Symmetric Saliency-Based Adversarial Attack to Speaker Identification

Jiadi Yao, Xing Chen, Xiao-Lei Zhang, *Senior Member, IEEE*, Wei-Qiang Zhang, *Senior Member, IEEE*, and Kunde Yang

***Abstract*—Adversarial attack approaches to speaker identification either need high computational cost or are not very effective, to our knowledge. To address this issue, in this letter, we propose a novel generation-network-based approach, called symmetric saliency-based encoder-decoder (SSED), to generate adversarial voice examples to speaker identification. It contains two novel components. First, it uses a novel saliency map decoder to learn the importance of speech samples to the decision of a targeted speaker identification system, so as to make the attacker focus on generating artificial noise to the important samples. It also proposes an angular loss function to push the speaker embedding far away from the source speaker. Our experimental results demonstrate that the proposed SSED yields the state-of-the-art performance, i.e. over 97% targeted attack success rate and a signal-to-noise level of over 39 dB on both the open-set and close-set speaker identification tasks, with a low computational cost.**

***Index Terms*—Adversarial attack, speaker identification, saliency map decoder, angular loss.**

I. Introduction

SPEAKER recognition is vulnerable to spoofing attacks [1]. Many spoofing attack techniques to speaker recognition, including replay, voice conversion, impersonation and text-to-speech synthesis, and adversarial attacks [2], have been developed. On the contrary, various detection [3], [4], [5] and countermeasures [6] against spoofing attacks are in full swing. In this letter, we focus on developing adversarial attacks to speaker identification. An adversarial attack to speaker identification aims to make an identification system wrongly recognize the adversarial voice of a source speaker as a targeted imposter speaker, where the adversarial voice, a.k.a. *adversarial example*, is produced by adding human-imperceptible noise to the speech of the source speaker. It shows great threat to modern speaker identification systems based on deep learning. Recently, many adversarial attacks, which were first studied in computer vision [7], [8], [9], [10], can be used to speaker recognition as well. Existing adversarial noise generation methods to speaker recognition can be categorized roughly into: i) gradient-based approaches [11], such as FGSM [12] and BIM [13], ii) optimization-based approaches, such as the C & W attack [14], Quasi-Newton [15], FoolHD [16] and AdvPulse [17], iii) query-based approaches [18], and iv) generation-network-based approaches, such as the Universal Adversarial Perturbations (UAPs) [19] and FAPG [20].

The first three classes of the aforementioned methods are able to generate adversarial examples effectively with the expense of high computational cost, since that they need to search the optimal perturbation for each test utterance. To address the issue, the generation-network-based methods were proposed, which are able to generate an adversarial example of a test utterance in a single forward inference pass. However, to our knowledge, their performance might not be as high as the other classes in terms of targeted attack successful rate (TASR) and signal-to-noise ratio (SNR).

To generate adversarial attacks to speaker identification efficiently that are able to achieve both high TASR and high SNR, in this letter, inspired by [21], we propose a generation-network-based method, named Symmetric Saliency-based Encoder-Decoder (SSED), for the targeted attack to a speaker identification system. It has two novel components:

1) A novel *saliency map decoder* makes the attacker focus on the important samples of a test utterance that will strongly affect the decision of the targeted speaker identification system.
2) A novel *angular loss* makes the speaker embedding of an adversarial example far apart from the source speaker, which is a supplement to the mainstream of making the adversarial example close to the targeted speaker.

Experimental results on both close-set speaker identification and open-set speaker identification [18], [22] show that the proposed method achieves comparable performance to representative gradient- and optimization-based approaches on TASR and SNR with much higher efficiency than the latter. It also significantly outperforms a representative generation-network-based method with similar efficiency.

II. Preliminaries

Speaker identification aims to detect the speaker identity of a test utterance from an enrollment database [23]. It contains an enrollment phase and a test phase. In the enrollment phase, the system enrolls a group of k speakers, with speaker identities $\mathcal{U} =$

Manuscript received 7 December 2022; accepted 4 January 2023. Date of publication 12 January 2023; date of current version 3 February 2023. This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62176211 and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China under Grants JCYJ20210324143006016 and JSGG20210802152546026. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiangui Kang. *(Corresponding author: Xiao-Lei Zhang.)*

Jiadi Yao, Xing Chen, and Xiao-Lei Zhang are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Research and Development Institute of Northwestern Polytechnical University, Shenzhen 710072, China (e-mail: yaojiadi@mail.nwpu.edu.cn; xing.chen@mail.nwpu.edu.cn; huoshan6@126.com).

