The defender's perspective on automatic speaker verification: An overview

Haibin Wu[1], Jiawen Kang[2], Lingwei Meng[2], Helen Meng[2] and Hung-yi Lee[1]

[1]*Graduate Institute of Communication Engineering, National Taiwan University*

[2]*Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong*

Abstract

Automatic speaker verification (ASV) plays a critical role in security-sensitive environments. Regrettably, the reliability of ASV has been undermined by the emergence of spoofing attacks, such as replay and synthetic speech, as well as adversarial attacks and the relatively new partially fake speech. While there are several review papers that cover replay and synthetic speech, and adversarial attacks, there is a notable gap in a comprehensive review that addresses defense against adversarial attacks and the recently emerged partially fake speech. Thus, the aim of this paper is to provide a thorough and systematic overview of the defense methods used against these types of attacks.

Keywords

Automatic speaker verification, Replay and synthetic speech, Adversarial attack, Partially fake speech, Review

1. Introduction

The past few years have witnessed significant advances in ASV, and this technique is now widely integrated into daily life, including voice activation in smartphones and e-banking authentication. However, ASV is serious vulnerable to malicious spoofing attacks includes tactics such as replay and synthetic speech, adversarial attacks and recently emerged partially fake speech.

While there are several review papers that cover replay and synthetic speech [1, 2, 3, 4], and adversarial attacks [5], there is a notable gap in a comprehensive review that addresses defense methods against adversarial attacks and the recently emerged partially fake speech. The objective of this thesis is to provide a thorough and systematic overview of the defense methods used against these two types of attacks. It is hoped that they will inspire further researches within the ASV community.

2. Attacks

2.1. Partially fake speech

The first Audio Deep Synthesis Detection challenge (ADD 2022) [6] releases a kind of brand new attack, known as the partially fake speech attack [7]. The ASVspoof challenge [1, 2, 3, 4] focuses on generating spoofing speech in its entirety, ignoring the scenario of partially fake speech, where small fake clips are hidden within a piece of real speech. The generation of partially fake audio involves the insertion of only small clips of synthetic speech into the real speech as shown in Figure 1, resulting in even more stealthy fake speech containing a significant amount of the genuine user's audio.

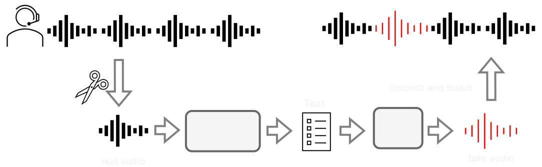

Figure 1: The partially fake audio generation process. A small clip is selected from the user's utterance, the content is recognized using Automatic Speech Recognition (ASR), and the recognized content is modified to manipulate the meaning of the entire speech. The fake clip is then generated using Text-to-Speech (TTS) or Voice Conversion (VC), and inserted into the genuine utterance to generate the partially fake speech.

Previous studies [7, 8] have shown that it is challenging to differentiate between partially fake and genuine audios by directly using existing state-of-the-art countermeasure models fostered by the ASVspoof challenge [1, 2, 3, 4]. These countermeasure models address the problem of identifying whether an entire audio utterance is genuine or fabricated. However, they are not equipped to identify anomalous regions within a single utterance.

2.2. Adversarial attacks

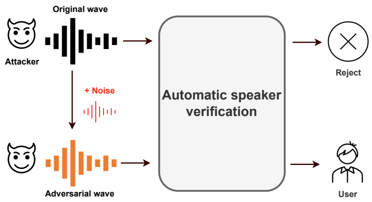

Figure 2: A tiny adversarial noise is added to the original wave to get the adversarial one to fool the ASV falsely accept.

