INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Pseudo-Siamese Network based Timbre-reserved Black-box Adversarial Attack in Speaker Identification

*Qing Wang, Jixun Yao, Ziqian Wang, Pengcheng Guo, Lei Xie**

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University (NPU), Xi'an, China

qingwang@nwpu-aslp.org, {yaojx, zq_wang}@mail.nwpu.edu.cn, guopengcheng1220@gmail.com, lxie@nwpu.edu.cn

Abstract

In this study, we propose a timbre-reserved adversarial attack approach for speaker identification (SID) to not only exploit the weakness of the SID model but also preserve the timbre of the target speaker in a black-box attack setting. Particularly, we generate timbre-reserved fake audio by adding an adversarial constraint during the training of the voice conversion model. Then, we leverage a pseudo-Siamese network architecture to learn from the black-box SID model constraining both intrinsic similarity and structural similarity simultaneously. The intrinsic similarity loss is to learn an intrinsic invariance, while the structural similarity loss is to ensure that the substitute SID model shares a similar decision boundary to the fixed black-box SID model. The substitute model can be used as a proxy to generate timbre-reserved fake audio for attacking. Experimental results on the Audio Deepfake Detection (ADD) challenge dataset indicate that the attack success rate of our proposed approach yields up to 60.58% and 55.38% in the white-box and black-box scenarios, respectively, and can deceive both human beings and machines.

Index Terms: speaker identification, adversarial attack, black-box, timbre-reserved

1. Introduction

Speaker identification (SID) [1, 2] is a process of determining the identity of the person who spoke a particular speech. As a type of biometric identification, it is critical to ensure the security of the speaker identification system. SID system usually confronts various kinds of attacks, such as spoofing attacks [3–5] and adversarial attacks [6, 7]. Spoofing attack commonly includes impersonation, replay, voice conversion, and speech synthesis. Recently, adversarial attack has emerged as a significant threat to the accuracy and reliability of speaker identification systems, which can be defined as malicious attempts to deceive a machine learning model, including speaker identification systems, by exploiting their vulnerabilities.

Many researchers have successfully conducted adversarial attacks on SID systems [8–16]. Das *et al.* [17] gave an overview of various types of attack on speaker verification focusing on potential threats of adversarial attacks and spoofing countermeasures from the attacker's perspective. Li *et al.* [18] launched a practical and systematic adversarial attack against speaker recognition systems and integrated the estimated room impulse response into the adversarial example training for over-the-air attack. To constrain the perceptibility of the adversarial perturbation and perform targeted speaker attack, in [19], Wang *et al.* generated inaudible adversarial perturbations based on the psychoacoustic principle of frequency masking. The gradients via back-propagation are essential when crafting the adversarial examples in these methods. However, in the black-box settings, it's hard to generate an adversarial example by estimating the gradient. Zhang *et al.* [20] performed black-box waveform-level targeted adversarial attacks against speaker recognition systems by generating imperceptible adversarial perturbations based on auditory masking. In [21], Chen *et al.* proposed FAKEBOB to craft adversarial examples and conducted a comprehensive and systematic study of the adversarial attacks on speaker recognition systems to understand their security weakness in the practical black-box setting.

Moreover, inspired by the aforementioned spoofing attack and adversarial attack, generating quality fake audio to effectively attack the SID model requires the ability to deceive both machines and humans simultaneously. To deceive machines, we need to consider the downstream SID task and ensure that the fake audio possesses a distinctive speaker attribute that makes the SID model give the desired decision. From the human perception perspective, fake audio with noticeably different timbre or text from the target speaker's real audio is easily detectable. Thus, timbre and text information also need to be taken into account when conducting attacks on the SID model.

To this end, in this study, we propose a novel timbre-preserved adversarial attack approach for speaker identification to exploit the weakness of the SID model while preserving the timbre of the target speaker in the black-box attack setting. To achieve this, we generate timbre-preserved fake audio by incorporating an adversarial constraint during the training of the voice conversion (VC) model. We then utilize a pseudo-Siamese network architecture to learn from the black-box SID model, constraining both intrinsic similarity and structural similarity simultaneously. The intrinsic similarity loss is aimed at achieving an intrinsic invariance, while the structural similarity loss ensures that the substitute speaker classifier shares a similar decision boundary to the fixed black-box speaker classifier. By using the substitute speaker classifier as a proxy, we can generate timbre-preserved fake audio for the black-box classifier. Experimental results on a partial Aishell-3 dataset [22] indicate that the substitute speaker classifier is proximate to the black-box speaker classifier. Meanwhile, the results of experiments conducted on the Audio Deepfake Detection (ADD) challenge dataset [23] show that the fake audio generated based on our proposed method, which preserves the original timbre, can successfully deceive both humans and machines with a comparable attack success rate.

