Interspeech 2022
18-22 September 2022, Incheon, Korea

NRI-FGSM: An Efficient Transferable Adversarial Attack Method for Speaker Recognition System

Hao Tan[1,2], *Junjian Zhang*[1], *Huan Zhang*[1,2], *Le Wang*[1,2,*], *Yaguan Qian*[3], *Zhaoquan Gu*[1,2,*]

[1] Guangzhou University, Guangzhou, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Zhejiang University of Science and Technology, Hangzhou, China
{th198,zhhuan,2112106069}@e.gzhu.edu.cn,{wangle,zqgu}@gzhu.edu.cn
qianyaguan@zust.edu.cn

Abstract

Deep neural network (DNN), though widely applied in Speaker Recognition Systems (SRS), is vulnerable to adversarial attacks which are hard to detect by humans. The black-box model vulnerability against adversarial attacks is crucial for the robustness of SRS, especially for latest models such as x-vector and ECAPA. The state-of-the-art transferable adversarial attack methods start with generating the adversarial audio from white-box SRS, then utilize this audio to attack the black-box SRS. However, these methods often have a lower success rate in SRS than in the image processing domain. To improve the attack performance on SRS, we propose an efficient Nesterov accelerate and RMSProp optimization based Iterative-Fast Gradient Sign Method (NRI-FGSM), which integrates the Nesterov Accelerated Gradient method and the Root Mean Squared Propagation optimization method with adaptive step size. Through extensive experiments on both close-set speaker recognition (CSR) and open-set speaker recognition (OSR) tasks, our method achieves higher attack success rates of 97.8% for CSR and 61.9% for OSR tasks than others, and meanwhile maintains a lower perturbation rate with signal-to-noise ratio (SNR) and perceptual evaluation of speech quality (PESQ) metrics. It is worth mentioning that our work is the first to attack the ECAPA SRS model successfully.

Index Terms: speaker recognition, transferable attack, nesterov accelerated gradient, root mean squared propagation

1. Introduction

Deep neural networks (DNNs) have been shown vulnerable to adversarial attacks in many fields, such as image classification [1], natural language processing [2] and speech recognition [3]. Recent studies [4, 5, 6, 7, 8] have shown that adversarial attacks also exist in speaker recognition system (SRS). Different from traditional attacks against SRSs, such as speech synthesis [9] and voice conversion attacks [10], the adversarial attacks generate the audio to fool the SRS which is highly similar to the original audio.

As a pioneering work, it is firstly verified in [4] that adversarial attacks exist in SRS by simply adopting the classical Fast Gradient Sign Method (FGSM) [11] in image domain. Following that, more methods have been proposed to attack different types of SRSs, such as i-vector [5], d-vector [6] and x-vector [7]. Generally speaking, the adversarial attacks can be divided into white-box and black-box attacks according to the attackers' prior knowledge of the system. White-box attacks assume that all information of the system including model architecture, parameters and data is transparent and obtainable. On the contrary, black-box attacks assume that the attacker has no or partial information about the system, such as only the output results being obtainable. Obviously, black-box attacks are more realistic and challenging, so we focus on efficient black-box attacks against DNN-based SRSs in this paper.

Transferable attack is a commonly adopted method for black-box attacks, which generates an adversarial audio from a white-box model first and then uses this audio to attack a black-box model. For example, the Momentum-based Iterative FGSM (MI-FGSM) [12] method, which is adopted from the image domain to attack the anti-spoofing models [13], has been demonstrated very effective against the black-box models such as AFNet [14]. Some other gradient-based white-box attack methods have also been utilized to attack SRSs [15, 16].

However, the success rate of those transferable attacks is in a low range for the black-box SRS compared with the much higher rate they achieve in the image domain. The gradient-based attack method greedily perturbs the audios in line with the direction of the sign of the gradient at each iteration, which may easily fall into the local maximum and fail to attain global optimization, resulting in low performance in SRSs.

To improve the attack success rate and achieve global optimization for the black-box SRS, we propose a multi-step iterative attack method called NRI-FGSM, which represents Nesterov accelerate gradient (NAG) and Root Mean Squared Propagation (RMSProp) optimization based Iterative-Fast Gradient Sign Method. First, we adopt the NAG [17] optimization method to generate the adversarial audio that is close to the white-box model's decision boundary, which is similar to SI-NI-FGSM (Scale Invariant and Nesterov Accelerate Gradient Iterative FGSM) [18]. Compared to the traditional gradient-based methods, NAG can take a step forward and meanwhile stabilize the direction of the gradient, which will correct the previous accumulated gradient and thus avoid local maximum. Second, we combine the RMSProp [19] optimization method with adaptive step size and momentum to optimize the step size dynamically. This proposed NRI-FGSM can assist in studying the vulnerability of the black-box SRS under transferable attacks.