Speaker verification models are also subject to adversarial attacks [9, 5, 10] as shown in Figure 2. Kreuk et al. [11] are among the pioneers in studying the susceptibility of ASV models to adversarial attacks. Additionally, even the current state-of-the-art ASV models, including i-vector [12] and x-vector [13] systems, are not immune to adversarial attacks. [14] conducts a pioneering effort

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R
f07921092@ntu.edu.tw (H. Wu)

in exposing the adversarial weakness of countermeasure models and [15] further enhances the transferability of adversarial attacks through model ensemble.

3. Defense methods

3.1. Tackle partially fake speech attacks

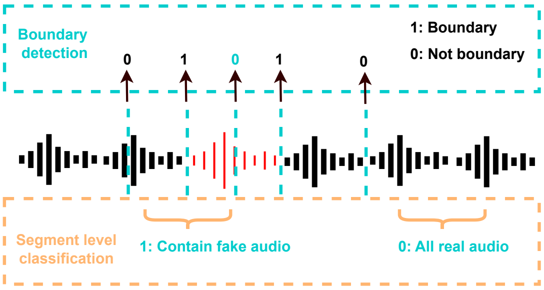

Figure 3: The two categories of methods to tackle partially fake speech attacks. The black and red parts of the utterance are real and fake, respectively. The first approach, illustrated in the blue block, focuses on detecting the transition boundaries between the genuine and fake segments. The second method, depicted in the orange block, endeavors to distinguish between genuine and fake short segments.

Partially fake speech attacks are generated as shown in Figure 1. As this kind of attack is brand new, there have been only a few initiatives to handle this attack, and we categorize these efforts into two categories as shown in Figure 3: transition boundary detection [16, 17, 18] and segment level classification [7, 8, 19, 20].

3.1.1. SSL-based feature extractor

Before delving into the two main approaches, let's first examine the feature engineering aspect of the task. Lv et al. [21] are the pioneers in utilizing self-supervised learning (SSL) models to tackle partially fake speech attacks. Rather than using traditional acoustic features, they instead adopt XLS-R [22], a self-supervised learning model, as the feature extractor. Their method [21], which involved simply adding a lightweight prediction head on top of the XLS-R model and fine-tuning the large XLS-R model, ultimately achieved first place out of 33 international teams in the ADD challenge [6].

Their efforts [21] have taught us a valuable lesson - the acoustic features extracted by a fine-tuned self-supervised learning model can be incredibly helpful for detecting partially fake speech. It's worth noting that the two main approaches introduced below can also harness the power of self-supervised learning models, provided there are sufficient computing resources available.

3.1.2. Transition boundary detection

[16] is the first to introduce the transition boundary detection task for partially fake audio detection. The transition boundaries contain artifacts, such as discontinuity in speech and inconsistencies in ambient noise. Inspired by the extraction-based question-answering models [23] used in natural language processing (NLP), we refer to the boundary detection task as a question-answering or fake span discovery proxy task. In this task, the model is required to answer the question “where is the fake clip?" in a piece of partially fake audio. Extraction-based question-answering models in NLP typically take a question and a passage as input, construct representations for the passage and the question, match the question and passage embeddings, and then output the start and end positions of the answer within the passage. In our case, the passage is the partially fake utterance, and the answer is the start and end time of the fake clip. As depicted in the blue block of Figure 3, when the model is presented with a boundary frame between a real (black) and a fake (red) clip, it should predict “1". Conversely, when the model is presented with a non-boundary frame, it should predict “0". By training the model on the question-answering proxy task, the model can learn to find the concatenation boundaries with discontinuity and identify fake clips within an utterance, thus improving its ability to distinguish between audios with and without fake clips. The proposed method placed the second out of 33 international teams in the ADD challenge [6], even without the assistance of self-supervised learning features.

Wang et al. [18] divide the entire utterance into several chunks, and extracted acoustic features from each chunk to feed into the deep learning model. The model is then tasked with determining whether a boundary exists within the given chunk by predicting “1" if the chunk contains a boundary, or a “0" if it does not. Through training, the model gains the ability to identify clues such as speech discontinuity or inconsistencies in ambient noise, allowing it to effectively highlight potential boundaries.

