DAoB: A Transferable Adversarial Attack via Boundary Information for Speaker Recognition Systems

Junjian Zhang[1,3], Binyue Deng[1], Hao Tan[2,3], Le Wang[1] and Zhaoquan Gu[2,3,*]

[1]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China
[2]School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[3]Department of New Networks, Peng Cheng Laboratory, Shenzhen, China

Abstract

Audio deepfakes pose significant security threats to speaker recognition systems (SRSs), particularly with the growing threat of adversarial attacks. Existing black-box attack methods mostly rely on ensembling multiple datasets and models to search for adversarial examples (AEs) with good transferability, but they ignore the limitations of such search algorithms. In this paper, we comprehensively analyze different iterative-based adversarial attack methods and explain different transferability from the perspective of optimizing the search space. Furthermore, we propose a diffusion-based attack method located on the boundary (DAoB for short), which takes boundary information into consideration to achieve better transferability. Specifically, DAoB starts searching for an appropriate AE from the boundary of the search space instead of the original example, then it guides the search process by diffused audio and the gradients of multiple white-box models to obtain better gradient directions. To validate the effectiveness, we conducted experiments on seven state-of-the-art SRSs and DAoB outperforms others. Remarkably, even in the black-box scenario, the attack success rate of DAoB attains an impressive 97.2%, in close proximity to the rate achieved in the white-box scenario.

Keywords

Speaker recognition, adversarial attack, boundary information, diffusion, transferability

1. Introduction

With the rapid development of information technologies, identity recognition have become more intelligent and convenient. Typical methods such as facial recognition, fingerprint recognition, iris recognition, and speaker recognition can quickly perform identity verification with high success ratio[1, 2]. Especially with the emergence of deep learning technologies, identity recognition has made significant progress. However, Deepfake technology verifies the challenge of identity recognition systems' reliability [3, 4]. Compared to image-based deepfake technology, speech-based deepfake technology is a relatively new field. For example, recently emerging technologies [5, 6] such as speech recording and replay, Text-to-Speech (TTS), and Voice Conversion (VC) can deceive SRSs to a certain extent.

Various deep-learning models are shown to be vulnerable to adversarial attacks which add small amounts of noise to the benign example and mislead high-performance deep neural networks (DNNs) to produce incorrect output [7]. This attack can create examples that deceive deep learning models without being easily detected by humans. This reveals the common and serious vulnerabilities of deep learning models and promotes the further development of various intelligent technologies. Among them, adversarial attacks against SRSs are more difficult than those against images and have started later [2].

Commonly speaking, there are three attack scenarios: white-box, gray-box, and black-box. In the white-box scenario, attackers have access to all the details of the SRSs, so they often employ gradient-based attack methods to directly generate adversarial examples. Recent studies [8, 9, 10] show that adversarial attacks can fully overcome almost all white-box SRSs. In the gray-box scenario, attackers need to continuously query the victim system and obtain corresponding score vectors to optimize adversarial examples. For example, the FAKEBOB method proposed in [11] can effectively attack most SRSs in the gray-box situation. However, this strategy requires frequent access to the victim system, which may expose the attacking intent and weaken the attack's concealment. In real scenarios, the internal information of the model is often unknown, and a normal SRS only outputs a speaker identity label, not a score vector. Therefore, the black-box attack scenario has attracted more attention from researchers. Currently, effective attacks against black-box scenarios are mostly based on the transferability of adversarial examples, by generating adversarial examples on existing white-box models or retraining

IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA 2023), August 19, 2023, Macao, S.A.R
*Corresponding author.
2112106069@e.gzhu.edu.cn (J. Zhang); 2112106010@e.gzhu.edu.cn (B. Deng); tanhh198@gmail.com (H. Tan); wangle@gzhu.edu.cn (L. Wang); guzhaoquan@hit.edu.cn (Z. Gu)
0009-0002-1180-0786 (J. Zhang)

substitute models to generate the adversarial examples against the target black-box model. Therefore, improving the transferability of adversarial examples has become a unanimous goal of researchers.