The rest of the paper is organized as follows. In Section 2, we detail the proposed timbre-reserved black-box adversarial attack in the SID system. Datasets and experimental setup are described in Section 3. Section 4 presents the experimental re-

* Lei Xie is the corresponding author.

 10.21437/Interspeech.2023-1352

sults and analysis. Finally, we conclude in Section 5.

2. Methodology

2.1. Timbre-reserved Adversarial Attack

Figure 1 illustrates our timbre-reserved adversarial attack, which uses a non-autoregressive-based voice conversion model and an attack constraint process. We use a trained speaker identification model $f(\cdot)$ in the adversarial constraint process. Mel-spectrogram M is the input, representing the VC model predicted representation, and its corresponding speaker label predicted by the SID model is y, while the VC target speaker label is y'. Adversarial constraint δ can be defined as follows:

$$\min L_{CE}(f(M+\delta), y'), \qquad \text{s.t.} \quad \|\delta\| < \epsilon, \tag{1}$$

here, $L_{CE}(\cdot)$ is the loss function and ϵ is the hyperparameter. For each iteration, δ is updated as follows:

$$\delta \leftarrow \text{clip}_\epsilon(\delta - lr \cdot \text{sign}(\nabla_\delta L_{CE}(f(M+\delta), y'))). \tag{2}$$

The adversarial constraint is added during the VC model training to preserve target speaker information. For joint training with the adversarial constraint, we use the Mel-spectrogram predicted by the VC model to attack the speaker classifier. If the attack fails, we add a tiny adversarial perturbation to the predicted Mel-spectrogram to generate the adversarial Mel-spectrogram M_{adv} which can be defined as:

$$\begin{aligned} &M_{\text{adv}} = \hat{M} + \delta_{\text{adv}} \\ &\text{s.t.} \quad \|\delta_{\text{adv}}\| < \epsilon, \end{aligned} \tag{3}$$

here, δ_{adv} represents the tiny adversarial perturbation and ϵ is used to control the maximum adversarial perturbation generated. The tiny adversarial perturbation can be optimized by:

$$\min \mathcal{L}_{\text{CE}}\left(f(M_{\text{adv}}), y'\right), \tag{4}$$

where $\mathcal{L}_{\text{CE}}$ aims to make the M_{adv} fool the well-trained speaker identification system into predicting a specified target label. Therefore, the joint training with adversarial constraint can be optimized by the following loss function:

$$\mathcal{L}_{\text{adv}} = \begin{cases} \|M_{gt} - \hat{M}\|_1, & if \;\; succeeded, \\ \|M_{\text{adv}} - \hat{M}\|_1, & if \;\; failed. \end{cases} \tag{5}$$

When $f(\hat{M}) \neq y'$, in other words, the attack failed, we add an adversarial perturbation to the Mel-spectrogram as the adversarial constraint. In order to force the predicted Mel-spectrogram $\hat{M}$ of the VC model can be classified to the target speaker, we expect to minimize the L1 loss between the predicted Mel-spectrogram $\hat{M}$ and the Mel-spectrogram with the adversarial perturbation M_{adv} so that the VC model can fool the well-trained speaker identification system. The adversarial perturbation δ_{adv} is optimized by Equation 2 until the predicted label $f(\hat{M})$ is the target label. In contrast, when the speaker classifier gives a prediction of the target speaker label, which means the attack succeeds, the VC model is optimized only using the original reconstruction loss.

When in a more practical scenario, there is no prior knowledge of the black-box setting speaker classifier's architecture or parameters. As a result, we train a substitute speaker classifier to conduct the black-box adversarial attack.

