INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Pairing Weak with Strong: Twin Models for Defending against Adversarial Attack on Speaker Verification

Zhiyuan Peng[1], *Xu Li*[2], *Tan Lee*[1]

[1]Department of Electronic Engineering, The Chinese University of Hong Kong, China
[2]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, China

jerrypeng1937@gmail.com, xuli@se.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Vulnerability of speaker verification (SV) systems under adversarial attack receives wide attention recently. Simple and effective countermeasures against such attack are yet to be developed. This paper formulates the task of adversarial defense as a problem of attack detection. The detection is made possible with the verification scores from a pair of purposely selected SV models. The twin-model design comprises a fragile model paired up with a relatively robust one. The two models show prominent score inconsistency under adversarial attack. To detect the score inconsistency, a simple one-class classifier is adopted. The classifier is trained with normal speech samples, which not only bypasses the need of crafting adversarial samples but also prevents itself from over-fitting to the crafted samples, and hence makes the detection robust to unseen attacks. Compared to single-model systems, the proposed system shows consistent and significant performance improvement against different attack strategies. The false acceptance rates (FARs) are reduced from over 63.54% to 2.26% under the strongest attack. Our approach has practical benefits, e.g., no need to modify a well-deployed SV model even it is well-known and can be fully accessed by the adversary. Moreover, it can be combined with existing single-model countermeasures for even stronger defenses.

Index Terms: speaker verification, adversarial defense, twin models, one-class classifier

1. Introduction

Speaker verification (SV) is the task of verifying whether a spoken utterance comes from a claimed or hypothesized speaker identity. Biometric authentication requires highly secure SV systems, which are robust against malicious spoofing threats [1–3]. Very recently, a novel threat to SV systems, named as adversarial attack, has been gaining importance [4–6]. By injecting a subtle perturbation into a genuine voice, which is almost indistinguishable by human perception, adversarial attack could easily fool many existing SV systems, including i-vector speaker embedding systems [5], neural speaker embedding systems [7, 8] as well as end-to-end systems [4]. Moreover, it is found that the perturbed voice designed for fooling one SV system has the transferability of deceiving other SV systems, which may differ in training data, feature configuration, and model structure [4, 5, 9, 10], resulting in a strong urge to defend adversarial attack to state-of-the-art SV systems.

Counter-measures against adversarial attack have been investigated extensively in computer vision tasks [11–13]. The exploration on speech-related applications remains scarce. Existing defense mechanisms can be categorized into passive and proactive defenses. Passive defenses attempt to modify the input sample and remove adversarial perturbations by, for examples, spatial smoothing [12, 14], down-sampling [15], deep filtering [11,16,17], etc. The general idea behind these methods is that the perturbation is preferably noise-like and could be easily *removed* from the waveform or some feature representations. On the other hand, proactive defenses aim to improve model robustness through adversarial re-training [13, 14, 18] or incorporating an additional attack detector [8, 19, 20]. The underlying assumption is that the perturbation pattern can be *perceived* and *memorized* by a data-driven model.However, from the attacker's perspective, the perturbation pattern in the low-level feature representations could be flexibly changed by altering the perturbation algorithm along with its attack configurations [8]. Memorizing all possible perturbation patterns is essentially impossible for the defender.

Nevertheless, all kinds of perturbations for adversarial attack serve the same goal of fooling the target SV systems, e.g., to accept an adversarial impersonation voice with high confidence scores. For the same adversarial voice, the confidence scores from different systems may vary greatly and be mutually inconsistent, because of diverse robustness of the systems. In this regard, attack detection based on the scores from a group of diverse SV systems could be straightforward and robust against unseen attacks.

This paper proposes a strong countermeasure that adopts twin SV models for proactive defenses. The twins, made up of a premier SV model and a mirror one, give consistent scores on a genuine voice whilst disagree a lot on a perturbed voice. The twin models are designed to serve different functions. The premier model should be able to achieve state-of-the-art SV performance on genuine utterances. It is expected to be well-known to attackers and fragile under adversarial attack. The mirror model should be rarely-seen with some unique designs . It is hidden from the attacker and relatively robust against attack. Design of the twins will be detailed in Section 3.1. Their robustness gap is one of the reasons that cause score inconsistency under attack. We experimentally find that the score inconsistency generally exists under several representative adversarial attack algorithms. Furthermore, the more advanced the algorithm, the more prominent inconsistency it tends to cause. To detect the score inconsistency, a simple one-class classifier, named minimum covariance determinant (MCD) [21], is used. Only genuine speech samples are required for training this classifier, unlike most existing proactive defense approaches for which adversarial samples need to be crafted [8,13,14,18,19]. Moreover, the one-class training prevents the classifier from over-fitting to specific attack patterns, and hence improves the defense robustness against unseen attacks.