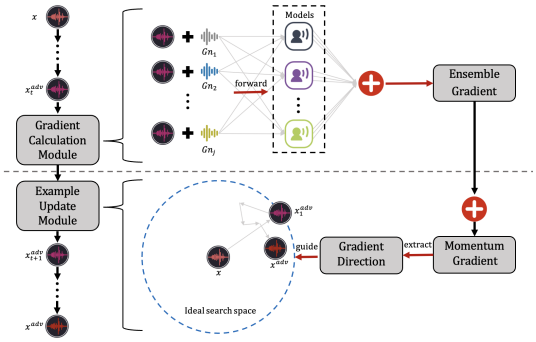

Figure 1: DAoB obtains gradient information from diffused examples and multiple models and uses this information to guide the search for adversarial examples starting from the boundary of the search space.

Many works on enhancing transferability [12, 13, 14] have achieved positive results, but there are still some unresolved or undiscovered issues. First, the development of transferability for adversarial examples is still insufficient, and the effectiveness of attacks is significantly different from that of white-box attacks. Second, with the limitation of the algorithmic constraints, searching and generating adversarial examples always in a smaller real search space than expected. Third, current works have limited research on model ensembles, mostly integrating only 2-3 models, and lack research on multi-model ensembles.

In this paper, we propose a diffusion attack located on the boundary (called DAoB for short), which adopts the boundary information to address the above-mentioned problems. The workflow of DAoB is described in Fig. 1. Specifically, we place the search region for adversarial examples directly on the attack boundary, which is more likely to produce high transferability at the beginning of each iteration. Secondly, when calculating gradient information, we obtain multiple gradient information that has been diffused by taking an information diffusion space near the example and taking the average of these gradients as guidance for this iteration. Finally, considering that the logits of different models have large differences, we adopt gradient-level model integration for the attack.

We summarize the paper's main contributions as follows:

- We conduct theoretical analysis and experimental comparisons of existing adversarial attack methods, propose definitions of ideal and actual search spaces, and explain the differences in transferability of different attack methods through the differences in search spaces;
- We propose DAoB, a powerful technique for generating adversarial examples, which can generate adversarial examples that pose a strong threat to black-box SRSs;
- We demonstrate the effectiveness of our method by attacking seven state-of-the-art SRSs. Our method can achieve attack effects that are close to the white-box scenario without accessing the target model in a fully black-box setting.

The rest of the paper is organized as follows. The next section highlights the related work in the field of adversarial attacks against speaker recognition systems. Then, we demonstrate DAoB and its theoretical analysis in Section 3. We describe the setup of the experiment in Section 4 and show the experiment results and discussions in Section 5. Finally, we conclude the paper in Section 6.

2. Related work

2.1. FakeBob

FakeBob [11] first proposed a black-box adversarial attack on SRSs. By assuming that the attacker can obtain the scoring of the model, and designing a loss function based on the score vector, the attack success rate can be close to that of a white-box attack. However, this method still requires model scores to attack in actual attacks, and it is not a complete black box. At the same time, a large number of query target models are required during the attack process, which is difficult to achieve in practical applications.

2.2. Transferable Adversarial Attack

Zhang et al. [13] proposed an integrated attack at the logits level can be used to achieve a black-box attack with a higher success rate, but they did not give a reasonable solution to the problem of inconsistent logits ranges between different models in the paper. The method in [12] uses spatial momentum to calculate the gradient for integrated attack, that is, the gradient information in the previous iteration process will be used in this iteration, and the use of momentum can effectively improve the aggressiveness and transferability of adversarial examples. The STA-MDCT [14] method introduces discrete cosine transform into the adversarial attack of speaker recognition, it firstly converts the audio to the frequency domain and adds random adversarial noise, and then converts it back to the time domain for model integration attack. This method improves the interpretability

and transferability of the attack in the black-box environment, but the additional transformation increases a lot of computing time, and the current models are all end-to-end input, they would not pay more attention to the frequency domain part of the audio.