The main contributions of this paper are as follows: (1) We have explored the robustness of the latest SRS models, such as the x-vector and the ECAPA [20]. To the best of our knowledge, we are the first to attack the ECAPA-based SRS model. (2) The proposed NRI-FGSM is an efficient method to attack the black-box SRS, which generates the adversarial audio with high transferability. (3) Extensive experiments have been done to evaluate the effectiveness of the proposed NRI-FGSM method com-

* Corresponding author.

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10499

pared to some other state-of-the-art transferable attack methods. The results show that NRI-FGSM achieves the attack success rates of 97.8% for close-set speaker recognition (CSR) tasks and 61.9% for open-set speaker recognition (OSR) tasks.

2. Methodology

2.1. Gradient-based Transferable attacks

Black-box transferable attacks can employ some particular white-box alternative models to craft adversarial examples using gradient-based attacks, which utilize the gradients to generate perturbations that can maximize the loss function of misrecognition. Existing works generate the adversarial example iteratively by Eq. (2), or plugs in the momentum as Eq. (1):

$$\mathbf{g}_{i+1} = m \cdot \mathbf{g}_i + \frac{\nabla L(f(x_i^{adv}), y^t)}{\| \nabla L(f(x_i^{adv}), y^t) \|_1}; \tag{1}$$

$$x_{i+1}^{adv} = x + clip_{\varepsilon}(x_i^{adv} + \sigma \cdot sign(\mathbf{g}_{i+1}) - x), \tag{2}$$

where $\sigma = \varepsilon/T$, ε is the budget of adversarial perturbations, T is the maximum iteration, $i \in T$, $\mathbf{g}_i$ is the computed gradient in the i-th iteration, m is the momentum factor, x_i^{adv} is the generated adversarial example in the i-th iteration. When $T = 1$, $m = 0$, it is reduced to FGSM [11], which is a single-step attack method. If we prepend single-step attacks by a small random perturbation, which is named R-FGSM [21]. When $T > 1$, $m = 0$, it is reduced to the basic iterative method (BIM) [22], which uses multi-step iteration with slight perturbation. When $T > 1$, $m = 0$, and σ starting with a random perturbation in ε-ball, it is reduced to the projected gradient descent method (PGD) [23]. When $T > 1$, $m \neq 0$, it is reduced to the momentum-based fast gradient sign method (MI-FGSM) [12].

2.2. NRI-FGSM

Previous methods of transferable attacks can easily fall into local maximum during optimization, such as BIM and PGD, due to their greedy search with a fixed step size. The momentum accelerated gradient methods such as MI-FGSM can eliminate the limitation of greedy search to some extent. However, the fixed step size during each iteration will also affect the quality of the generated adversarial examples. This quality loss is much more serious when it is adopted for attacking SRS. Hence, an efficient transferable attack is needed to improve the attack performance.

To address these problems, we propose an efficient transferable attack method called NRI-FGSM, which combines the NAG and the RMSProp optimization methods with an adaptive step size. Specifically, we adapt the NAG method to escape from the local maximum during optimization, as verified in [18] with the looking forward property. To improve the efficiency, we combine the RMSProp optimization method, which can accelerate the optimization process. We also introduce an adaptive factor μ during the process.

As shown in Alg. 1, before calculating the gradient in each iteration, we make a forward step in the direction of the previously accumulated gradient, which has the same step size as last iteration. The perturbation step is optimized based on the sum of squared gradient and accumulated gradients. The update processes of NRI-FGSM can be formalized as follows:

$$x_t^{nes} = x_t^{adv} + \beta_1 \cdot \mathbf{g_t} \cdot \alpha_t; \tag{3}$$

$$\mathbf{g}_{t+1} = \beta_1 \cdot \mathbf{g_t} + \frac{\nabla L(f(x_t^{nes}, y^t)}{\| \nabla L(f(x_t^{nes}), y^t) \|_1}; \tag{4}$$

Algorithm 1 NRI-FGSM

Input: White-Box model M, clean audio x with true label y, the budget of adversarial perturbation ε, Gaussian perturbation δ, step size α, max iteration T, momentum factor β_1, RMSProp factor β_2, adaptive factor μ, extra step κ.