http://dx.doi.org/10.21437/Interspeech.2021-1343

2. Adversarial Attack

In a speaker verification task, an adversarial voice is a speech waveform mixed with intentional perturbations that can fool a well-trained SV model to make mistakes. Common approaches to generate the perturbations are by performing gradient descent on an objective function w.r.t the input data. Specifically, given a test utterance $\boldsymbol{x}$ and its claimed speaker identity s, an SV model with parameters θ assigns the claim a confidence score $S_\theta(\boldsymbol{x}, s)$. A well-trained SV model usually outputs a high confidence score if the test utterance $\boldsymbol{x}$ is uttered by the claimed speaker s. If the score reaches a predefined threshold, the speaker claim will be accepted. From an adversary's perspective, it will optimize the perturbation $\boldsymbol{\delta_x}$ to manipulate the confidence score so that the whole system behaves incorrectly on the perturbed voice $\boldsymbol{x} + \boldsymbol{\delta_x}$, e.g., either falsely rejecting the true target speaker or accepting the imposter. This optimization problem can be formulated by,

$$\boldsymbol{\delta_x} = \underset{||\boldsymbol{\delta_x}||_p \leq \epsilon}{\arg\max}\, L_\theta(\boldsymbol{x} + \boldsymbol{\delta_x}, s) \tag{1}$$

$$L_\theta(\boldsymbol{x} + \boldsymbol{\delta_x}, s) = k \times S_\theta(\boldsymbol{x} + \boldsymbol{\delta_x}, s) \tag{2}$$

$$k = \begin{cases} -1, & \boldsymbol{x} \text{ is target's voice} \\ 1, & \boldsymbol{x} \text{ is imposter's voice} \end{cases} \tag{3}$$

, where L_θ is the score loss function to be maximized for score manipulation. The p-norm of $\boldsymbol{\delta_x}$ is bounded by the perturbation degree ϵ. In this paper, p is set to ∞. The perturbation degree ϵ balances between attack intensity and attack stealthiness. Increasing ϵ within a reasonable range can improve the attack success rate but the perturbation may be easily perceived by human [22] or well-designed attack detectors [8].

2.1. Perturbation algorithms

2.1.1. Fast gradient sign method (FGSM)

FGSM [23] perturbates the voice $\boldsymbol{x}$ by taking a single step ϵ in the direction of the signed gradient, as shown in Eq.4-5:

$$\boldsymbol{\delta_x} = \text{sign}(\nabla_{\boldsymbol{x}} L_\theta(\boldsymbol{x}, s)) \tag{4}$$

$$\boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon \times \boldsymbol{\delta}_x \tag{5}$$

FGSM can fool an SV system without difficulty as long as the system is fully accessed by the attacker (*white-box attack*), including its feature extraction process, the model structure, and parameters, etc. However, in many real scenarios, the victim systems are both unknown and unseen to its adversary (*black-box attack*), making the FGSM attack much less effective.

2.1.2. Iterative FGSM (I-FGSM)

Iterative FGSM [24], as its name implies, takes iterative smaller steps (ϵ/N) to update the adversarial utterance $\boldsymbol{x}^{(i)}$ by

$$\boldsymbol{\delta}_{\boldsymbol{x}}^{(i)} = \nabla_{\boldsymbol{x}} L_\theta(\boldsymbol{x}^{(i)}, s) \tag{6}$$

$$\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)} + \epsilon/N \times \text{sign}(\boldsymbol{\delta}_{\boldsymbol{x}}^{(i)}) \tag{7}$$

, where $\boldsymbol{x}^{(0)} = \boldsymbol{x}$ and $i = 0, \cdots, N-1$. N is the number of iterations. It has been shown that I-FGSM could perform stronger white-box attack than FGSM but still transfer inadequately to unseen systems [25].