3. METHODOLOGY

3.1. DAoB

As shown in the Fig. 1, we divide each iteration of DAoB into two phases: a module for computing the gradient of the current iteration, and another for using the gradient to guide the search for the adversarial example in this iteration. Specifically, in the gradient computation module, we input the diffused example into multiple white-box models to obtain their corresponding embeddings and calculate the gradient accordingly. In the example update module, we add the gradient obtained in this iteration to the momentum gradient. Then, we extract the direction of the momentum gradient to guide the update of the current example. Specifically, we update the original example starting from the search boundary with a small step size for multiple iterations, eventually obtaining an adversarial example with strong transferability, which can cause a black-box victim model to output incorrect results. Below, we will detail the theoretical basis and specific operations of DAoB.

Algorithm 1 DAoB

Require:
Set of several white-box models $M = \{M_j | j = 1, \cdots, m\}$, clean input x with target labels y, parameters $\theta = \{\epsilon, T, n, \beta, \gamma\}$

Ensure:
Adversarial example x^{adv};

$x_1^{adv} \leftarrow x, g_0 \leftarrow 0, \alpha_0 \leftarrow \epsilon$;
for iteration time $t \leftarrow 1$ to T **do**
  for number of sub-example $i \leftarrow 1$ to n **do**
    $x_i^d = D(x_t^{adv}, \gamma)$
    for number of model $j \leftarrow 1$ to m **do**
      Calculate gradient g_i^j of (x_i^d, M_j)
    end for
  end for
  Update g_t by Eq. (3-4)
  Update x_t^{adv} by Eq. (5-7)
end for
return x^{adv}

3.1.1. Take the boundary as the starting point

Previous attack methods generate adversarial examples by updating from the original example. Attackers set a boundary epsilon to prevent excessive noise. For multi-step attacks (using the most commonly used parameter selection, step size $\alpha = \epsilon/n$), as the gradient direction changes, the attacker is unable to search near the set boundaries.

For transfer-based black-box attacks, many experiments have shown that the single-step attack method has certain advantages in transferability. It can also be inferred from experience that the further an adversarial example is from the original example, i.e., the closer it is to the boundary, the greater the difference in speaker characteristics between the adversarial example and the original example. Therefore, we believe that the transferability of adversarial examples in the search space near the boundary is stronger than that of adversarial examples in the search space near the original example.

In summary, we propose an attack method that starts from the boundary. Based on the initial gradient information, we directly add a large amount of noise to the original example to reach the boundary. Then, we search for adversarial examples from the boundary with a smaller step size.

3.1.2. Diffusion attack

The diffusion model has played a huge role in the fields of image and speech generation. The process of a single diffusion step is to add noise to the example x_k to obtain x_{k+1}, while the reverse diffusion is the process of using x_k and x_{k+1} as guidance to learn the generation method. Inspired by the diffusion model, we optimize the process of generating adversarial examples iteratively obtained from x_i, by diffuse it to x_i^d, and then using x_i^d as guidance to iterate to obtain x_{i+1}. Considering that the process of generating adversarial examples lacks support from data volume, we perform multiple diffusions to better guide the direction of iteration.

3.1.3. DAoB with model ensembles

To improve transferability, we obtain gradients from multiple white-box models. Unlike ensemble methods based on logits level, we set the same loss function for each white-box model. We believe that the size of the gradient in any dimension reflects the influence that the gradient direction can have in that dimension. Therefore, we do not set weights for the gradients of different white-box models in the ensemble gradient. We simply add the gradients obtained from each white-box model and then take the direction of the ensemble gradient as the direction for the current iteration.

We introduced the details of DAoB, which is described in Algorithm 1. We formalize the specific attack process as follows:

The input consists of a set of several white-box models $M = \{M_j | j = 1, \cdots, m\}$,clean input x with target labels y, and a series of required parameters. In each iteration, the audio x_t^{adv} from the previous iteration is diffused into a x_i^d:

$$x_i^d = D(x_t^{adv}, \gamma) = x_t^{adv} + Gn(x_t^{adv}, \gamma), \quad (1)$$

where $Gn(x_t^{adv}, \gamma)$ is Gaussian noise with the same shape as x_t^{adv} and follows a normal distribution $N(0, \gamma^2)$.

