.-Q. Zhang, and K. Yang, "Lmd: A learnable mask network to detect adversarial examples for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2476–2490, 2023.

[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 Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 202–217, 2021.

[11] X. Chen, J. Du, H. Wu, J.-S. R. Jang, and H.-y. Lee, "Neural codec-based adversarial sample detection for speaker verification," in *Interspeech 2024*, 2024, pp. 522–526. DOI: 10.21437/Interspeech.2024-1191.

[12] X. Li, N. Li, J. Zhong, *et al.*, "Investigating robustness of adversarial samples detection for automatic speaker verification," in *Interspeech 2020*, 2020, pp. 1540–1544. DOI: 10.21437/Interspeech.2020-2441.

[13] S. Joshi, S. Kataria, J. Villalba, and N. Dehak, "Advest: Adversarial perturbation estimation to classify and detect adversarial attacks against speaker identification," in *Interspeech 2022*, 2022, pp. 5060–5064. DOI: 10.21437/Interspeech.2022-10985.

[14] Y. Bai, X.-L. Zhang, and X. Li, "Diffusion-based adversarial purification for speaker verification," *IEEE Signal Processing Letters*, 2024.

[15] S. Chen, Y. Bai, J. Yao, X.-L. Zhang, and X. Li, "Textual-driven adversarial purification for speaker verification," in *Proc. Interspeech 2024*, 2024, pp. 527–531.

[16] Y. Bai, X.-L. Zhang, and X. Li, "Adversarial purification for speaker verification by two-stage diffusion models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2024, pp. 1158–1164.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, PMLR, 2015, pp. 2256–2265.

[19] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 5329–5333.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *INTERSPEECH*, ISCA, 2020.

[21] Z. Chen, S. Chen, Y. Wu, *et al.*, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 6147–6151.

[22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*, Chapman and Hall/CRC, 2018, pp. 99–112.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[24] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2014.

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[26] H. Liu, Z. Chen, Y. Yuan, *et al.*, "Audioldm: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning*, PMLR, 2023, pp. 21 450–21 474.

[27] X. Li, Y. Ren, X. Jin, *et al.*, "Diffusion models for image restoration and enhancement–a comprehensive survey," *arXiv preprint arXiv:2308.09388*, 2023.

[28] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Ieee, 2022, pp. 7402–7406.

[29] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

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

[31] H. Wang, C. Liang, S. Wang, *et al.*, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[32] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *arXiv preprint arXiv:2305.02765*, 2023.

[33] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech language models," in *The Twelfth International Conference on Learning Representations*.

[34] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 980–27 993, 2023.