2.1.3. Moment iterative FGSM (MI-FGSM)

MI-FGSM [26] introduces a momentum term into the iterative process of I-FGSM. It improves Eq.6 by,

$$\boldsymbol{\delta}_{\boldsymbol{x}}^{(i)} = \mu \times \boldsymbol{\delta}_{\boldsymbol{x}}^{(i-1)} + \frac{\nabla_{\boldsymbol{x}} L_\theta(\boldsymbol{x}^{(i)}, s)}{||\nabla_{\boldsymbol{x}} L_\theta(\boldsymbol{x}^{(i)}, s)||_1} \tag{8}$$

where $\boldsymbol{\delta}_{\boldsymbol{x}}^{(i)}$ accumulates all the $L1$- normalized gradients in the first i iterations with a moment decay factor μ. Compared to I-FGSM, MI-FGSM has much better transferability and performs well on black-box attack [26].

2.1.4. Ensemble MI-FGSM (ens-MI-FGSM)

To achieve more transferable attack, ensemble learning has been leveraged for adversarial attack [13]. Given a group of K SV models $\{\theta_k\}_{k=1}^K$, ens-MI-FGSM improves Eq.8 by collecting and averaging gradients from multiple SV models,

$$\boldsymbol{\delta}_{\boldsymbol{x}}^{(i)} = \mu \times \boldsymbol{\delta}_{\boldsymbol{x}}^{(i-1)} + \frac{1}{K}\sum_{k=1}^{K} \frac{\nabla_{\boldsymbol{x}} L_{\theta_k}(\boldsymbol{x}^{(i)}, s)}{||\nabla_{\boldsymbol{x}} L_{\theta_k}(\boldsymbol{x}^{(i)}, s)||_1} \tag{9}$$

The intuition is that if a perturbation remains adversarial for multiple models, it may capture an intrinsic direction that always fools these models and is more likely to transfer to the unseen. This fundamental ensemble approach has many variants [10, 26, 27] to further improve its attack transferability.

2.2. Attack evaluation

The criteria for benchmarking adversarial attack remain unclear in the SV community. Existing work usually leverages the common SV metrics, like equal error rate (EER) and minimum detection cost function (minDCF), to evaluate the performance degradation under attack. Essentially, these metrics adjust the system's decision threshold for a balance between false acceptance rate (FAR) and false rejection rate (FRR). For a practical SV system, once deployed, its decision threshold has been fixed even when it is under attack. Perhaps it would be more reasonable to separately evaluate the FAR increase under impersonation attack ($k = 1$) and the FRR increase under evasion attack ($k = -1$). Additionally, in most realistic attack situations, we are concerned more with impersonation attack where the adversary could get unauthorized access to its target system.

With these considerations, we report the FAR increase of SV systems under impersonation attack. The systems' decision thresholds are fixed at which FAR = FRR before attack.

3. Adversarial Defense

Fig.1 describes the components of our proposed defense system. It consists of two stand-alone SV models scoring the input speaker claim in parallel, followed by an attack detector making the accept/reject decision. The two SV models act as twins such that their scores (premier score and mirror score) on a genuine speech utterance will be in general consistent with each other. Meanwhile, an adversarial utterance mixed with intentional perturbations could give rise to significant score disagreements from the twins. This abnormal pattern could be captured by the attack detector and leveraged for making the final decision. Specifically, only if the premier score reaches its acceptance threshold and the mirror score is validated to be consistent with the premier one, the speaker claim could be finally accepted. In this way, adversarial impersonation attack is defended.

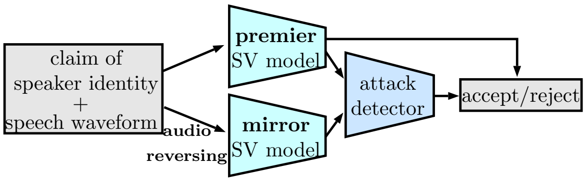

Figure 1: *Proposed system to defend against adversarial impersonation attack. For the sake of simplicity, feature extraction is omitted in this figure and regarded as part of SV models.*

Note that we report only our design and experiments on impersonation attack due to the realistic concern explained in section 2.2.

3.1. Twin SV models

The two representative deep SV models, xvector [28] and rvector [29], could be good candidates of the twins. As shown in Fig. 2, we plot their confidence scores of an imposter's voice $\boldsymbol{x}^*$ added with two different perturbations, i.e., $\boldsymbol{x}^* = \boldsymbol{x} + \epsilon_1\boldsymbol{\delta}_1 + \epsilon_2\boldsymbol{\delta}_2$. The two perturbations, $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$, are crafted on the xvector and rvector models, respectively, according to Eq.4. As either of the perturbation degrees, ϵ_1 or ϵ_2, increases, the verification scores from both models rise, reflecting their growing confidence in accepting the imposter's voice. In the meanwhile, the score gap gets enlarged, suggesting their prominent score inconsistency under attack.