Wei-Qiang Zhang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wqzhang@tsinghua.edu.cn).

Kunde Yang is with the Ocean Institute of Northwestern Polytechnical University, Xi'an 710072, China (e-mail: ykdzym@nwpu.edu.cn).

Digital Object Identifier 10.1109/LSP.2023.3236509

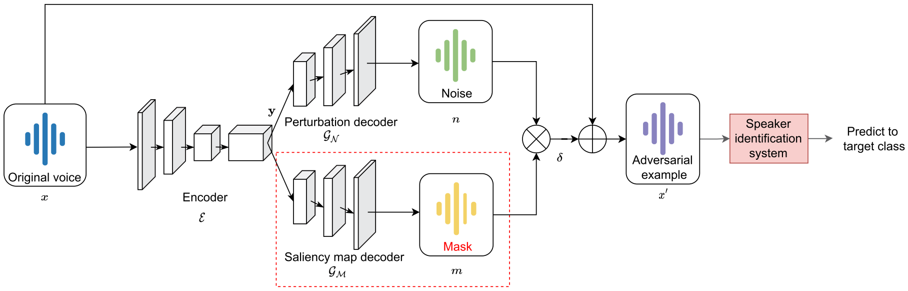


Fig. 1. Overview of the proposed symmetric saliency-based encoder-decoder. The saliency map decoder is in the red dot box.

$\{1, 2, \ldots, k\}$. In the test phase, the system aims to recognize the speaker identity of a test voice $\boldsymbol{x}$, which has the following two situations:

Close-Set speaker Identification (CSI): It classifies $\boldsymbol{x}$ into one of the enrolled speakers. The decision module of CSI $D(\boldsymbol{x})$ is defined as:

$$D(\boldsymbol{x}) = \underset{i \in \mathcal{U}}{\operatorname{argmax}} [S(\boldsymbol{x})]_i, \tag{1}$$

where $[S(\boldsymbol{x})]_i$ denotes the similarity score between $\boldsymbol{x}$ and the i-th enrollment speaker produced by the speaker identification.

Open-Set speaker Identification (OSI): It either identifies $\boldsymbol{x}$ as one of the enrolled speakers or determines that $\boldsymbol{x}$ does not belong to any of the speakers in $\mathcal{U}$. The decision module of OSI $D(\boldsymbol{x})$ is defined as:

$$D(\boldsymbol{x}) = \begin{cases} \underset{i \in \mathcal{U}}{\operatorname{argmax}} [S(\boldsymbol{x})]_i, & \text{if } \underset{i \in \mathcal{U}}{\max} [S(\boldsymbol{x})]_i \geq \theta; \\ \text{reject}, & \text{otherwise}. \end{cases} \tag{2}$$

where θ is a predefined decision threshold θ. If the largest score is less than θ, then the voice $\boldsymbol{x}$ is not uttered by any of the enrolled speakers, i.e., it is rejected by the system.

III. Proposed Method

A. Problem Formulation

A targeted adversarial attack to the speaker identification system $D(\boldsymbol{x})$ aims to generate a perturbation $\boldsymbol{\delta}$ to the input voice $\boldsymbol{x}$ of a source speaker s, such that the system may misclassify a generated adversarial example $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{\delta}$ as a target speaker $t \in \mathcal{U} = \{1, \ldots, k\}$, i.e. $t = D(\boldsymbol{x}')$, while the difference between $\boldsymbol{x}$ and $\boldsymbol{x}'$ is as small as possible and may be unaware by humans. Usually, the perturbation $\boldsymbol{\delta}$ is generated by an adversarial attacker $\boldsymbol{\delta} = G_\alpha(\boldsymbol{x})$, where α is the learnable parameters of the attacker. To design an effective attacker, the core is to design a loss function, denoted as L, such that, when L is minimized, $t = D(\boldsymbol{x}')$ is achieved.