The gradients computed using x_i^d and M in each iteration can be formalized as:

$$g_t^w = \sum_{i=1,\cdots,n}^{j=1,\cdots,m} \nabla_x J_j \left(x_i^d, y\right), \quad (2)$$

where $J_j\left(x_i^d, y\right)$ is the derivative of the loss function with respect to the input example x_i^d.

Then we obtain the gradient for this iteration as:

$$g_t = \mu \cdot g_{t-1} + \frac{g_t^w}{\|g_t^w\|_1}. \quad (3)$$

We use the first obtained gradient g_1 to guide the original example to reach the boundary of the search space:

$$x_1^{adv} = x + \epsilon \cdot \operatorname{sign}(g_1), \quad (4)$$

where ϵ is the constraint of pertubation.

Afterwards, we use the momentum gradient to generate smaller perturbation, which is then used to create the adversarial audio. This process can be formalized as follows:

$$\alpha_t = \beta \cdot \alpha_{t-1}, \quad (5)$$

$$x_t^{adv} = \operatorname{Clip}_x^{\epsilon}\left\{x_{t-1}^{adv} + \alpha_t \cdot \operatorname{sign}(g_{t+1})\right\}, \quad (6)$$

where α_t is the attack step size, and β is the attenuation factor.

4. Experiment set up

4.1. Datasets and Models

The datasets generated according to libriSpeech [15] are consistent with the Spk10-enroll, Spk10-test, and Spk10-imposter datasets published in [16].To validate the effectiveness of our method, we selected seven strong victim models: Res34-L [17], Res34-V [17], TDy_HR [18], TDy_QR [18], TDy_VGG [18], XV-plda [19], ECAPA [20].

4.2. Evaluation Metrics

We use the attack success rate (ASR) of adversarial audio on black-box victim models as the metric to evaluate the transferability of adversarial audio.

In addition to focusing on the transferability of adversarial audio, we also pay attention to the imperceptibility of adversarial audio. Therefore, we refer to [12] and use Signal-to-Noise Ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ), and $L_2 - norm$ as the indicators of the imperceptibility of adversarial audio.

4.3. Parameter Details

During the process of adversarial attacks, all attacks were targeted attacks in the open-set recognition scenario, and the attack targets were fixed simple targets, with the specific settings being the same as in [21]. The number of iterations T for all iterative attack methods was 10. We conducted experiments with different perturbation limits, but due to space constraints, the perturbation limits ϵ shown in the experimental results below are all set to 0.002. The fixed step size α is set to ϵ/T. In particular, there is some randomness in diffusion of DAoB, so all results reported for our methods are the average of 10 repeated experiments.

5. Experiment Results and Analysis

5.1. Results of Different Attack Methods

We systematically studied the transferability of adversarial examples generated by DAoB and reported the experimental results in Tab. 1, where the DAoB's parameter setting is $\{n = 3, \beta = 0.8\}$. Specifically, all attack methods adopted the strategy of the gradient-level model ensemble. The gray part in the table is the attack success rate of adversarial examples on the white-box model. We can see that for any victim model, the attack success rate of DAoB is higher than that of MI-FGSM, with the highest increase of 16.9 percentage points. Compared with the PGD and MI-FGSM, the average attack success rate from six group black-box attack experiments has improved by 43.65% and 23.04%, respectively. At the same time, the audio quality of DAoB is between that of MI-FGSM and FGSM, which is consistent with our description of the search space in the previous text: DAoB can search the area near the boundary, while MI-FGSM cannot. Nevertheless, the adversarial audio obtained by DAoB still achieves a high signal-to-noise ratio and PESQ. In addition, we did not include xv-plda as a white-box model in the table because we found during the experiment that xv-plda would have a negative impact on the attack. We

Table 1
The target attack success rate (ASR%) results of different attack methods against seven victim speaker recognition models.