Output: Adversarial example x_{adv}

$\alpha \leftarrow \varepsilon$ / T
$\beta_1 \leftarrow 1$; $\beta_2 \leftarrow 0.9$; $\mu \leftarrow 1$; $\mathbf{g}_0 \leftarrow 0$
for $t = 0$ to $T - 1$ **do**
  Add a small random perturbation δ in the ε-ball
  Get x_t^{nes} by Eq. (4)
  Get the forward gradients by $\nabla L(f(x^{nes}, y')$
  Accumulate the gradients by Eq. (5)
  Update the step size α_{t+1} by Eq. (7)
  Update x_{t+1}^{adv} by $clip_{\varepsilon}\{x_t^{adv} + \alpha_{t+1} \cdot sign(\mathbf{g}_{t+1})\}$
  Get the adversarial label y' from $\boldsymbol{M}(x_{t+1}^{adv})$
  if y' = Target **then**
    Continue extra κ iterations
    return $x_{t+1+\kappa}^{adv} \leftarrow x_{t+1}^{adv}$
  end if
end for
return x_T^{adv}

$$r_{t+1} = \beta_2 \cdot r_t + (1 - \beta_2) \cdot \mathbf{g}_{t+1}^2; \tag{5}$$

$$\alpha_{t+1} = \mu \cdot \alpha_t \cdot \frac{\mathbf{g}_{t+1}}{\sqrt{r_{t+1} + \delta}}; \tag{6}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha_{t+1} \cdot sign(\mathbf{g}_{t+1}), \tag{7}$$

where β_1 is the momentum factor, $\mathbf{g}_t$ is used to accumulate the gradient and $\mathbf{g}_0 = 0$. β_2 is the exponential decay rate which is similar to the RMSProp method. r is used to accumulate the gradient for adaptive step size. μ is the adaptive factor for the step size, which plays a key role in improving the transferability of the generated adversarial audio. α is the adaptive step size with an initial value $\alpha = \varepsilon/T$, and T is the maximum number of iterations. During each iteration, α changes with the current gradient and the cumulative gradient, such as a big step forward on a large gradient.

The proposed method takes advantage of the forward-looking property of NAG and the real-time optimization step of RMSProp to generate more robust adversarial audio. In order to improve the transferablity of attacking black-box models, our algorithm stops κ iterations after the adversarial audio is misrecognized to the target label. Although many attack methods terminate upon a successful attack, it is found from extensive experiments that the generated adversarial audio has better attack performance with extra κ steps.

3. Experiment setup

3.1. Datasets

The experiments are conducted on the LibriSpeech [24] dataset, which is follow [25]. For each speaker, 10 audio are selected for the registration phase. For the imposter dataset, there are 10 speakers with different identities from the 10 speakers in the enrolling set, each speaker has 100 utterances.

3.2. Tasks

Speaker recognition include close-set and open-set tasks. The CSR task can be regarded as a classification problem, which

identifies a specific speaker from the corpus of a set of enrolled speakers. The system always identifies an input audio as a specific speaker in the training dataset. In the OSR task, the model obtains a threshold by the probabilistic linear discriminant analysis (PLDA) [26] or the cosine similarity (cos-sim) algorithm, and recognizes the test utterance as an enrolled speaker by comparing the calculated similarity score and the preset threshold. Besides, OSR tasks can also identify unknown speakers. A speaker who is not in the original training dataset can also be enrolled in the OSR system to generate specific feature vectors. In the verification process, if the maximum score from the back-end scoring module is below the preset threshold, the speaker will be identified as an unenrolled speaker and rejected.

3.3. Victim models

Among the existing DNN-based SRS models, we choose four advanced models which are x-vector [7], Res34-L [27], Res34-V [28] and ECAPA [20] based models. Specifically, all of them can be adopted for the CSR and OSR tasks. The details of these models are summarized in Table 1.

Table 1: *Details and performance of four advanced models. Arch and "θ" represent architecture and the threshold in the OSR task, respectively.*

Arch	Model	Feature	Score	CSR	OSR	
				ACC(%)	EER(%)	θ
CNN	**Res34-L**	FBank	cos-sim	99.9	2.8	0.72
	Res34-V	FBank	cos-sim	100	0.9	0.62
TDNN	**x-vector**	MFCC	PLDA	100	6	18.72
	ECAPA	FBank	cos-sim	100	1	0.51

3.4. Evaluation metrics

Generally, a high-quality adversarial example should own a strong attack ability with undetectable noise. In our experiments, we use attack success rate (ASR), signal-to-noise ratio (SNR), and perceptual evaluation of speech quality (PESQ) to evaluate the adversarial examples.