This letter focuses on adversarial attack techniques in the time domain, where the attacker $G_\alpha(\cdot)$ takes the waveform voice $\boldsymbol{x}$ as its input and generates $\boldsymbol{\delta}$ in the time domain. In the following of this section, we describe the proposed attacker, i.e. SSED, in detail.

B. Framework

The proposed SSED aims to generate adversarial perturbations with high TASR and high SNR. The architecture of the proposed SSED is shown in Fig. 1. It consists of an encoder $\mathcal{E}$, a perturbation decoder $\mathcal{G}_{\mathcal{N}}$, and a saliency map decoder $\mathcal{G}_{\mathcal{M}}$, where the saliency map decoder is one of the core novelties of the proposed method.

SSED generates the perturbation $\boldsymbol{\delta}$ for the voice of the source speaker $\boldsymbol{x}$ in the following process. First, the encoder $\mathcal{E}$ encodes $\boldsymbol{x}$ into a latent vector $\mathbf{y}$. Then, the perturbation decoder $\mathcal{G}_{\mathcal{N}}$ creates noise $\boldsymbol{n}$ from $\mathbf{y}$, i.e. $\boldsymbol{n} = \mathcal{G}_{\mathcal{N}}(\mathbf{y})$. In the meantime, the saliency map decoder $\mathcal{G}_{\mathcal{M}}$ generates a mask $\boldsymbol{m}$ from $\mathbf{y}$. The final adversarial perturbation $\boldsymbol{\delta}$ is generated by:

$$\boldsymbol{\delta} = c(\boldsymbol{n} \odot \boldsymbol{m}), \tag{3}$$

where c is a manually-tunable scaling factor for controlling the amplitude of the perturbation, and the symbol "$\odot$" denotes the dot-product operator. Note that, the mask $\boldsymbol{m}$ is designed to emphasize the important samples of $\boldsymbol{x}$ that affect the decision $D(\boldsymbol{x})$. Also, $\boldsymbol{x}, \boldsymbol{n}$, and $\boldsymbol{m}$ are all voice signals in the time domain. The above modules are all convolutional residual networks. See Section IV-A3 for the detailed settings of the network structures.

SSED jointly trains the above modules by minimizing the following loss function:

$$L = L_{\text{SNR}} + L_{\text{TASR}}, \tag{4}$$

where L_{SNR} and L_{TASR} are two components for improving the SNR and TASR of the adversarial examples respectively. In the following two subsections, we describe the two components respectively.

C. Saliency Map Decoder

Unlike existing adversarial attack approaches for speaker identification that simply minimizes $L_{\text{norm}} = [\max(\boldsymbol{x} - \boldsymbol{x}', 0)]^2$ to maximize the SNR [19], here we propose an additional saliency map decoder for further improving the SNR. The keyword "saliency map," which was originally used in image processing, e.g. [24], is used to make SSED attend to highly sensitive regions of $\mathbf{y}$ that affect the decision of $D(\boldsymbol{x})$. The saliency map decoder minimizes the following objective:

$$L_f = \sqrt{(\boldsymbol{m}')^T \boldsymbol{m}'}, \tag{5}$$

where $\boldsymbol{m}'$ is a normalized saliency map vector whose elements are variables in the range of [0,1]: $\boldsymbol{m}' = \frac{\boldsymbol{m} - \min(\boldsymbol{m})}{\max(\boldsymbol{m}) - \min(\boldsymbol{m})}$.

Making SSED focus on the highly sensitive regions of the spectral representation $\mathbf{y}$ has the following two merits. First, SSED preserves the spectral characteristics of the original voice $\boldsymbol{x}$, and makes $\boldsymbol{\delta}$ difficult to be detected in the frequency domain. At the same time, the distribution of the perturbation is concentrated to the high energy region of $\boldsymbol{x}$, which makes the adversarial voice $\boldsymbol{x}'$ less noisy and more inconspicuous.