Cai et al. [17] propose to introduce the self-supervised learning model for frame-level boundary detection to detect partially fake speech. They modify the method in [16] to further boost the detection performance: 1). Instead of solely focusing on transition boundaries that indicate inconsistency and discontinuity, [17] proposes setting nearby frames of the boundaries as boundaries to increase robustness. 2). [17] employs wav2vec 2.0 [24], a self-supervised learning model as feature extractor and also fine-tunes the feature extractor during training. Utilizing the features from wav2vec 2.0 improves the performance by a relative 58.25% compared to traditional acoustic features extracted by digital signal processing front-ends.

The main takeaway message from this subsection is that the transition boundaries can serve as a useful cue to identify partially fake audio, as it indicates discontinuity and inconsistency in speech. By tasking models with detecting these boundaries, they can learn to identify these cues and detect partially fake speech.

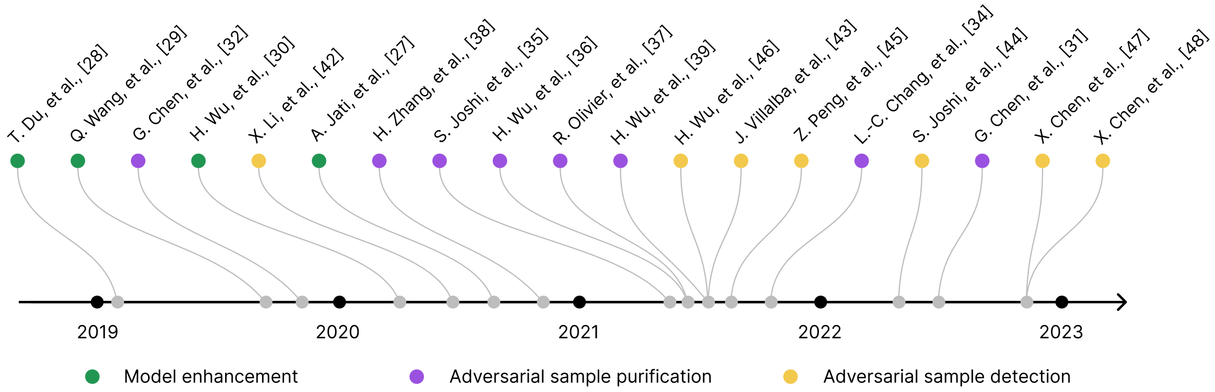

Figure 4: The timeline of defense methods for ASV against adversarial attacks.

3.1.3. Segment level classification

The goal of segment level classification is to distinguish between genuine and fake segments. The short segments have different time resolutions, ranging from 1 frame (around 20 ms) to the entire utterance. Segments that only contain genuine speech will be labeled as “1", while all other segments will be labeled as “0" as shown in the orange block of Figure 3. Zhang et al. [8] do the initial attempt to conduct segment level classification for partially fake speech detection with a fixed time resolution. In their subsequent works [19], they propose to train the countermeasure model by both the utterance level classification and segment level classification. To further boost the countermeasure’s performance, they [20] introduce the self-supervised learning models [25, 24] as the front-end feature extractor, and enable the model to learn segment level classification with different time resolutions, ranging from 1 frame to the entire utterance.

The time resolution used in segment level classification is a crucial hyperparameter for training. If the segment’s frame number is too small, the model may not extract enough information to distinguish between genuine and fake segments. On the other hand, if the frame number is too large, the proportion of fake frames may be too small, resulting in fake frames being dominated by genuine frames. Enabling the model learn from different time resolutions [20] is a reasonable solution to bypass the hyperparameter search. Note that in Figure 1, the inserted red clip can be from other genuine users. The segment level classification [8, 19, 20] does not consider this condition into account as in their produced dataset, the inserted clips are always fake.