Methods	ASR(%)							Stealthiness		
	Res34-V	Res34-L	TDy-HR	TDy-QR	TDy-VGG	ECAPA	xv-plda	L2↓	SNR↑	PESQ↑
FGSM	19.7	47.2	38.3	37.4	20.8	28.6	8.9	0.716	28.054	2.383
PGD	61.9	99.8	98.9	98.7	95.3	99.1	28.3	**0.273**	**36.511**	**3.668**
MI-FGSM	73.6	99.8	99.2	99.1	95.2	99.3	34.4	0.528	30.718	2.825
DAoB	**87.8**	**100**	**100**	**100**	**100**	**99.9**	**48.7**	0.63	29.171	2.715
FGSM	32	26.6	40	37.8	20.5	29.5	7.5	0.716	28.054	2.383
PGD	99.5	82.5	99.8	99.3	95.4	99.3	26.9	**0.277**	**36.414**	**3.669**
MI-FGSM	99.7	90.2	99.7	99.2	95.3	99.3	33.8	0.526	30.75	2.837
DAoB	**100**	**96.4**	**99.9**	**100**	**100**	**100**	**51.5**	0.627	29.212	2.729
FGSM	32.1	24.6	22.2	37.3	22.9	30.5	7.8	0.717	28.051	2.378
PGD	99.5	77.2	86.8	99.2	95.4	99.3	25.4	**0.275**	**36.459**	**3.668**
MI-FGSM	99.9	86.4	92.6	99.2	95.3	99.3	31.3	0.525	30.751	2.824
DAoB	**100**	**95.3**	**95.1**	**99.9**	**100**	**100**	**46.7**	0.627	29.209	2.713
FGSM	34.1	46	39	23.7	21.1	30.5	8.1	0.716	28.054	2.381
PGD	99.4	99.6	99.7	88.5	95.4	99.3	28.2	**0.278**	**36.4**	**3.664**
MI-FGSM	99.8	99.6	99.7	93.4	95.3	99.3	35.4	0.527	30.738	2.83
DAoB	**100**	**100**	**100**	**97.2**	**100**	**100**	**53.7**	0.625	29.222	2.728
FGSM	36.8	48.2	44	40.7	8.4	33.6	7.8	0.716	28.054	2.395
PGD	99.7	99.6	99.8	99.2	57	99.3	31	**0.283**	**36.255**	**3.66**
MI-FGSM	99.8	99.6	99.7	99.2	67.8	99.3	37.3	0.527	30.728	2.852
DAoB	**100**	**100**	**100**	**100**	**78**	**100**	**56**	0.631	29.153	2.723
FGSM	32.3	43.6	39.6	37.2	21	16.2	6.9	0.716	28.054	2.384
PGD	99.3	99.3	99.7	99.2	95.2	41.8	21.3	**0.269**	**36.681**	**3.676**
MI-FGSM	99.7	99.6	99.7	99.2	95.2	58.7	28.2	0.527	30.731	2.831
DAoB	**100**	**99.9**	**100**	**99.9**	**100**	**76.6**	**44.9**	0.625	29.238	2.734

have not found a specific reason for this phenomenon and we will try to explain it in future work.

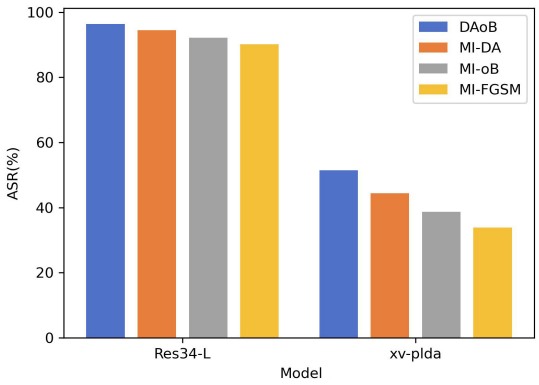

Figure 2: The experimental results of the ablation study.

5.2. Ablation Study

We conducted ablation studies by combining MI-FGSM with diffusion attack and boundary attack, and Fig. 2 reports some experimental results. We use Res34-L, Res34-V, TDy_HR, TDy_VGG, and ECAPA as white-box models to generate adversarial examples. The height of the bars in the chart represents the success rate of adversarial examples on the corresponding black-box model. We found that both diffusion attacks and boundary attacks can improve the transferability of attacks, and there is no conflict between these two strategies.