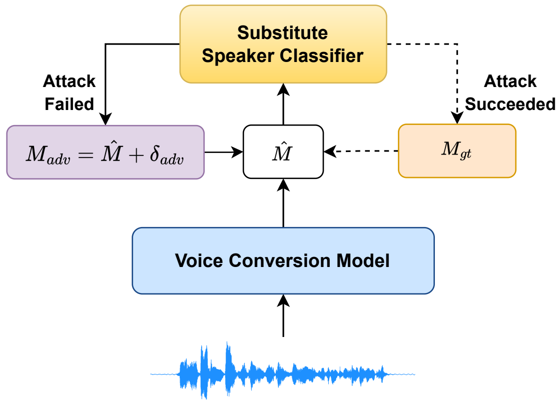

Figure 1: *An overview of timbre-reserved adversarial attack.*

2.2. Pseudo-Siamese Network based Substitute Speaker Classifier

Figure 2 is an overview of the pseudo-Siamese network based architecture to learn both the **intrinsic similarity** and **structural similarity** from the black-box speaker classifier. In particular, the intrinsic similarity enables learning of intrinsic invariance, while the structural similarity guarantee that the substitute speaker classifier shares a similar decision boundary to the fixed black-box classifier. Since there is no prior knowledge of the target SID system's architecture or parameters, we can only access the input and output of the black-box speaker classifier but not the internal processing.

The intrinsic similarity loss is minimized by comparing the original speech and transformed speech, with the aim of assisting the substitute speaker classifier in learning with disturbed labels. It is expected that the posterior probability of the same speaker should remain consistent before and after any transformations. Specifically, for a given speech sample x_0, a transformation was applied by adding Gaussian noise, resulting in a transformed speech sample x_1, which increases the diversity of the data distribution. The intrinsic similarity loss is calculated by comparing the output probabilities of x_0 and x_1:

$$\mathcal{L}_{\text{ins}} = KL(p_1 || {p_1}'), \tag{6}$$

where $\mathcal{L}_{\text{ins}}$ is the intrinsic similarity loss and p_1 and ${p_1}'$ is the output posterior probability of x_0 and x_1 from the substitute speaker classifier, while $KL(\cdot)$ is the Kullback-Leibler (KL) divergence [24].

The structural similarity loss is calculated by comparing the posterior probability distributions of speech samples from substitute and black-box speaker classifiers. In particular, for a given speech sample x_0 and transformed speech sample x_1, its structural similarity loss is defined as follows:

$$\mathcal{L}_{\text{aux}} = KL({p_1}' || p_2), \tag{7}$$

$$\mathcal{L}_{\text{str}} = KL(p_1 || p_2) + \mathcal{L}_{\text{aux}}, \tag{8}$$

where $\mathcal{L}_{\text{str}}$ is the structural similarity loss, $\mathcal{L}_{\text{aux}}$ is the auxiliary similarity loss, and p_2 is the posterior probability of original speech x_0 from the black-box speaker classifier. The final training objective is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ins}} + \mathcal{L}_{\text{str}}, \tag{9}$$

where $\mathcal{L}_{\text{total}}$ is the total loss to constrain both structural and intrinsic similarities.

The pseudo-Siamese network architecture is utilized because it can effectively capture the similarity between the different outputs of the same speech samples from different classifiers. Overall, the training procedure for the substitute speaker classifier aims to learn a model that behaves similarly to the black-box speaker classifier in terms of predicting speaker identities, while also being able to withstand adversarial attacks. After the substitute speaker classifier is trained, the attack loss is calculated as described in Section 2.1.

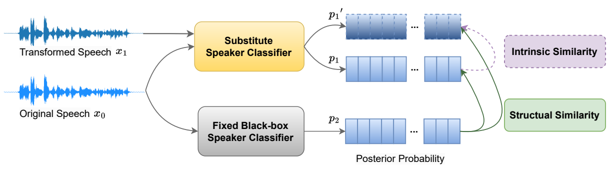

Figure 2: *An overview of pseudo-Siamese network based substitute speaker classifier.*

2.3. Generation of Timbre-reserved Fake Audio

After the VC model with adversarial constraints is trained, the generation of timbre-reserved black-box fake audio is shown in Figure 3. Firstly, given a text, we generate audio for a random speaker using a TTS system based on the FastSpeech [25] model. The encoder and decoder structure of the TTS model is modified to incorporate the conformer block from DelightfulTTS 2 [26]. The speaker ID and the TTS-generated audio are then used as inputs for the adversarially constrained VC model to predict the attack Mel-spectrogram. Finally, a HifiGAN vocoder [27] is used to reconstruct the waveform from the Mel-spectrogram, resulting in timbre-reserved fake audio that can be used for the adversarial attack against the fixed black-box SID model.