3.2. Defense against adversarial attacks

We propose to classify the defense methods into three categories and the timeline for related works is shown in Figure 4. 1). Model enhancement focuses on developing robust models during the training phase by modifying the models’ internals, making the attackers difficult to find effective adversarial examples. 2). Adversarial sample purification aims to alleviate the superficial adversarial noise and transform adversarial samples into genuine samples. 3). Adversarial sample detection aims to distinguish between adversarial and genuine samples, allowing the identification and removal of adversarial samples.

3.2.1. Model enhancement

[26, 27, 28] adopt adversarial training to alleviate the vulnerability of ASV against adversarial attacks. Wu et al. [29] also investigate improving the adversarial robustness for countermeasures by adversarial training.

Model enhancement methods involve modifying the model’s parameters, and they can usually work together with purification and detection methods.

3.2.2. Adversarial sample purification

Previous efforts for purification can be classified into 5 categories: Lossy pre-processing, adding noise, generative method, denoising method and filtering.

The “Lossy pre-processing" approach treats adversarial perturbations as redundant information and discards it to improve the model’s adversarial robustness. Chen et al. [30] consider adversarial perturbations as redundant information and use lossy speech compression techniques to mitigate these perturbations. Quantization [31, 30] involves rounding each audio sample point to the nearest integer multiple of a factor q, which can impact the fragile adversarial perturbations. Chen et al. [30] propose to do k-means [32] on the acoustic features to get clusters of acoustic features, and use the clusters to represent the acoustic features.

The “adding noise" approach aims to disrupt and neutralize adversarial perturbations, by introducing additional noise, typically Gaussian. Randomized smoothing [31, 33, 30, 34] involves adding random Gaussian noise to the input utterances before sending them to the ASV to counter the adversarial perturbations. [35] adopts to the idea of “voting for the right answer" to prevent risky decisions of ASV in blind spot areas. To achieve this, they samples the neighbors of a given utterance by random sampling using Gaussian noise, and allow the neighbors to vote on whether the utterance should be accepted by the ASV model or not, rather than relying solely on the prediction of the single utterance. Olivier et al. [36] is an enhanced version by adding Gaussian noise to the high-frequency region rather than the entire utterance.

The “Denosing method" treats adversarial noise as a specific kind of noise and aims to estimate and eliminate it. Chang et al. [33] suggest using a denoising algorithm tailored for Gaussian noise and they contend that the denoising algorithm can also cleanse the adversarial noise. Zhang et al. [37] propose to employ an adversarial separation network, which is trained using the adversarial-genuine data pairs, to estimate and purify the adversarial noise. This method requires prior knowledge of adversarial sample generation.

The “generative method" approach typically involve training a generative model to model the genuine data manifolds and using this model to pull the adversarial samples towards the genuine data manifolds. Wu et al. [38] propose the SSLM-based reconstruction to alleviate the superficial adversarial noise and maintain key information for genuine samples. They [38] utilize the self-supervised learning models to extract key features from the adversarial samples, and do reconstruction to pull the inputs to the genuine data manifold. Joshi et al. [34] use the encoder of a VAE [39] to project testing data onto a latent posterior that aligns with the genuine manifold. They then use the decoder to re-generate the input data based on the hidden embedding sampled by the latent posterior, thereby purifying superficial adversarial noise. Joshi et al. [34] borrow the DefenseGAN from computer vision [40]. The DefenseGAN projects the testing data, either adversarial or genuine, into the low-dimensional manifold of genuine data to get the hidden embeddings and then re-generate the testing data by the generator using such embeddings.

“Filtering", also known as local smoothing, helps smooth and alleviate the superficial adversarial perturbations. Local smoothing involves applying Gaussian, mean, and median filters to the waveform to purify the adversarial noise. [31, 30] and [29] utilize local smoothing to defend ASV and countermeasures, respectively.

3.2.3. Adversarial sample detection

The detection methods can be classified into two categories based on whether they require prior knowledge about adversarial sample generation: attack-dependent or attack-independent detection methods.