Functional requirements on the twins are different. (1) The premier model should achieve state-of-the-art performance on genuine utterances. It is expected to be well-known to attackers and fragile under adversarial attack. Its vulnerability is taken as an advantage for the defense. This requirement coincides with the aim of adversarial attack (to find/attack vulnerable models). Therefore, existing work on searching and attacking vulnerable SV models [4, 5, 7, 9, 10] can be adopted to design the premier model. (2) The mirror model should be rarely seen and hidden from its adversary. In addition, its robustness against attack needs to be better than the premier model. An intuitive approach is to modify the input voice (passive defense) such that the score function $S_\theta(\boldsymbol{x}, s)$ alters while the speaker information remains.

With the considerations above and experiments in Section 4.2, the premier model is set as the standard TDNN xvector [28], and the mirror is the ResNet-34 rvector [29]. In addition, input voice to the rvector model is time-reversed before feature extraction, as shown in Fig.1. Both models are trained with additive angular margin (AAM)-softmax loss with hyperparameters $\{m = 0.3, s = 30\}$. Dimensions of speaker embeddings extracted from xvector and rvector models are set to 600 and 256, respectively. The extracted embeddings are centered and L2-normalized to unit length before cosine similarity scoring.

3.2. Attack detector

Score pairs generated from the twins are fed into an attack detector for consistency validation. A critical question is how to formulate this task: binary classification versus one-class classification. It's common to treat the task as binary classification, e.g. crafting adversarial samples to train a binary classifier for detection [8, 19, 30]. In this way, the classifier's decision boundary depends on the adversarial samples crafted, which means that the classifier may only correctly detect the attack it has seen and could not generalize well to other novel adversarial attacks, as experimentally found in [8, 19].

We argue that one-class classification (OCC) could be a better and simpler alternative. It requires only genuine samples to generate score pairs for training. The decision boundary is irrelevant to adversarial samples, and therefore the classifier could generalize well on unseen attacks. In this work, minimum covariance determinant (MCD) method [21, 31] is adopted. Simply speaking, it fits a single Gaussian model to the data (score pairs of genuine target trials, in our case), with robust statistical estimation of the Gaussian mean and covariance against outliers. From the Gaussian model, a robust tolerance ellipse is then derived as the decision boundary of the detector. By combining the decision boundaries from the premier model and the detector, the trial acceptance region is reduced to defend against impersonation attack.

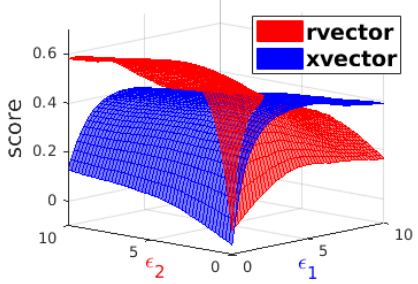

Figure 2: *Score inconsistency between the two SV models, x-vector and r-vector, under adversarial attack*

4. Experiments

4.1. Experimental setup

The experiments are conducted on the Voxceleb1 [32] dataset, which comprises short clips of human speech collected from YouTube. Following its default setup, the training data with $148,642$ utterances from 1211 speakers are utilized to train SV models. In addition, a random training subset with $6,207$ utterances from 50 speakers is utilized to generate $10,000$ target trials for the training of the attack detector. The test data, consisting of $4,874$ utterances from 40 speakers, serves to generate $187,860$ target trials and $18,7860$ nontarget trials for the evaluation in terms of FRR and FAR. The decision thresholds of single-model SV systems are fixed at which FAR = FRR before attack. To conduct impersonation attack, adversarial utterances are crafted only from the nontarget trials of the test data.

Table 1 lists the configuration of four SV models adopted in our experiments. Training details about xvector and rvector have been illustrated in section 3.1. Input speech waveform can be reversed in time (*audio reversed*) before feature extraction. The acoustic features used in our experiments are 30-dim filterbanks (FBanks) and 65-dim log-power magnitude spectrums (LPMSs). Both features are preprocessed by mean normalization over a sliding window of up to 3 seconds, followed by energy-based voice activity detection and global mean-variance normalization. Adversarial utterances are crafted on LPMSs and inverted into speech waveform to conduct attack.