Finally, the SNR of the perturbation is improved by minimizing:

$$L_{\text{SNR}} = \lambda_f L_f + \lambda_n L_{\text{norm}}, \tag{6}$$

where λ_f and λ_n are two hyperparameters.

D. Angular Loss Function

To improve the TASR of adversarial examples to speaker identification, we need to make the speaker embedding of $\boldsymbol{x}'$ as dissimilar as possible to that of the source speaker s, and as similar as possible to the target speaker t, where a speaker embedding of an utterance is the input of the softmax layer of the speaker identification system. However, existing approaches only explored the latter [25], leaving the former far from studied, to our knowledge. In this letter, we propose a new loss function, named angular loss, to enlarge the dissimilarity between the speaker embeddings of $\boldsymbol{x}'$ and the source speaker s.

Specifically, we denote the speaker embedding of $\boldsymbol{x}'$ as $\mathbf{z}'$, and the speaker embedding of $\boldsymbol{x}$ as $\mathbf{z}$. The angular loss minimizes:

$$L_{\text{angular}} = \frac{\mathbf{z}^T \mathbf{z}'}{\max\left(\|\mathbf{z}\|_2 \left\|\mathbf{z}'\right\|_2, \epsilon\right)}, \tag{7}$$

where ϵ is a small constant, which is set to 10^{-12} in this letter.

We also adopt the conventional speaker loss [18], [22] as part of L_{TASR}. It has different forms for CSI and OSI respectively. For the integrity of the letter, we present them briefly as follows. The speaker loss for CSI is defined as:

$$L_{\text{speaker}} = \max_{i \in \mathcal{U}, i \neq t} [S(\boldsymbol{x}')]_i - [S(\boldsymbol{x}')]_t, \tag{8}$$

which aims to make the targeted speaker identification system wrongly identify $\boldsymbol{x}'$ as the target speaker t. The speaker loss for OSI also have to guarantee that the decision score $[S(\boldsymbol{x}')]_t$ is larger than the threshold θ; otherwise, x' will not be determined as any of the enrolled speakers:

$$L_{\text{speaker}} = \max\left\{\max_{i \in \mathcal{U}, i \neq t} [S(\boldsymbol{x}')]_i, \theta\right\} - [S(\boldsymbol{x}')]_t. \tag{9}$$

Finally, the TASR of SSED can be improved by minimizing:

$$L_{\text{TASR}} = \lambda_s L_{\text{speaker}} + \lambda_a L_{\text{angular}}, \tag{10}$$

where λ_s and λ_a are two hyperparameters. Substituting (6) and (10) into (4) derives the objective of the proposed SSED.

IV. Experiments

A. Experimental Setup

1) Dataset: We used VoxCeleb1 and VoxCeleb2 datasets [26], where the development set of VoxCeleb2 was used for the training of the speaker identification system. All comparison methods were evaluated in both the CSI and OSI scenarios. For CSI, the system was enrolled by 1211 speakers from the development set of VoxCeleb1. In the evaluation process, we randomly chose one of the enrolled speakers as the targeted speaker of adversarial attacks. We selected 35 utterances from each speaker of the development set of VoxCeleb1 to train adversarial attackers, and another 5 utterances of each speaker as imposters for testing.

For OSI, the system is enrolled by 5 random speakers from the development set of VoxCeleb2 [18]. In the evaluation process, we randomly chose one of the 5 enrolled speakers as the targeted speaker. We trained each adversarial attacker with 10 utterances per speaker from the 5,994 speakers of the development set of VoxCeleb2, and chose 40 speakers from the test set of VoxCeleb1 as imposters to attack the OSI system.

2) Targeted Speaker Identification System: The Fast ResNet-34 [27], [28] trained with the AAM-Softmax [29] objective was used as the targeted speaker identification system. We used voxceleb_trainer[1] to train the system. To verify that the targeted system is a state-of-the-art system, we first evaluated the system on the VoxCeleb1 development set which consists of 1,211 speakers. The system achieves a classification accuracy of 93.15%. We further transformed the speaker identification system to a verification system by using the cosine scoring back-end, and evaluated the verification system on the VoxCeleb1 Original trial list[2] that consists of 37 k trials from 40 speakers. Its equal error rate is 2.34%.