The attack-dependent methods usually leverage the deep learning models to implicitly find cues to differentiate between specific kinds of adversarial samples and genuine samples using both adversarial and genuine data. Li et al. [41] propose to train a detector using the binary classification loss to distinguish the adversarial and genuine samples. They find their detector is unable to detect unseen adversarial samples derived by other adversarial attack algorithms that are not used during training. Based on that different kinds of adversarial samples attain different attack signatures, Villalba et al. [42] propose to train an x-vector [13] system to extract the bottleneck features as the attack signatures using various types of adversarial samples. After training the x-vector system, attack signatures will be extracted for different types of attacks. During inference, the testing utterance is inputted, and the x-vector feature extractor will extract the hidden embeddings. These embeddings are then compared with the enrolled attack signatures to determine whether the testing utterance is an adversarial sample or not. To further improve the performance of the attack signature extractor, Joshi et al. [43] propose training the attack signature extractor using adversarial perturbations instead of adversarial examples. They argue that the adversarial perturbations eliminate redundant information from the adversarial samples. They then train an adversarial perturbation estimator to extract adversarial perturbations from the input utterance and use the attack signature extractor to extract hidden features to detect the adversarial samples.

Attack-independent methods treat the detection of adversarial samples as an anomaly detection problem. Genuine data samples always exhibit some properties that are absent or different for adversarial samples. Therefore, attack-independent detection methods can exploit the inconsistency of these internal properties to distinguish between adversarial and genuine samples. Wu et al. [38] leverage the ASV score difference before and after putting the testing utterance into SSLMs as an indicator to differentiate between adversarial and genuine samples. Specifically, for genuine samples, the ASV score difference before and after putting the utterance into SSLMs is small, while for adversarial samples, the difference is large. Peng et al. [44] propose to detect adversarial samples using twin ASV models, including one premier model that is exposed to attackers and is fragile under adversarial attack, and one mirror model that is robust to adversarial attacks and cannot be accessed by attackers. When a genuine sample is inputted, both the premier and mirror models produce similar predictions. However, when an adversarial sample is inputted, the models produce different predictions. Peng et al. [44] leverage the score inconsistency between genuine and adversarial samples to detect adversarial samples. Wu et al. [45] utilize the vocoders to re-synthesize the input utterance and find that the difference between the ASV scores for the original and re-synthesized utterance is a good indicator for discrimination between genuine and adversarial samples. To be specific, the score difference for adversarial samples is large, while it is small for genuine samples. Chen et al. [46] utilize two kinds of hand-crafted masks to detect adversarial samples: they mask parts of the input speech features. They claim the masked parts contain less speaker information and won't affect the ASV scores for genuine samples two much, but will greatly impact the adversarial samples. By comparing the absolute difference of scores before and after masking, they are able to detect adversarial examples. The two masks used are MLFB-H, which masks the high frequencies of LogF-Bank, and MLFB-D, which masks the time-frequency bins whose absolute values of their one-order difference along the frequency axis are smaller than a threshold. Chen et al. [47] further enhance the detection performance by learning such mask matrix by a deep recurrent networks, rather than using hand-crafted masks.

4. Future directions

For the future directions of partially fake speech attacks: 1). Data collection. The collection of data is a crucial component in developing an effective defense system against partially fabricated speech. Only 100k utterances are collected by [6] for partially fake detection and the transition boundaries are not stealthy enough. To this end, there exists a pressing need to investigate the generation of more data with discreet transition boundaries, while carefully considering the linguistic and acoustic characteristics involved. This undertaking is of great significance and warrants further exploration. 2). Reduce training efforts. The state-of-the-art (SOTA) methodology for partially fake speech detection involves the fine-tuning of the entire SSLMs. The SSLM in [21] is with 2 billion parameters, which presents a challenge for academic researchers when attempting to fine-tune the model. Several works have emerged that offer promising avenues for minimizing training efforts while maximizing the benefits of SSLMs, including linear probing, adapter, and prompt techniques. Exploring these approaches may significantly enhance the efficiency of adopting SSLMs for partially fake speech detection. 3). Model compression. The current state-of-the-art detection method relies heavily on large-scale SSLMs. The parameter number of the SSLM used in [21] is 2 billion parameters. Therefore, investigating approaches to reduce the model size is a crucial research endeavor. This issue warrants considerable attention as it has significant implications for the scalability, computational efficiency, and generalizability of partially fake speech detection systems.