Table 1: *List of SV models used. The back-end is cosine scoring.*

model abbr.	front-end	audio reversed	feature	feat. to waveform
xvec-lpms	xvector	no	LPMS	yes
xvec-inv-fbank	xvector	yes	FBank	no
rvec-lpms	rvector	no	LPMS	yes
rvec-inv-fbank	rvector	yes	FBank	no

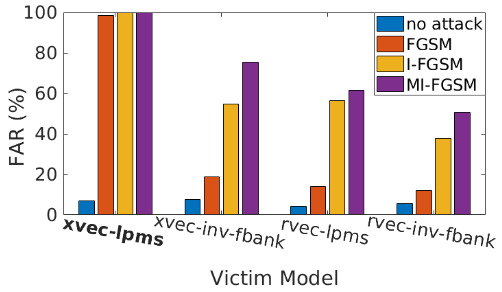

Figure 3: *Transfer impersonation attack crafted from* ***xvec-lpms***

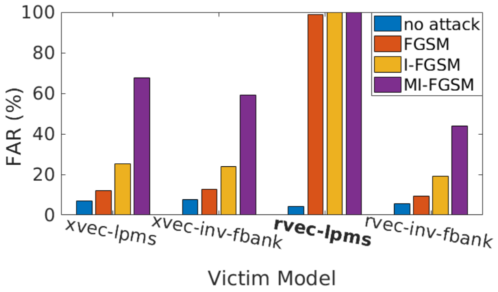

Figure 4: *Transfer impersonation attack crafted from* ***rvec-lpms***

4.2. Adversarial attack on single SV models

Attack results on the four single SV models are shown in Fig. 3-4. Their FARs surge seriously as the attack algorithm progresses from FGSM to MI-FGSM. It is in general hard to defend against adversarial attack with a single SV model, especially when the victim is fully accessed by its adversary. It can also be observed that the attack crafted from xvec-lpms achieves higher FARs on black-box victims (xvec-inv-fbank and rvec-inv-fbank) than the attack from rvec-lpms. In addition, among the victim models, rvec-inv-fbank has the lowest FARs under both attack, showing its robustness against adversarial attack.

4.3. Adversarial defense with the twin SV models

Table 2 shows the superiority of twin SV models over single models in defending against adversarial attack. In this experiment, xvec-lpms and rvec-inv-fbank are set as the premier and the mirror models of the twins. FGSM and MI-FGSM attacks are crafted from xvec-lpms. The strongest ens-MI-FGSM attack mixes perturbations from both xvec-lpms and rvec-lpms by Eq. 9. As shown in Table 2, both the single models suffer from serious FAR increases under attack. For the strongest ens-MI-FGSM attack, FAR of rvec-inv-fbank jumps from 5.72% (before attack) to 63.54%. In contrast to the vulnerable single models, the proposed system with twin models (in the last row) gives both consistent and dramatic robustness improvements. For the strongest ens-MI-FGSM attack, it reduces the FAR from over 63% (on single models) to 2.26%, which is even lower than the FAR before attack (4.48%). It means the decision errors are actually reduced under attack. As a result, the attack is completely defended.

To probe more behind the results in Table 2, score pairs produced by the twins on the test trials are visualized in Fig. 5. The x and y axes represent the scores from xvec-lpms and rvec-inv-fbank, respectively. The genuine nontarget (blue) and target trials (red) lie in the bottom-left and top-right corners. Along either of the axes, the two sets of trials are well-separated, indicating the good performance of single SV models before attack. By impersonation attack, the nontargets are pushed towards the targets so that they intersect more along the x-/y-axis, giving rise to the significant FAR increases of single SV models in Table 2. Nevertheless, they are still, or even more, separable in the 2D space, due to the strong score inconsistency between the twins. The inconsistency is not only because of the different robustness degrees of the twins but also due to the large intensity gap of attack on the twins. In our experiments, ens-MI-FGSM has the strongest attack intensity on single models. But its intensity gap between the twin models is also the largest, resulting in the most separable nontarget clusters (yellow). The attack detector provides an extra decision boundary, the red ellipse curve in Fig.5, to lessen the trial acceptance region, and finally resulting in the dramatic FAR decrease shown in Table 2.

The reduction of acceptance region also means the enlargement of rejection region, which leads to the cost of a slight FRR increase on genuine target trials (7.04% → 8.50%).