3) Comparison Methods: The parameter settings of the proposed SSED are as follows. The encoder $\mathcal{E}$ consists of 3 convolution blocks and 6 residual blocks. The 1-D convolution, batch normalization, and ReLU were applied in every convolution block. The kernel sizes for all convolution layers were set to 7, 3, and 3, respectively. The perturbation decoder $\mathcal{G}_{\mathcal{N}}$ consists of 3 transposed convolution blocks, whose kernel sizes are 3, 3, and 7, respectively. The saliency map decoder $\mathcal{G}_{\mathcal{M}}$ has a similar structure with $\mathcal{G}_{\mathcal{N}}$ except the last transposed convolution layer. We trained SSED by 10 epochs with the Adam optimizer and a learning rate of 10^{-3}.

We compared the proposed method with four well-known adversarial attack methods which are described as follows. FGSM [7] is a gradient-based one-step l_∞ attack. BIM-10 [8] is an iterative version of FGSM with 10 iterations. C & W [9] is an optimization-based attack. UAPs [19] is a universal generation-network-based attack approach. Note that, UAPs is similar to SSED. Besides the two novel contributions of SSED, SSED also adopts different generative network structures, given the recent fast development of deep architecture designs.

4) Evaluation Metrics: To evaluate the effectiveness of an attack, we adopt classification error rate (CER) and TASR. CER is the probability of misclassifying an adversarial example. TASR is the probability of identifying an imposter voice as the target speaker by the targeted speaker identification system. To measure the concealment of adversarial voices, we use SNR and Perceptual Evaluation of Speech Quality (PESQ). SNR is defined as $\text{SNR} = 10\log_{10}(P_x/P_\delta)$ where P_x and P_δ are the signal powers of $\boldsymbol{x}$ and $\boldsymbol{\delta}$ respectively. PESQ, which quantifies the voice quality by human perception, is in the range of $[-0.5, 4.5]$. Larger SNR and PESQ indicates better stealthiness. To measure the efficiency for generating the adversarial examples, we also recorded the running time.

B. Results

1) Main Results: Tables I and II list the comparison results of the proposed SSED with the four referenced methods on the CSI and OSI tasks respectively. From the tables, we see that the proposed SSED is comparable to BIM-10, and significantly outperforms the other comparison methods in terms of TASR and SNR. However, SSED needs much less time to generate the perturbations. Comparing to UAPs

[1] [Online]. Available: https://github.com/clovaai/voxceleb_trainer

[2] [Online]. Available: https://www.robots.ox.ac.uk/\;vgg/data/voxceleb/meta/veri_test.txt

TABLE I
COMPARISON RESULTS ON THE CSI TASK

Attack Type	c	CER (%)	TASR (%)	SNR (dB)	PESQ	Time (s)
None	-	6.85	0.09	-	-	-
FGSM	0.001	41.75	5.42	41.40	4.41	0.9
	0.002	49.63	10.03	35.42	4.31	
	0.005	59.88	14.91	27.47	4.05	
	0.01	67.86	16.09	21.45	3.73	
BIM-10	0.001	80.31	63.68	43.49	4.39	**5.86**
	0.002	95.04	90.23	38.48	4.30	
	0.005	**99.62**	**99.39**	**31.65**	**4.12**	
C&W-L2	-	65.54	52.88	63.97	4.45	130.21
UAPs [19]	-	96.66	65.6	38.56	2.48	0.004
SSED (proposed)	0.01	88.18	45.4	52.07	3.29	**0.41**
	0.03	96.65	90.0	43.19	2.49	
	0.05	**99.77**	**97.3**	**39.07**	**2.29**	

The performance of the top two methods are marked in bold.