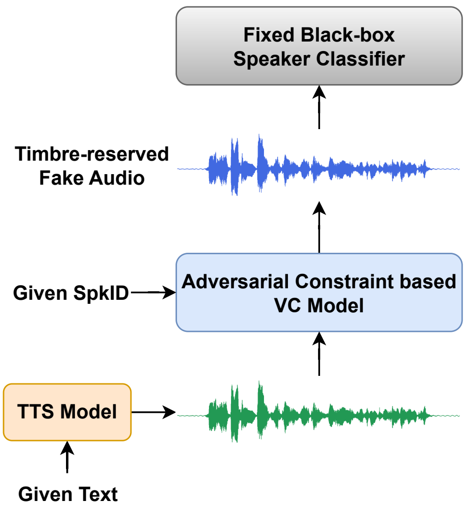

Figure 3: *The generation of timbre-reserved fake audio.*

3. Experimental Setup

3.1. Datasets

In this study, we use AISHELL-3 [22] to train the pseudo-Siamese network, speaker identification model, and voice conversion model. AISHELL-3 is a multi-speaker Mandarin Chinese audio corpus containing 88,035 recordings from 218 native speakers. The test set of AISHELL-1 [28] is used to evaluate the performance of the SID model. To evaluate our proposed method for the timbre-reserved adversarial attack, we employed the dataset of the audio deepfake detection (ADD) challenge [23], which is an open-source dataset designed for deep fake audio attack and detection. We then evaluate our proposed method on the test set of ADD challenge Track 3.1, which consists of 10 speaker IDs and 500 texts used to generate fake audio. Our generated attack samples are required to meet certain standards of intelligibility and similarity while still being able to fool the SID model. By using ADD challenge dataset, we develop a solution to address adversarial attacks while maintaining timbre and allowing for customized texts.

3.2. Setup

The detailed experimental setup of all the models shown in Figure 3 is described as follows:

- TTS model: 6-layer conformer encoder-decoder structure similar to DelightfulTTS 2 [26] is used to generate waveforms from given text with a random speaker timbre.
- VC model: Pre-trained ASR model [29] is used to extract speaker-independent linguistic information. A non-autoregressive VC model is adopted using an 8-layer transformer encoder-decoder structure similar to FastSpeech2 [25] and HifiGAN vocoder with multi-band processing [27]. The speaker classifier is added to adversely constrain distinctive speaker information to generation.
- Speaker identification model: ECAPA-TDNN [30] black-box SID model used in this study, is also used as a speaker classifier in voice conversion. EER on AISHELL-1 test set is 1.91%. Incorporates 3 SE-Res2Block modules with channel size and bottleneck dimension set to 1024 and 256, respectively. AAM-softmax loss function [31] with a margin of 0.2 and scale of 30. The substitute speaker classifier consists of six conformer [32] blocks arranged in a stack followed by a statistics pooling layer which is similar to x-vector [33].
- Mel-spectrogram adversarial constraint experiments: VC model learning rate in Equation 2 is set to 8e-4 and δ is updated 1000 times for each mini-batch. l_∞ norm is used to measure perturbation bound. ϵ starts from 0.8. The substitute speaker classifier is added to adversely constrain distinctive speaker information to generation.

4. Experimental Result

4.1. Performance of Substitute SID Model

The performance of the substitute speaker classifier is evaluated by comparing its results to both the ground truth speaker label and the predictions made by the black-box speaker classifier, shown in Table 1. We can observe that 91.44% of the predictions of substitute and black-box classifiers are the same, while compared with the ground truth speaker label, the accuracy of the substitute classifier is 90.16%. This indicates the substitute SID is proximate to the black-box SID.