The re-synthesis-based adversarial sample detection methods achieves the SOTA [45, 46, 47]. An effective audio re-synthesis method for adversarial sample detection must possess two critical properties. Firstly, the score variations between the original and re-synthesized utterances should be minimal for genuine samples. Secondly, the score variations between the original and re-synthesized utterances for adversarial samples should be substantial. Investigating approaches for refining the design of audio re-synthesis methods to further optimize these properties represents a valuable research direction. By enhancing the efficacy of the audio re-synthesis method, it would be possible to improve the reliability and accuracy of detection systems.

5. Conclusion

This paper reviews the defense methods against adversarial attacks and partially fake speech attacks that have recently emerged. We hope the comprehensive review and comparisons can inspire future works to boost the robustness of ASV. Further investigation is needed to explore future directions as in Section 4

References

[1] J. Yamagishi, et al., Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection, arXiv preprint arXiv:2109.00537 (2021).

[2] M. Todisco, et al., Asvspoof 2019: Future horizons in spoofed and fake audio detection, arXiv preprint arXiv:1904.05441 (2019).

[3] T. Kinnunen, et al., The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection (2017).

[4] Z. Wu, et al., Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Sixteenth Annual Conference of the ISCA, 2015.

[5] H. Tan, et al., Adversarial attack and defense strategies of speaker recognition systems: A survey, Electronics 11 (2022) 2183.

[6] J. Yi, et al., Add 2022: the first audio deep synthesis detection challenge, in: IEEE ICASSP, 2022.

[7] J. Yi, et al., Half-truth: A partially fake audio detection dataset, in: Interspeech, 2021, pp. 1654–1658.

[8] L. Zhang, et al., An initial investigation for detecting partially spoofed audio, arXiv preprint arXiv:2104.02518 (2021).

[9] H. Abdullah, et al., Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems, in: IEEE SP, 2021, pp. 730–747.

[10] R. K. Das, et al., The attacker's perspective on automatic speaker verification: An overview, arXiv preprint arXiv:2004.08849 (2020).

[11] F. Kreuk, et al., Fooling end-to-end speaker verification with adversarial examples, in: IEEE ICASSP, 2018, pp. 1962–1966.

[12] X. Li, et al., Adversarial attacks on gmm i-vector based speaker verification systems, in: IEEE ICASSP, 2020, pp. 6579–6583.

[13] J. Villalba, et al., x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification, ISCA Interspeech (2020) 4233–4237.

[14] S. Liu, et al., Adversarial attacks on spoofing countermeasures of automatic speaker verification, in: IEEE ASRU, 2019, pp. 312–319.

[15] Y. Zhang, et al., Black-box attacks on spoofing countermeasures using transferability of adversarial examples., in: ISCA Interspeech, 2020, pp. 4238–4242.

[16] H. Wu, et al., Partially fake audio detection by self-attention-based fake span discovery, in: IEEE ICASSP, IEEE, 2022, pp. 9236–9240.

[17] Z. Cai, et al., Waveform boundary detection for partially spoofed audio, arXiv preprint arXiv:2211.00226 (2022).

[18] L. Wang, et al., Synthetic voice detection and audio splicing detection using se-res2net-conformer architecture, arXiv preprint arXiv:2210.03581 (2022).