TABLE II
COMPARISON RESULTS ON THE OSI TASK

Attack Type	c	TASR (%)	SNR (dB)	PESQ	Time (s)
None	-	0.84	-	-	-
FGSM	0.001	15.61	41.42	4.4	0.14
	0.002	21.5	35.40	4.3	
	0.005	21.34	27.44	4.03	
	0.01	16.45	21.42	3.69	
BIM-10	0.001	77.9	43.42	4.38	**1.16**
	0.002	95.9	38.42	4.28	
	0.005	**99.79**	**31.68**	**4.07**	
C&W-L2	-	93.66	64.02	4.46	9.25
UAPs [19]	-	56.0	33.66	2.48	0.004
SSED (proposed)	0.01	74.8	51.88	3.34	**0.41**
	0.03	92.8	43.1	2.63	
	0.05	**98.0**	**39.2**	**2.36**	

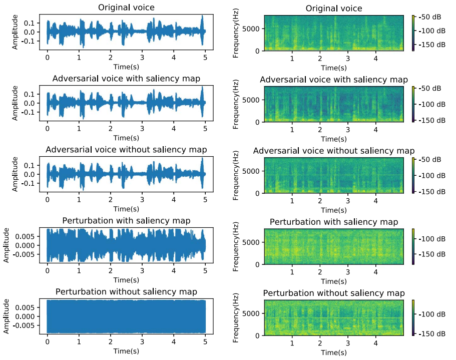
TABLE III
PERFORMANCE COMPARISON OF THE SSED WITH OR WITHOUT THE SALIENCY MAP DECODER

Task	Type	TASR (%)	SNR (dB)	PESQ
CSI	with saliency map decoder	**97.3**	**39.07**	**2.29**
	without saliency map decoder	95.1	35.62	2.25
OSI	with saliency map decoder	**98.0**	**39.2**	**2.36**
	without saliency map decoder	96.7	37.1	2.32

which is also a generation-network-based approach, SSED yields higher TASR and SNR than UAPs, with the expense of higher computational complexity. The phenomenon demonstrates the effectiveness of the novel points of SSED. Finally, SSED, which does not optimize for PESQ, behaves not well in PESQ.

2) Effect of the Saliency Map Decoder on Performance: We compared SSED with a variant of SSED that removes the saliency map decoder. From the comparison results in Table III, we see that the saliency map decoder leads to higher TASR, SNR and PESQ. Fig. 2 gives an example on how the saliency map decoder affects adversarial examples. From the figure, we see that both the waveform and the spectrogram of the adversarial example produced by SSED are more similar to the original voice than those produced by the SSED variant, while the perturbation produced by SSED shows lower global energy and clearer local patterns than that produced by the SSED variant. These phenomena are caused by that the saliency map decoder assigns different weights to different samples according to their importance to speaker identification.

3) Effect of the Angular Loss on Performance: Table IV lists the results of SSED and its variant without the angular loss. From the table, we see that the angular loss leads to better performance for TASR, SNR and PESQ.



(a) Waveform (b) Spectrogram

Fig. 2. A comparison example between the original voice and its adversarial examples generated by either SSED or the SSED variant without the saliency map decoder. The SNR of the adversarial examples generated with and without the saliency map decoder are 51.79 dB and 43.47 dB. The PESQ of the adversarial examples generated with and without the saliency map decoder are 3.15 and 2.54.

TABLE IV
PERFORMANCE COMPARISON OF THE SSED WITH OR WITHOUT THE ANGULAR LOSS

Task	Type	TASR (%)	SNR (dB)	PESQ
CSI	with angular loss	**97.3**	**39.07**	**2.29**
	without angular loss	96.8	38.85	2.18
OSI	with angular loss	**98.0**	**39.2**	**2.36**
	without angular loss	96.9	38.92	2.31

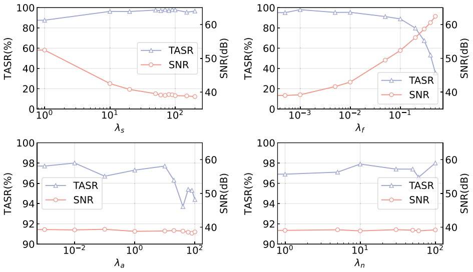


Fig. 3. Effects of the hyper-parameters on the OSI task.