Table 1: *Accuracy (%) of substitute speaker classifier prediction comparing with ground truth speaker label and black-box speaker classifier prediction.*

	Black-box SID	Ground Truth
Black-box SID	-	97.42
Substitute SID	91.44	90.16

Table 2: *Attack success rates (%) of different kinds of generation methods.*

	Method	Acc (%) ↑
Baseline	VC	29.60
Upper limit	VC+adv [19]	76.50
Proposed	White-box	**60.58**
	Black-box ($\mathcal{L}_{\text{str}}$)	**52.86**
	Black-box ($\mathcal{L}_{\text{total}}$)	**55.38**

4.2. Attack Success Rate

The attack success rate, denoted as 'Acc' in Tabel 2, is used to measure the performance of targeted attacks in speaker identification. It represents the accuracy predicted from the SID and is calculated by dividing the number of successfully attacked audios by the total number of generated fake audios. A higher Acc indicates a better attack.

The first method used in the comparison is generated only by the vanilla VC model, which serves as the baseline for all other methods and is denoted as 'VC' in Table 2. The second method involves the direct addition of adversarial perturbation to the fake audios generated by the VC model, using the approach proposed in [19]. As the perturbation is optimized by the SID system and directly added to the waveform, it is still perceptible to human beings. This method is sub-optimal in the scenarios of deceiving both humans and machines, though represents the upper limit of adversarial attack and is denoted as 'VC+adv' in Table 2. The remaining three proposed methods are based on the VC model trained with multiple types of adversarial constraints, as outlined in Section 2, are denoted as 'White-box', 'Black-box ($\mathcal{L}_{\text{str}}$)', and 'Black-box ($\mathcal{L}_{\text{total}}$)', respectively. As shown in Table 2, the attack success rate of vanilla VC model based fake audio is 29.60%, while the VC audio with direct adversarial perturbation [19] achieves 76.50%. Meanwhile, the Acc results of our proposed timbre-reserved strategies are 60.58%, 52.86%, and 55.38%, respectively. We can observe that the fake audios generated based on all these three proposed timbre-reserved adversarial strategies are significantly improved compared to the fake audios generated by the original vanilla VC model, with improvements of 30.98%, 23.26%, and 25.78%, respectively. Furthermore, our proposed methods yield comparable results to those obtained through direct adversarial perturbations to the fake audios generated by the vanilla VC model.

4.3. Ablation Study

To further investigate the effectiveness of each component in the pseudo-Siamese framework, we conduct ablation studies of different components in the loss function. As shown in Table 3, the accuracy of the substitute speaker classifier prediction is evaluated by comparing it to the ground truth speaker label and the black-box speaker classifier prediction. In the ablation

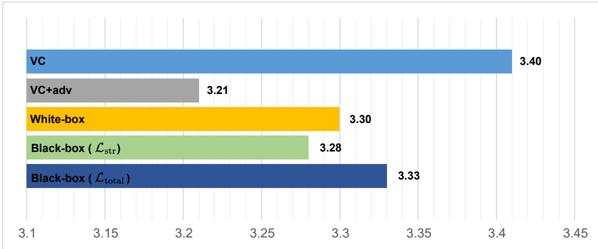

Figure 4: *The o-MOS of different kinds of generation methods.*

studies, we first remove the intrinsic similarity loss "$\mathcal{L}_{\text{int}}$", only remaining the structural similarity "$\mathcal{L}_{\text{str}}$", the performance degrades to 82.33% and 85.11%. Then, when continue removing the auxiliary similarity loss "$\mathcal{L}_{\text{aux}}$", the performance degrades to 67.19% and 66.22%. Both removed losses are calculated between the posterior probability of transformed speech and of substitute and black-box speaker classifiers, which proves the effectiveness of our proposed loss function and data augmentation method.

Table 3: *Ablation studies of different components in the loss function.*

Variants	Ground Truth	Black-box SID
Proposed ($\mathcal{L}_{\text{total}}$)	90.16%	91.44%
− Intrinsic $\mathcal{L}_{\text{int}}$	82.33%	85.11%
−Auxiliary $\mathcal{L}_{\text{aux}}$	67.19%	66.22%

4.4. Generated Audio Quality Evaluation

We employ MOSNet [34] prediction of objective Mean Opinion Score (o-MOS) to measure the quality of fake audio generated by different methods, as depicted in Figure 4. Notably, the o-MOS results of fake audios generated through our proposed strategies are consistently superior to those generated audios with direct adversarial perturbation addition. This indicates higher quality of fake audio generated through our proposed strategies, as we integrate adversarial constraints into the training process of the VC model, thereby avoiding the direct introduction of additional human-perceptible perturbations to the fake audio.