The ASR is the success rate at which the attacker fools SRS to make a wrong decision. The SNR is the ratio between the power values of audio and noise, such as $SNR = 10 \cdot \log_{10}(P_x/P_\delta)^2$, where P_x is the power of the original example and P_δ is the power of the perturbation. PESQ [29] is an objective index of speech quality, which can directly and truly reflect the real situation of speech quality. The rating range of the PESQ is from 0 to 5. In addition, we measure the Euclidean distance between the adversarial and original audio. Normally, higher SNR and PESQ values, smaller L_2 distance indicate better quality of the audio and can be used to evaluate whether the audio can be heard clearly.

4. Experiment results and analysis

4.1. CSR target attack

We set up the CSR tasks in three models: ECAPA, Res34-L and x-vector. We make one of them as the white-box model while the others are black-box models. We then generate the adversarial audio against the white-box model with a target label, and use the audio to attack the other two black-box models. Both attacks are conducted under $\varepsilon = 0.001$ with 10 iterations and $\kappa = 10$. As we can see in Table 2, PGD, MI-FGSM and NRI-FGSM show great aggressiveness for white-box attack, which achieve the success rate at 100%. However, compared to PGD and MI-FGSM in transfer attack, our proposed NRI-FGSM has the highest transferability among all models.

Table 2: *The ASR(%) of CSR target attack. We generate adversarial audio from the alternative model (Alter-Model), and use them to attack the victim models. "underline" means our methods own the same performance as PGD and MI-FGSM.*

Alter-Model	Methods	Victim Model			
		ECAPA	**Res34-L**	**Res34-V**	**x-vector**
ECAPA	FGSM	96.7	75	84.8	78.6
	PGD	100	81.1	89	83.4
	MI-FGSM	100	83.1	90.1	87.7
	NRI-FGSM*	100	**85.8**	**93.1**	**92.4**
Res34-L	FGSM	70.9	83.4	82.6	79.4
	PGD	79.2	100	91.1	87.4
	MI-FGSM	82.7	100	93	89.4
	NRI-FGSM*	**93**	100	**97.8**	**92.6**
x-vector	FGSM	65.6	74.6	72.9	79.6
	PGD	73.5	84.4	81	100
	MI-FGSM	75	86.4	82.4	100
	NRI-FGSM*	**84.6**	**92**	**90.9**	100

When we set $\mu = 1$, the average SNR of the adversarial audio by our method is 34.26 dB, and the PESQ is 3.35, which implies the audio has good quality and is insensible to human.

4.2. OSR target attack

Different from the CSR task, the OSR target attack not only assumes the target label has the maximum confidence score of the adversarial audio, but also requires the score is larger than a preset threshold. Transfer attacks in the OSR task are more difficult than that in CSR task under the same settings. For instance, as shown in Fig. 1, under the same settings in Sec 4.1, the ASR for the CSR task is significantly higher than that for the OSR task. To find the best κ, we use the adversarial examples generated from ECAPA model with different κ to attack other models, such as the Res34-L model. As shown in Fig. 2, we can easily observe that the transfer attack success rate with different algorithms reach the highest value when $\kappa \in [20, 40]$.

We set up the OSR tasks in four models, i.e. ECAPA, Res34-L, Res34-V and x-vector, and use the adversarial audio that is generated against the white-box model to attack the other three black-box models. Specifically, we tested L_∞ norms ε between 0.001 and 0.1 with 100 iterations and $\kappa = 40$. Due to the page limits, we only show the results of $\varepsilon = 0.002$ in Table 3.

When attacking the black-box models, as expected the ASR is decreased for all adversarial attacks, while our method can still achieve the best performance. The average of SNR and PESQ are 27.98 dB and 2.55.

4.3. Imperceptibility

Some transferable attacks can achieve high attack rates, but meanwhile greatly increase the adversarial perturbation. In our experiments, we set $\varepsilon = 0.001$, $\mu = 0.7$, $T = 10$, $\kappa = 0$ to evaluate the performance of the proposed NRI-FGSM method. As shown in Table 4, our method achieves higher attack success rate while maintaining a lower perturbation rate of the adversarial audios according to the SNR and PESQ metrics.