4) Effects of Hyperparameters on Performance: We tuned the hyperparameters λ_s, λ_f, λ_a and λ_n respectively on the OSI task. For tuning each hyperparameter, we fixed the other three to their default values. From Fig. 3, we see that, when λ_f is enlarged which emphasizes the importance of the saliency map decoder L_f, the SNR of SSED is improved while its TASR drops. When λ_s is enlarged which emphasizes the speaker loss L_s, the TASR of SSED is improved while its SNR drops. Interestingly, λ_a which emphasizes the angular loss L_a has slight effects on TASR and SNR, after tuned in a wide range, so as to λ_n which emphasizes L_{norm}. To summarize, SSED is sensitive to λ_f and λ_s, which needs carefully selections in practice; while it is insensitive to λ_a and λ_n. Similar phenomena were observed on CSI as well.

V. CONCLUSION

In this letter, we propose SSED to generate adversarial attack against speaker identification. SSED adopts a novel saliency map decoder to assign different weights to the samples of an utterance according to their importance to the decision of the targeted speaker identification system. It also uses a novel angular loss to push the adversarial example away from its source speaker. Our experimental results on both the CSI and OSI tasks demonstrate the effectiveness of the two novel points with a low computational cost.

REFERENCES

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

[2] H. Tan, L. Wang, H. Zhang, J. Zhang, M. Shafiq, and Z. Gu, "Adversarial attack and defense strategies of speaker recognition systems: A survey," *Electronics*, vol. 11, no. 14, 2022, Art. no. 2183.

[3] S. Cui, B. Huang, J. Huang, and X. Kang, "Synthetic speech detection based on local autoregression and variance statistics," *IEEE Signal Process. Lett.*, vol. 29, pp. 1462–1466, 2022.

[4] A. Luo, E. Li, Y. Liu, X. Kang, and Z. J. Wang, "A capsule network based approach for detection of audio spoofing attacks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6359–6363.

[5] Z. Wang, S. Cui, X. Kang, W. Sun, and Z. Li, "Densely connected convolutional network for audio spoofing detection," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 1352–1360.

[6] H. Wu et al., "Tackling spoofing-aware speaker verification with multi-model fusion," in *Proc. Odyssey*, 2022, pp. 92–99.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–11.

[8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–17.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[10] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Artif. Intell. Saf. Secur*. Boca Raton, FL, USA: Chapman and Hall/CRC, 2018, pp. 99–112.

[11] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Comput. Speech Lang.*, vol. 68, 2021, Art. no. 101199.

[12] 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. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6579–6583.

[13] S. Joshi, J. Villalba, P. Żelasko, L. Moro-Velázquez, and N. Dehak, "Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4811–4826, 2021.

[14] J. Villalba, Y. Zhang, and N. Dehak, "x-Vectors meet adversarial attacks: Benchmarking adversarial robustness in speaker verification," in *Proc. Interspeech*, 2020, pp. 4233–4237.

[15] K. Goto and N. Inoue, "Quasi-Newton adversarial attacks on speaker verification systems," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 527–531.

[16] 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 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6159–6163.

[17] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "AdvPulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2020, pp. 1121–1134.

[18] G. Chen et al., "Who is real bob? Adversarial attacks on speaker recognition systems," in *Proc. IEEE Symp. Secur. Privacy*, 2021, pp. 694–711.

[19] J. Li et al., "Universal adversarial perturbations generative network for speaker recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*. 2020, pp. 1–6.

[20] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Enabling fast and universal audio adversarial attack using generative model," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14129–14137.

[21] S. Lu et al., "Discriminator-free generative adversarial attack," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1544–1552.

[22] 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 Intell. Syst.*, pp. 1–15, 2022.

[23] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, 2021.

[24] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[25] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems," *J. Signal Process. Syst.*, vol. 93, no. 10, pp. 1187–1200, 2021.

[26] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "VoxCeleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, 2020, Art. no. 101027.

[27] J. S. Chung et al., "In defence of metric learning for speaker recognition," in *Proc. Interspeech*, 2020, pp. 2977–2981.

[28] Z. Bai, J. Wang, X.-L. Zhang, and J. Chen, "End-to-end speaker verification via curriculum bipartite ranking weighted binary cross-entropy," *IEEE/ACM Trans. Audio Speech, Lang. Process*., vol. 30, pp. 1330–1344, 2022.

[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. pattern Recognit.*, 2019, pp. 4690–4699.