Table 2: *FAR% under impersonation attack*

model	no attack	FGSM	MI-FGSM	ens-MI-FGSM
①xvec-lpms	7.04	98.54	100.00	100.00
②rvec-inv-fbank	5.72	12.14	43.90	63.54
twins: ①+②	**4.48**	**8.28**	**2.58**	**2.26**

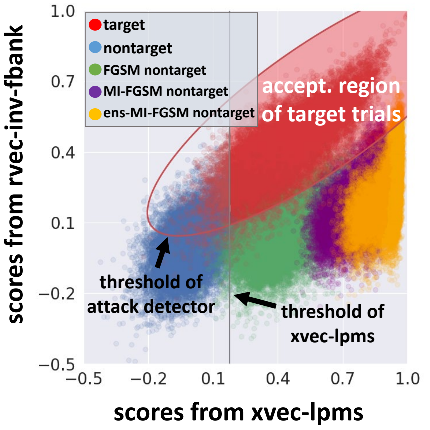

Figure 5: *Score pairs generated from the twins.*

5. Conclusion

Motivated by the score inconsistency between SV models under adversarial attack, we pair two SV models as twins for defense. A vulnerable premier model, together with a relatively-robust mirror model as the twins, can reveal a stealthy attack by their inconsistency of the output verification scores. A simple one-class classifier is adopted to detect the inconsistency. It uses only genuine speech samples for training and circumvents the need of crafting adversarial samples, which improves its defense robustness against unseen attacks. Experiments conducted on VoxCeleb found that the proposed twins system could give both consistent and dramatic performance improvements over single-model systems under different attack algorithms. This pairing method also has practical benefits, e.g., no need to modify a well-deployed SV model even it has been fully-accessed by adversary. Moreover, it can be combined with existing single-model countermeasures for stronger defense.

6. References

[1] A. Nautsch, X. Wang, N. Evans, T. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "Asvspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[2] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[3] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with res2net architecture," *arXiv preprint arXiv:2010.15006*, 2020.

[4] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April, pp. 1962–1966, 2018.

[5] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial Attacks on GMM I-Vector Based Speaker Verification Systems," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, pp. 6579–6583, 2020.

[6] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *arXiv preprint arXiv:2004.08849*, 2020.

[7] J. Villalba, Y. Zhang, and N. Dehak, "x-Vectors Meet Adversarial Attacks: Benchmarking Adversarial Robustness in Speaker Verification," pp. 4233–4237, 2020.

[8] X. Li, N. Li, J. Zhong, X. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Investigating robustness of adversarial samples detection for automatic speaker verification," *arXiv*, pp. 3–7, 2020.

[9] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? adversarial attacks on speaker recognition systems," *arXiv preprint arXiv:1911.01840*, 2019.

[10] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-Box Attacks on Spoofing Countermeasures Using Transferability of Adversarial Examples," pp. 4238–4242, 2020.

[11] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *arXiv preprint arXiv:1710.10766*, 2017.

[12] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[13] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[14] H. Wu, S. Liu, H. Meng, and H.-y. Lee, "Defense against adversarial attacks on spoofing countermeasures of asv," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6564–6568.

[15] X. Yuan, Y. Chen, Y. Zhao, Y. Long, X. Liu, K. Chen, S. Zhang, H. Huang, X. Wang, and C. A. Gunter, "Commandersong: A systematic approach for practical adversarial voice recognition," in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 49–64.

[16] H. Wu, A. T. Liu, and H. yi Lee, "Defense for black-box attacks on anti-spoofing models by self-supervised learning," *arXiv*, pp. 2–6, 2020.

[17] H. Wu, X. Li, A. T. Liu, Z. Wu, H. Meng, and H.-y. Lee, "Adversarial defense for automatic speaker verification by cascaded self-supervised learning models," *arXiv preprint arXiv:2102.07047*, 2021.

[18] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-Septe, 2019, pp. 4010–4014.

[19] S. Samizade, Z.-H. Tan, C. Shen, and X. Guan, "Adversarial example detection by classification for deep speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3102–3106.

[20] X. Li, X. Wu, H. Lu, X. Liu, and H. Meng, "Channel-wise gated res2net: Towards robust detection of synthetic speech attacks," *Proc. Interspeech 2021*, 2021.

[21] M. Hubert, M. Debruyne, and P. J. Rousseeuw, "Minimum covariance determinant and extensions," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 10, no. 3, p. e1421, 2018.

[22] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *arXiv*, 2020.

[23] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[24] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, "Adversarial examples in the physical world," 2016.

[25] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[26] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[27] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[28] 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.

[29] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[30] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," *arXiv preprint arXiv:1704.04960*, 2017.

[31] "Outlier detection on a real data set," https://scikit-learn.org/stable/auto_examples/applications/plot_outlier_detection_wine.html#sphx-glr-auto-examples-applications-plot-outlier-detection-wine-py, accessed 2021-02-16.

[32] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.