Table 3: *The ASR(%) of OSR target attack on four models by six gradient-based iterative attacks.*

Alter-Model	Methods	Victim Model			
		ECAPA	Res34-L	Res34-V	x-vector
ECAPA $\theta = 0.51$	FGSM	64.4	18.1	12.2	6.9
	R-FGSM	75.4	20.7	13.3	7.4
	BIM	100	31.4	17	17.3
	PGD	100	32.1	22.6	16.9
	MI-FGSM	100	44.8	32.4	19.5
	NRI-FGSM*	100	**54.9**	**46.1**	**24.4**
Res34-L $\theta = 0.72$	FGSM	8	62.1	8	4.4
	R-FGSM	16	73.9	8.9	5.2
	BIM	14	100	10.2	5.6
	PGD	15.2	100	11.8	6.1
	MI-FGSM	24.7	100	21.5	11.8
	NRI-FGSM*	**28.9**	100	**27.5**	**14.2**
Res34-V $\theta = 0.62$	FGSM	14	17.2	47.6	3.6
	R-FGSM	18.3	20.2	58.9	3.5
	BIM	22.4	39.8	100	7.8
	PGD	23.7	40.7	100	8.2
	MI-FGSM	36.1	60.9	100	15.9
	NRI-FGSM*	**42.6**	**61.9**	100	**19.4**
x-vector $\theta = 18.72$	FGSM	16.1	21.5	7.2	50.5
	R-FGSM	18.2	23.1	6.9	64
	BIM	17.5	20.6	7.2	100
	PGD	17.7	19.4	7.1	100
	MI-FGSM	25.1	30.7	12.4	100
	NRI-FGSM*	**31.5**	**38.5**	**16.3**	100

Table 4: *The performance of OSR target attack.*

Methods	Victim Model		Imperceptibility		
	Res34-L	x-vector	SNR ↑	PESQ ↑	L_2 ↓
PGD	12.7	7.1	47.83	4.39	0.07
MI-FGSM	12.9	7.1	46.1	4.27	0.09
NRI-FGSM*	**15.6**	**8**	**48.96**	**4.42**	**0.06**

4.4. Performance Under Defense Settings

To further validate the effectiveness of the proposed method, we attack the advanced model with a series of defense modules, including Quantization (QT) [30], Audio Turbulence (AT) [25], Median Smoothing (MS) [30], Down Sampling (DS) [30], Low Pass Filter (LPF) [31] and Band Pass Filter (BPF) [32]. The adversarial examples are crafted from the ECAPA model to attack Res34-L with the same settings as Sec 4.1, and the details of defense parameters are following [25].

As shown in Table 5, compared to MI-FGSM, the adversarial samples generated by the proposed method can achieve better performance on the defense model with an average success rate improvement of 14.34%.

Table 5: *The ASR(%) of OSR target attack against six advanced defense methods. "-" denotes no defense.*

Methods	-	QT	AT	MS	DS	LPF	BPF
PGD	32	1.9	6.3	14.8	9.6	31.2	20.4
MI-FGSM	41.2	3.4	12.6	20.7	13.5	40.4	28
NRI-FGSM*	**57.2**	**13.1**	**35.4**	**31.5**	**28.5**	**56.7**	**39.4**

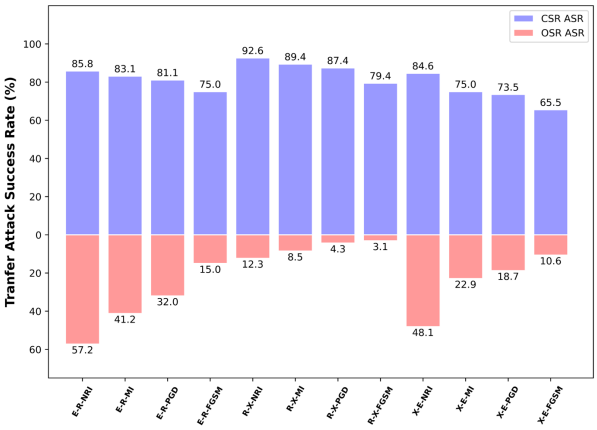


Figure 1: *Attack success rate of different attacks in CSR and OSR with the same parameters. "A-B-C" indicates that using the adversarial audio generated from model A based on algorithm C to attack black-box model B; "E" denotes ECAPA, "R" denotes Res34-L, "X" denotes x-vector; "NRI" and "MI" are short for NRI-FGSM and MI-FGSM.*