5. Conclusion

In this study, we propose a timbre-reserved adversarial attack on speaker identification in a black-box attack scenario. The proposed method involves using a timbre-reserved adversarial constraint during the training of the VC model to generate fake audio, along with a pseudo-Siamese network architecture to learn from the black-box SID model. The intrinsic similarity loss ensures intrinsic invariance, while the structural similarity loss ensures a similar decision boundary between the substitute speaker classifier and the fixed black-box speaker classifier. The substitute speaker classifier is used as a proxy to generate timbre-reserved fake audio for the black-box classifier. The experiments on the ADD challenge corpus show that our approach significantly improves the attack success rate compared to the vanilla VC model, without affecting the quality of the audio generated by the VC model or introducing extra noise. The proposed method offers a new perspective on adversarial attacks in speaker identification, emphasizing the importance of preserving the timbre of the target speaker during the attack.

6. References

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[3] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. ISCA*, 2012.

[4] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[5] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, N. Evans, M. Todisco, and H. Delgado, "ASVspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[6] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

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

[8] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. ICASSP*. IEEE, 2018, pp. 1962–1966.

[9] Q. Wang, P. Guo, S. Sun, L. Xie, and J. H. Hansen, "Adversarial regularization for end-to-end robust speaker verification," in *Proc. INTERSPEECH*, 2019, pp. 4010–4014.

[10] H. Abdullah, W. Garcia, C. Peeters, P. Traynor, K. R. Butler, and J. Wilson, "Practical hidden voice attacks against speech and speaker recognition systems," *arXiv preprint arXiv:1904.05734*, 2019.

[11] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 4889–4893.

[12] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Universal adversarial perturbations generative network for speaker recognition," in *Proc. ICME*. IEEE, 2020, pp. 1–6.

[13] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *Proc. ICASSP*. IEEE, 2020, pp. 1738–1742.

[14] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on GMM i-vector based speaker verification systems," in *Proc. ICASSP*. IEEE, 2020, pp. 6579–6583.

[15] Q. Wang, W. Rao, P. Guo, and L. Xie, "Adversarial training for multi-domain speaker recognition," in *Proc. ISCSLP*. IEEE, 2021, pp. 1–5.

[16] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Computer Speech & Language*, vol. 68, p. 101199, 2021.

[17] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," *Proc. INTERSPEECH*, pp. 4213–4217, 2020.

[18] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proc. ACM HotMobile*, 2020, pp. 9–14.

[19] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," *Proc. INTERSPEECH*, pp. 4228–4232, 2020.

[20] X. Zhang, X. Zhang, M. Sun, X. Zou, K. Chen, and N. Yu, "Imperceptible black-box waveform-level adversarial attack towards automatic speaker recognition," *Complex & Intelligent Systems*, pp. 1–15, 2022.

[21] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real bob? Adversarial attacks on speaker recognition systems," in *Proc. SP*. IEEE, 2021, pp. 694–711.

[22] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin TTS corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[23] J. Yi, R. Fu, J. Tao, S. Nie, H. Ma, C. Wang, T. Wang, Z. Tian, Y. Bai, C. Fan *et al.*, "ADD 2022: The first audio deep synthesis detection challenge," in *Proc. ICASSP*. IEEE, 2022, pp. 9216–9220.

[24] P. Moreno, P. Ho, and N. Vasconcelos, "A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications," *Advances in neural information processing systems*, vol. 16, 2003.

[25] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc ICLR*, 2021.

[26] Y. Liu, R. Xue, L. He, X. Tan, and S. Zhao, "DelightfulTTS 2: End-to-end speech synthesis with adversarial vector-quantized auto-encoders," in *Proc. INTERSPEECH*, 2022, pp. 1581–1585.

[27] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[28] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*. IEEE, 2017, pp. 1–5.

[29] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. INTERSPEECH*, 2021, pp. 4054–4058.

[30] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[31] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented Transformer for speech recognition," in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[33] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[34] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. INTERSPEECH*, 2019, pp. 1541–1545.