[19] L. Zhang, et al., Multi-task learning in utterance-level and segmental-level spoof detection, arXiv preprint arXiv:2107.14132 (2021).

[20] L. Zhang, et al., The partialspoof database and countermeasures for the detection of short fake speech segments embedded in an utterance, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2022).

[21] Z. Lv, et al., Fake audio detection based on unsupervised pretraining models, in: IEEE ICASSP, IEEE, 2022, pp. 9231–9235.

[22] A. Babu, et al., Xls-r: Self-supervised cross-lingual speech representation learning at scale, arXiv preprint arXiv:2111.09296 (2021).

[23] A. M. N. Allam M. H. Haggag, The question answering systems: A survey, IJRRIS 2 (2012).

[24] A. Baevski, et al., wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[25] S. Chen, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[26] A. Jati, et al., Adversarial attack and defense strategies for deep speaker recognition systems, Computer Speech & Language 68 (2021) 101199.

[27] T. Du, et al., Sirenattack: Generating adversarial audio for end-to-end acoustic systems, in: ACM ASIACCS, 2020.

[28] Q. Wang, et al., Adversarial regularization for end-to-end robust speaker verification., in: Interspeech, 2019.

[29] H. Wu, et al., Defense against adversarial attacks on spoofing countermeasures of asv, arXiv preprint arXiv:2003.03065 (2020).

[30] G. Chen, et al., Towards understanding and mitigating audio adversarial examples for speaker recognition, IEEE TDSC (2022).

[31] G. Chen, et al., Who is real bob? adversarial attacks on speaker recognition systems, arXiv preprint arXiv:1911.01840 (2019).

[32] J. A. Hartigan M. A. Wong, Algorithm as 136: A k-means clustering algorithm, Journal of the royal statistical society. series c (applied statistics) 28 (1979) 100–108.

[33] L.-C. Chang, et al., Defending against adversarial attacks in speaker verification systems, in: IEEE IPCCC, 2021.

[34] S. Joshi, et al., Adversarial attacks and defenses for speaker identification systems, arXiv preprint arXiv:2101.08909 (2021).

[35] H. Wu, et al., Voting for the right answer: Adversarial defense for speaker verification, arXiv preprint arXiv:2106.07868 (2021).

[36] R. Olivier, et al., High-frequency adversarial defense for speech and audio, in: IEEE ICASSP, 2021.

[37] H. Zhang, et al., Adversarial separation network for speaker recognition., in: Interspeech, 2020.

[38] H. Wu, et al., Improving the adversarial robustness for speaker verification by self-supervised learning, IEEE/ACM Transactions on Audio, Speech, and Language Processing 30 (2021) 202–217.

[39] D. P. Kingma M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114 (2013).

[40] P. Samangouei, et al., Defense-gan: Protecting classifiers against adversarial attacks using generative models, arXiv preprint arXiv:1805.06605 (2018).

[41] X. Li, et al., Investigating robustness of adversarial samples detection for automatic speaker verification, arXiv preprint arXiv:2006.06186 (2020).

[42] J. Villalba, et al., Representation learning to classify and detect adversarial attacks against speaker and speech recognition systems, arXiv preprint arXiv:2107.04448 (2021).

[43] S. Joshi, et al., Advest: Adversarial perturbation estimation to classify and detect adversarial attacks against speaker identification, arXiv preprint arXiv:2204.03848 (2022).

[44] Z. Peng, et al., Pairing weak with strong: Twin models for defending against adversarial attack on speaker verification., in: Interspeech, 2021.

[45] H. Wu, et al., Adversarial sample detection for speaker verification by neural vocoders, in: IEEE ICASSP, 2022.

[46] X. Chen, et al., Masking speech feature to detect adversarial examples for speaker verification, in: IEEE APSIPA ASC, 2022.

[47] X. Chen, et al., Lmd: A learnable mask network to detect adversarial examples for speaker verification, arXiv preprint arXiv:2211.00825 (2022).