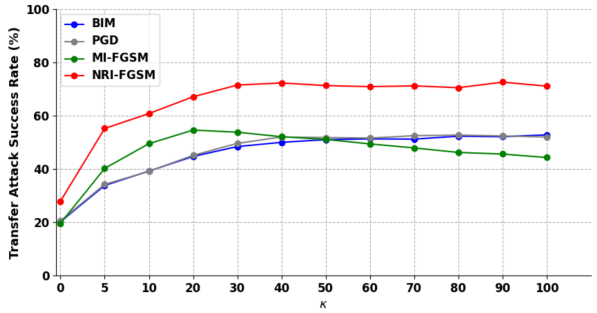


Figure 2: *Transfer Attack Success Rate (%) under different settings of κ. The adversarial examples are crafted from ECAPA.*

5. Conclusions

In this paper, we propose an efficient transferable adversarial attack method denoted NRI-FGSM, which generates the adversarial audio based on the NAG and RMSProp methods. By setting the adaptive step size, we are able to achieve a higher success rate to attack the black-box SRS compared to other transferable attack methods. In addition, the generated adversarial audio, with a high SNR of 48 dB, is quite similar to the original audio, which is a significant improvement compared to the existing attack methods. In the future, we will further improve the transferability of the adversarial audio by combining some other optimization algorithms like ensemble learning.

6. Acknowledgements

This work is supported in part by the Major Key Project of PCL (No. PCL2022A03), the Natural Science Foundation of China (No. 61902082), Guangzhou Science and technology planning project (No. 202102010507), the Key Laboratory of MIIT for Intelligent Products Testing and Reliability 2021 Key Laboratory Open Project Fund (No. CEPRE2022-02), and the Guangzhou University Graduate Student Innovation Ability Cultivation Funding Program (No. 2021GDJC-M34).

7. References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014*, Y. Bengio and Y. LeCun, Eds., 2014.

[2] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020.

[3] M. Cissé, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured prediction models," *CoRR*, vol. abs/1707.05373, 2017.

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

[5] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*. ISCA, 2009, pp. 1559–1562.

[6] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014*. IEEE, 2014, pp. 4052–4056.

[7] 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 2018*. IEEE, 2018, pp. 5329–5333.

[8] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso, "Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021*. IEEE, 2021, pp. 6159–6163.

[9] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013*, pp. 7962–7966.

[10] Z. Wu, T. Virtanen, T. Kinnunen, E. Chng, and H. Li, "Exemplar-based unit selection for voice conversion utilizing temporal information," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association 2013*, pp. 3057–3061.

[11] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 39–57.

[12] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 9185–9193.

[13] Y. Zhang, Z. Jiang, J. Villalba, and N. Dehak, "Black-box attacks on spoofing countermeasures using transferability of adversarial examples," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, H. Meng, B. Xu, and T. F. Zheng, Eds., pp. 4238–4242.

[14] C. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019*, pp. 6316–6320.

[15] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020*, pp. 6579–6583.

[16] J. Villalba, Y. Zhang, and N. Dehak, "x-vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 4233–4237.

[17] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2)," in *Doklady an ussr*, vol. 269, 1983, pp. 543–547.

[18] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

[19] S. Ruder, "An overview of gradient descent optimization algorithms," *CoRR*, vol. abs/1609.04747, 2016.

[20] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, pp. 3830–3834.

[21] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017,*. OpenReview.net, 2017.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018,*. OpenReview.net, 2018.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015*. IEEE, 2015, pp. 5206–5210.

[25] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "SEC4SR: A security analysis platform for speaker recognition," 2021. [Online]. Available: https://arxiv.org/abs/2109.01766

[26] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castán, and A. Lawson, "Analysis of critical metadata factors for the calibration of speaker recognition systems," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 4325–4329.

[27] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, and I. Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[28] H. Zeinali, S. Wang, A. Silnova, P. Matejka, and O. Plchot, "BUT system description to voxceleb speaker recognition challenge 2019," *CoRR*, vol. abs/1910.12592, 2019.

[29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[30] Z. Yang, B. Li, P. Chen, and D. Song, "Characterizing audio adversarial examples using temporal dependency," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[31] H. Kwon, H. Yoon, and K. Park, "POSTER: detecting audio adversarial example through audio modification," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*. ACM, 2019, pp. 2521–2523.

[32] K. Rajaratnam, B. Alshemali, K. Shah, and J. Kalita, "Speech coding and audio preprocessing for mitigating and detecting audio adversarial examples on automatic speech recognition," 2018.