6. Conclusion

We study effective adversarial attacks for black-box SRSs. By adopting the boundary information, we propose DAoB, a new adversarial example generation strategy that effectively improves the transferability of adversarial examples. Experimental results show that DAoB can achieve a high success rate under the black-box scenario that is close to that of white-box attacks. In future work, we will try to reduce the size of adversarial perturbations and limit the area of adversarial perturbations to enhance the stealth of attacks without affecting the success rate of attacks.

Acknowledgement

This work is supported in part by the Major Key Project of PCL (Grant No. PCL2022A03) and the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

[1] J. Priesnitz, C. Rathgeb, N. Buchmann, C. Busch, M. Margraf, An overview of touchless 2d fingerprint recognition, EURASIP Journal on Image and Video Processing 2021 (2021) 1–28.

[2] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, Z. Gu, Adversarial attack and defense strategies of speaker recognition systems: A survey, Electronics 11 (2022) 2183.

[3] Z. Gu, W. Hu, C. Zhang, H. Lu, L. Yin, L. Wang, Gradient shielding: towards understanding vulnerability of deep neural networks, IEEE transactions on network science and engineering 8 (2020) 921–932.

[4] Z. Gu, H. Li, S. Khan, L. Deng, X. Du, M. Guizani, Z. Tian, Iepsbp: A cost-efficient image encryption algorithm based on parallel chaotic system for green iot, IEEE Transactions on Green Communications and Networking 6 (2021) 89–106.

[5] L. Li, Y. Chen, D. Wang, T. F. Zheng, A study on replay attack and anti-spoofing for automatic speaker verification, arXiv preprint arXiv:1706.02101 (2017).

[6] Y. Zhang, F. Jiang, Z. Duan, One-class learning towards synthetic voice spoofing detection, IEEE Signal Processing Letters 28 (2021) 937–941.

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

[8] Z. Li, Y. Wu, J. Liu, Y. Chen, B. Yuan, Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations, in: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020, pp. 1121–1134.

[9] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, X. Hu, Attack on practical speaker verification system using universal adversarial perturbations, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 2575–2579.

[10] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, B. Yuan, Real-time, universal, and robust adversarial attacks against speaker recognition systems, in: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2020, pp. 1738–1742.

[11] G. Chen, S. Chenb, L. Fan, X. Du, Z. Zhao, F. Song, Y. Liu, Who is real bob? adversarial attacks on speaker recognition systems, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 694–711.

[12] H. Tan, Z. Gu, L. Wang, H. Zhang, B. B. Gupta, Z. Tian, Improving adversarial transferability by temporal and spatial momentum in urban speaker recognition systems, Computers and Electrical Engineering 104 (2022) 108446.

[13] Y. Zhang, Z. Jiang, J. Villalba, N. Dehak, Black-box attacks on spoofing countermeasures using transferability of adversarial examples., in: Interspeech, 2020, pp. 4238–4242.

[14] J. Yao, H. Luo, X.-L. Zhang, Interpretable spectrum transformation attacks to speaker recognition, arXiv preprint arXiv:2302.10686 (2023).

[15] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2015, pp. 5206–5210.

[16] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, Y. Liu, Sec4sr: a security analysis platform for speaker recognition, arXiv preprint arXiv:2109.01766 (2021).

[17] J. S. Chung, J. Huh, S. Mun, M. Lee, H. S. Heo, S. Choe, C. Ham, S. Jung, B.-J. Lee, I. Han, In defence of metric learning for speaker recognition, arXiv preprint arXiv:2003.11982 (2020).

[18] S.-H. Kim, H. Nam, Y.-H. Park, Temporal dynamic convolutional neural network for text-independent speaker verification and phonemic analysis, in: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2022, pp. 6742–6746.

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

[20] B. Desplanques, J. Thienpondt, K. Demuynck, Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, arXiv preprint arXiv:2005.07143 (2020).

[21] J. Zhang, H. Tan, B. Deng, J. Hu, D. Zhu, L. Huang, Z. Gu, Nmi-fgsm-tri: An efficient and targeted method for generating adversarial examples for speaker recognition, in: 2022 7th IEEE International Conference on Data Science in Cyberspace (DSC), IEEE, 2022, pp. 167–174.